



Proceedings of the 11th International Conference on GeoComputation

University College London, UK
20th – 22nd July 2011





Proceedings of the 11th International Conference of Geocomputation 2011

University College London
20th – 22nd July 2011

Editors: Tao Cheng
Paul Longley
Claire Ellul
Andy Chow



Proceedings of the 11th International Conference on GeoComputation

University College London
20th – 22nd July 2011

© 2011 the authors of the papers, except where indicated.

All rights reserved. The copyright on each of the papers published in these proceedings remains with the author(s). No part of these proceedings may be reprinted or reproduced or utilized in any form by any electronic, mechanical or other means without permission in writing from the relevant authors.

July 2011, University College London

Programme Committee

Bob Abrahart	University of Nottingham, UK
Mike Batty	University College London, UK
Itzhak Benenson	Tel Aviv University, ISRAEL
Dan Brown	University of Michigan, USA
Chris Brunsdon	University of Leicester, UK
Tao Cheng	University College London, UK
Arzu Coltekin	University of Zurich, Switzerland
Catherine Dibble	University of Maryland, USA
Andrew Evans	University of Leeds, UK
Claire Ellul	University College London, UK
Stewart Fotheringham	National University of Ireland (Maynooth), Ireland
Jianya Gong	Wuhan University, China
Bin Jiang	University of Gävle, Sweden
Shawn Laffan	University of New South Wales, Australia
Brian Lees	Australian National University, Australia
Paul Longley	University College London, UK
Victor Mesev	Florida State University, USA
Alex Singleton	University of Liverpool, UK
Daniel Sui	Ohio State University, USA
Paul Torrens	Arizona State University, USA
Ian Turton	Pennsylvania State University, USA
Peter Whigham	University of Otago, New Zealand

Local Organising Committee

Tao Cheng (Chair)	Civil, Environmental and Geomatic Engineering, UCL
Paul Longley	Geography, UCL
Andy Chow	Civil, Environmental and Geomatic Engineering, UCL
Claire Ellul	Civil, Environmental and Geomatic Engineering, UCL

Welcome

The GeoComputation community has convened a series of international research conferences, initiated in 1996. The conference moves to a different location each year, often alternating between Europe or North America and the Antipodes. The 11th International Conference on Geocomputation 2011 is being held at University College London on 20th - 22nd July 2011, organised jointly by the Department of Civil, Environmental and Geomatic Engineering and the Department of Geography. Scholars from 18 countries and regions registered for this symposium – from Australia, Austria, Canada, China, Finland, Iran, Ireland, Israel, Japan, Malta, Portugal, The Netherlands, Spain, Slovenia, Sweden, Switzerland, Turkey, UK and USA.

The Local Organising Committee received 104 submissions which were each reviewed by a minimum of two reviewers. These proceedings assemble the 78 scheduled for full paper presentation at the conference, following revisions made in the light of reviewer comments. We would like to thank the 50 reviewers and members of the Programme Committee who helped in evaluating the papers.

The strength and breadth of Geocomputation is reflected in the 16 parallel sessions of the 2011 Conference. Eight sessions present advances in methods and algorithms: “*Agent-based Modelling*” (two sessions), “*Genetic Algorithms & Cellular Automata Modelling*”, “*Geographically Weighted Regression*”, “*Geostatistics*”, “*Machine Learning*”, and “*Space-Time Modelling and Analysis*” (two sessions). Four sessions contribute to important domain specific applications of Geocomputation: “*Geodemographics*”, “*Network Complexity*”, “*Location-Based Services*”, and “*GeoVisual and Terrain Analysis*”, with two further sessions with more general focus on environmental and urban studies. The broader environment to Geocomputation provides the focus for two other sessions – one on “*Uncertainty and Accuracy*”, and the other on cloud computation – “*VGI and Computational Infrastructure*”. One poster session consisting of 14 papers is also included.

Five world-renowned scholars have kindly agreed to give keynotes at the Conference. They address ‘*Digital Environments and ‘Real World’ Geographies*’ (Peter Nijkamp, Free University Amsterdam); ‘*Does Visualization with Geocomputation Offer Anything We Didn’t Know Already?*’ (Jo Wood, City University London); ‘*Geographically Weighted Regression and Geocomputation: an Overview of Recent Developments*’ (Stewart Fotheringham, National University of Ireland at Maynooth); and ‘*The Future of Geocomputation*’ (Keith Clarke, University of California, Santa Barbara, USA). Another keynote presentation - ‘*Visualising Space-Time Dynamics: Graphs and Maps, Plots and Clocks*’ is contributed by Mike Batty, CASA, UCL, as a joint event with the ISPRS-sponsored International Symposium on Spatio-Temporal Analysis and Data Mining, which runs immediately before the Conference at same location. This session, and the joint reception scheduled to follow it, is intended as a forum to foster closer dialogue between these two groups.

We are grateful to all the keynote speakers, all the members of the Programme Committee and all the reviewers for their contributions to what we hope will be a very successful conference. Our thanks also go to the conference sponsors Ordnance Survey (GB), Esri UK, John Wiley & Sons, Taylor & Francis, Pion and the STANDARD Project. Special thanks are also due to STANDARD team members: Berk Anbaroglu, Adel Bolbol, James Haworth, Ed Manley, Ioannis Tsapakis, Garavig Tanaksaranond, Artemis Skarlatidou, and Jiaqiu Wang, for their hard work for this event. The help from Lee Philips, Richard Sharp, and from the volunteers of UCL MSc in GIScience group is highly appreciated.

Welcome to London and to UCL!

Tao Cheng, Paul Longley, Claire Ellul, Andy Chow
Local Organising Committee
University College London, 2011

Contents

	Page
Session 1A <i>Geodemographics</i>	
Creating Realistic Synthetic Populations at Varying Spatial Scales: a Comparative Critique of Population Synthesis Techniques	<i>Alison Heppenstall, Kirk Harland, Dianna Smith and Mark Birkin</i> 1
Building Geodemographics on Parallel Graphics Processing Unit Architecture	<i>Muhammad Adnan, Alex Singleton and Paul Longley</i> 7
The Use of Consensus Clustering in Geodemographics	<i>James Cheshire, Muhammad Adnan and Paul Longley</i> 14
Session 1B <i>Genetic Algorithms & Cellular Automata Modelling</i>	
A Comparison of Genetic Algorithms and Reinforcement Learning for Optimising Sustainable Forest Management	<i>Verena Rieser, Derek T. Robinson, Dave Murray-Rust and Mark Rounsevell</i> 20
Evolving Simulation Modeling: Calibrating SLEUTH Using a Genetic Algorithm	<i>M. Clarke-Lauer and Keith Clarke</i> 25
Calibration of a Cellular Automata Model with the Particle Swarm Algorithm	<i>Nuno N. Pinto, António P. Antunes and Josep R. Cladera</i> 30
A Macroscale Cellular Automata Model for Simulating Urban Change in Regional Urban Systems	<i>Nuno N. Pinto, António P. Antunes and Josep R. Cladera</i> 35
Session 2A <i>Agent-Based Modelling (1)</i>	
Simulation of Cholera Diffusion to Compare Transmission Mechanisms	<i>Ellen-Wien Augustijn, Juliana Useya, Raul Zurita-Milla and Frank Osei</i> 39
Modified Navigation Algorithms in Agent-Based Modelling for Fire Evacuation Simulation	<i>Tyng-Rong Roan, Muki Haklay and Claire Ellul</i> 43
Integrating an Agent-Based Model and a Population Microsimulation to Explore Crime Patterns	<i>Nick Malleson and Mark Birkin</i> 50
Understanding Route Choice by Using Agent-based Simulation	<i>Ed Manley, Tao Cheng and Andy Emmonds</i> 54

Session 2B Geostatistics	<i>Page</i>
Merging Areal and Point Data in Medical Geography and Soil Mapping	<i>Pierre Goovaerts</i> 59
Reducing Aggregation Error in Spatial Interaction Models by Location Sampling	<i>Alex Hagen-Zanker and Ying Jin</i> 65
Incorporating Environmental data into Poisson Kriging Approaches for Mapping Patterns of Herbivore Species Abundance in Kruger National Park, South Africa	<i>Ruth Kerry, Pierre Goovaerts, Izak Smit and Ben Ingram</i> 69
Evaluation of Geostatistical Analysis Capability in Wireless Signal Propagation Modeling	<i>Samira Kolyaei and Marjan Yaghooti</i> 76
Session 3A <i>Space-Time Modelling and Analysis (1)</i>	
Where were you? A Time-geographic Approach to Activity Destination Reconstruction	<i>Mark Horner and Joni Downs</i> 84
SimTraj: An Approach to Similar Queries over Trajectories Metric Spaces	<i>Fábio Afonso, Fernanda Barbosa and Armando Rodrigues</i> 87
Measuring Population Shift Bias in Tests of Spatio-Temporal Interaction	<i>Nicholas Malizia, Elizabeth Mack and Sergio Rey</i> <i>Luana Chetcuti Zammit, Kenneth Scerri, Maria Attard, Therese Bajada and Mark Scerri</i> 93
Spatio-Temporal Analysis of Air Pollution Data in Malta	<i>Nicholas Malizia, Elizabeth Mack and Sergio Rey</i> <i>Luana Chetcuti Zammit, Kenneth Scerri, Maria Attard, Therese Bajada and Mark Scerri</i> 99
Session 3B <i>Uncertainty and Accuracy</i>	
Toponym Disambiguation of Landscape Features Using Geomorphometric Characteristics	<i>Curdin Derungs, Ross Purves and Bettina Waldvogel</i> 106
On Estimating Ecotone Occurrence from Land Cover Data Using Type 2 Fuzzy Sets	<i>Pavel Tuček, Jan Caha and Vilem Pechanec</i> 112
Modeling Spatial Relevancy in Context-Aware Systems Using Fuzzy Intervals	<i>Najmeh Samany, Mahmoud Reza Delavar, Nicholas Chrisman and Mohammad Reza Malek</i> 116
Map Comparison for the Evaluation of Spatial Bayesian Models	<i>Colin Robertson</i> 123
On the Use of Grey Information Theory as a Conceptual Framework for Treating Uncertainty in Spatial Systems	<i>Mark Horner</i> 127

Session 4 Posters	<i>Page</i>
GeoComputation and Dialect Lexicography: Methods to Increase Insight in an Interdisciplinary Partnership	<i>Eveline Wandl-Vogt</i> 130
Evaluation of ASTER and LISS III Data in Identification of Saline Soils, Case Study: Regions of Iran	<i>Seyed Kazem Alavipanah, Hamidreza Matinfar, Nader Sarmasti, Mansour Jafarbeglou, Saeed Goodarzimehr, Bahere Khakbaz and Farzam Khosravi</i> 134
Accuracy Assessment for Fuzzy Classification in Tripoli, Libya	<i>Abdulhakim Khmag, Alexis Comber and Pete Fisher</i> 147
Application of Data Mining In Micro-scale Urban Feature Analysis	<i>Ahu Sokmenoglu, Gulen Cagdas and Sevil Sariyildiz</i> 154
MetaHeuristics for a Non-Linear Spatial Sampling Problem	<i>Eric Delmelle</i> 161
Spatial Determinants of Quality of Life in Urban Areas: Does Metropolitan Contiguity Effect?	<i>Ali Goli</i> 168
Fractal Perspectives of GIScience Based on the Leaf Shape Analysis	<i>Pavel Tuček, Lukáš Marek, Vít Pászto, Zbyněk Janoška and Martin Dančák</i> 169
Building a Web-based Cancer Atlas for Saudi Arabia	<i>Khalid Al-Ahmadi, Alison Heppenstall, Linda See and A Al-Zahrani</i> 177
Modelling the Humanitarian Relief through Crowdsourcing, Volunteered Geographical Information and Agent-based modelling: A Test Case - Haiti	<i>Andrew Crooks and Sarah Wise</i> 183
Vector-based Mathematical Morphology	<i>Huayi Wu, Wenxiu Gao</i> 188
Towards Using Geovisual Analytics to Interpret the Output of Geographically Weighted Discriminant Analysis	<i>Peter Foley</i> 195
Discovering Different Regimes of Biodiversity Support Using Decision Tree Learning	<i>Tomasz Stepinski, Denis White and Josue Salazar</i> 203
Knowledge Discovery for Exploring the Relations between Climate Change and Population Dynamics	<i>Budhendra Bhaduri, Xiaohui Cui, Cheng Liu, Jennifer Santos-Hernandez, Benjamin Preston, Jack Schryver, James Nutaro, Stan Hadley, Richard Medina and Hoe Kyoung Kim</i> 208
Location Based Social Networks-Tracking Activity in an Urban Environment in Using Twitter Data	<i>Fabian Neuhaus</i> 212

Session 5A <i>Network Complexity</i>	<i>Page</i>
Modelling Dynamic Space-Time Autocorrelations of Urban Transport Network	<i>Tao Cheng, Jiaqiu Wang, James Haworth, Benjamin Heydecker and Andy Chow</i> 215
Road Network Analysis using Geometric Graphs of β -skeleton	<i>Toshihiro Osaragi and Yuko Hiraga</i> 221
The Head/tail Division Rule for Characterizing the Scaling of Geographic Space	<i>Bin Jiang and Xintao Liu</i> 227
Distance Dependence in the Spatial Structure of China Aviation System: A Complex Network Perspective	<i>Jingyi Lin</i> 231
Session 5B <i>VGI and Computational Infrastructure</i>	
An Automated Method to Assess Data Completeness and Positional Accuracy of OpenStreetMap	<i>Thomas Koukoletsos, Muki Haklay and Claire Ellul</i> 236
Improving Global Land Cover through Crowd- sourcing and Map Integration	<i>Linda See, Steffen Fritz, Ian Mccallum, Christian Schill, Christoph Perger and Michael Obersteiner</i> 242
Geospatial Service Web-Cyber Infrastructure for Service-oriented Geospatial Science	<i>Jianya Gong, Huayi Wu, Wenxiu Gao, Peng Yue, Xinyan Zhu</i> 247
A Universal Framework for Parallel Processing Massive Spatial Data using a Split-and-Merge Paradigm	<i>Xuefeng Guan and Huayi Wu</i> 256
Session 6A <i>Machine Learning</i>	
Using a Moving Window SVMs Classification to Infer Travel Mode from GPS Data	<i>Adel Bolbol, Tao Cheng and James Haworth</i> 262
Putting the Geographical Analysis Machine on the Internet Revisited	<i>Ian Turton and Andy Turner</i> 271
Inverse Estimation of the Point Position from an Image of Kernel Density Estimation	<i>Atsushi Takizawa</i> 275
Kernel Regression for Traffic Prediction Under Missing Data	<i>James Haworth, Tao Cheng, John Shawe-Taylor</i> 280

Session 6B <i>GeoVisual and Terrain Analysis</i>	Page
Selective Progressive Transmission of Vector Data	<i>Fangli Ying, Peter Mooney, Padraig Corcoran and Adam Winstanley</i> 285
Software Prototyping of A Heuristic and Visualized Modeling Environment for Digital Terrain Analysis	<i>Cheng-Zhi Qin, Yan-Jun Lu, A-Xing Zhu and Weili Qiu</i> 290
Interactive Visualisation of Spatial Turnover	<i>Shawn Laffan</i> 295
Automatic Terrain Analysis in Railway Transportation Corridors with Regard to Asset Line-of-Sight from Monocular Video	<i>Thomas Warsop</i> 299
Session 7A <i>Geographically Weighted Regression</i>	
Model Selection in GWR: the Development of a Flexible Bandwidth GWR	<i>Wenbai Yang, A. Stewart Fotheringham and Paul Harris</i> 307
Spatial planning – an inverse problem?	<i>Ricardo Crespo and Adrienne Grêt-Regamey</i> 313
A Spatial Analysis of Perceptions of Health Services Accessibility, Health Status and Geographic Access using GWR	<i>Alexis Comber, Chris Brunsdon and Robert Radburn</i> 318
Distance Metric Selection for Calibrating a Geographically Weighted Regression model	<i>Binbin Lu and Martin Charlton</i> 323
Session 7B <i>Space-Time Modelling and Analysis (2)</i>	
Eight Challenges for Social Flows in a GIS Framework	<i>Clio Andris and Joseph Ferreira</i> 327
A Flexible Model for Haptic-assisted Pedestrian Navigation Mobile Applications	<i>Ricky Jacob, Peter Mooney, Padraig Corcoran and Adam Winstanley</i> 333
Trajectory Data Mining: Classification and Spatio-Temporal Visualization of Mobile Objects	<i>Atsushi Nara and Paul Torrens</i> 338
An Adaptive-Velocity Time-geographic Density Estimator	<i>Joni Downs and Mark Horner</i> 345

Session 8A <i>Location-based Services</i>	<i>Page</i>
Dynamic Planning of Ambulance Location in Leicestershire Locating-allocating Schools Using Metaheuristics and GIS A New Variable for Spatial Accessibility Measurement in Social Infrastructure Planning	<i>Emeka Chukwusa, Alexis Comber and Chris Brunsdon</i> 349 <i>Raul Zurita-Milla, Md. Shamsul Arifin and Otto Huisman</i> 353 <i>Yang Li and Allan Brimicombe</i> 357
 Session 8B <i>Applications (1)</i>	
Temporal Limits to Urban Growth Modelling High-Resolution Estimation and Analysis of Transport Accessibility in the City Agent-based Modeling of Sustainable Urban Development along the Mid-Section of Silk Road in Northwest China	<i>Gargi Chaudhuri and Keith Clarke</i> 363 <i>Itzhak Benenson, Amit Rosental and Karel Martens</i> 372 <i>Yichun Xie and Xiaojin Tan</i> 375
 Session 9A <i>Agent-Based Modelling (2)</i>	
Using Agent-Based Complex Systems to Model Impacts of Policy Decisions on Climate Change Scenarios An Agent-Based Model of Woodland Caribou Habitat- Selection in West Central Alberta: A Behavioural and Ecological Approach Agent Based Modelling and GIS for Community Resource Management: Acequia-based Agriculture SAFEPED: Agent-Based Environment for Estimating Accident Risks at the Road Black Spots	<i>Marta Vallejo</i> 377 <i>Christina Semeniuk, Marco Musiani, Mark Hebblewhite, Scott Grindal and Danielle Marceau</i> 382 <i>Sarah Wise and Andrew Crooks</i> 388 <i>Gennady Waizman, Eilon Blank-Baron and Itzhak Benenson</i> 395

Session 9B <i>Applications (2)</i>	<i>Page</i>
Polyline Averaging Using Distance Surfaces: a Spatial Hurricane Climatology	<i>Kelsey Scheitlin, Victor Mesev and James Elsner</i> 397
Using Fine Resolution Population Data and Spatial Interaction Modeling to Estimate Risk from Airborne Toxic Releases	<i>Jamison Conley and Robert Stewart</i> 402
Defining Spatial Weights Matrices in Ecological Research	<i>Trisalyn Nelson and Colin Robertson</i> 407
Geometric Techniques to Speed up Geospatial Feature Matching	<i>Constantinos Tsirogiannis</i> 411
Author Index	416

Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques

A.J.Heppenstall¹, K.Harland¹, D.M.Smith² and M.H. Birkin¹

¹School of Geography, University of Leeds, Leeds, LS2 9JT

Telephone: (+44) 113 343-3392

Fax: (+44) 113 343-3308

Email:[a.j.heppenstall]; [k.harland98]; [m.h.birkin] @leeds.ac.uk

² Department of Geography, Queen Mary, University of London

Mile End Road, London E1 4NS

Telephone: (+20) 7882 2750

Fax: (+20) 20 7882 7479

Email: d.smith@qmul.ac.uk

1. Introduction

Recent years have seen a rise in the number of methods and applications which require realistic individual-level data/synthetic populations. This trend can be attributed to a number of factors including increases in computational power and storage, a wealth of individual level data (for example, the British Household Panel Survey) and the development of new computational paradigms, such as cellular automata and agent-based modelling (ABM).

Static spatial microsimulation samples a synthetic population (a population built from anonymous survey data at the individual level) which realistically matches the observed population in a geographical zone for a given set of criteria. There is a diverse set of research and policy applications that use synthetic populations in a spatial setting, including: health (Smith et al, 2009, Tomintz and Clarke, 2008, Brown and Harding, 2002), transportation (see, for example, McFadden et al, 1977; Beckman et al, 1996) and water demand estimation (Williamson and Clarke, 1996).

ABM can also use synthesised data as a base population. There has been a rapid uptake in the use of ABM in Geography with applications ranging from simulating the movement of burglars (Malleson et al, 2009) to replicating dynamics in spatial retail markets (Heppenstall et al, 2006). Although the construction of an ABM does not require a complete individual data set, creating an agent population from a realistic synthesised individual dataset can only improve the realism of these models.

There are several established methodologies for generating synthetic populations. The focus of this paper will be on deterministic reweighting (Smith et al, 2009), conditional probability (Monte Carlo simulation) (Birkin and Clarke, 1988, 1989) and simulated annealing (combinatorial optimisation) (Openshaw, 1995; Williamson, Birkin and Rees 1998; Voas and Williamson, 2000, 2001). These methods were selected due to their common application in geography. Many recent spatial microsimulation studies including Anderson (2007), Ballas *et al.* (2005), Voas and Williamson (2000, 2001), Tomintz *et al.* (2008) Smith *et al.* (2009) and Morrissey *et al.* (2008) have adopted a variation on at least one of the three approaches examined here.

The work within this paper critically compares each approach as they are used to generate a synthetic individual level population at three different spatial scales, extending the initial work reported in Voas and Williamson (2000, 2001).

2. Spatial Microsimulation Algorithms

There are numerous algorithms that have been designed or adapted to produce synthetic populations. Here, three approaches that have commonly been adopted in recent years, each one taking a broadly different methodological approach, are reviewed. The three approaches are deterministic reweighting, a large iterative proportional fitting routine, conditional probabilities, which uses statistical joint probabilities, and simulated annealing, a combinatorial optimisation method.

Table 1 provides a summary comparison of the three algorithms.

	Deterministic Reweighting	Conditional Probabilities	Simulated Annealing
Easy setup (is there much pre-processing)?	Yes	Yes	No
Sensitive to specification of constraint order?	Yes	Yes	No
Limit to number of constraints that can be used?	Yes	Yes	No
Requires a sample population?	Yes	No	Yes
Can take forward and backward steps to find an appropriate solution?	No	No	Yes
Stochastic?	No	Yes	Yes
Speed of execution	Fastest	Middle	Slowest

Table 1. Summary comparison of the three algorithms.

3. Data and Experiments

Each of the spatial microsimulation methods discussed is used to produce a synthetic population at the Output Area (OA), Lower Layer Super Output Area (LLSOA) and Middle Layer Super Output Area (MLSOA) spatial scales. The synthesised populations are tested against known Census information, produced at all three geographies to evaluate each algorithmic approach. In summary, each population produced will be tested to examine:

- (i) Reproduction of variables used to constrain each of the synthetic models at each of the different spatial scales.

- (ii) Evaluation of the populations produced against information extracted from the Census of Population 2001 using the constraint variables cross-tabulated against each other.
- (iii) Examination of how reliably information from the sample population **not** included in the model constraints can be captured.
- (iv) Aggregation of outputs from OA to LLSOA and MLSOA and a subsequent evaluation of the aggregated output against Census of Population 2001 at the appropriate geographical level.

The results of each of these experiments will be presented at the conference.

4. Selected Results

4.1 Representing Constraint Variables

Voas and Williamson (2000) stated that all constraint attributes should be well represented in a synthetic population. The purpose of this test is to evaluate how well the constraint attributes are reproduced in each of the algorithms. Populations are synthesised using each algorithm at each spatial scale OA, LLSOA and MLSOA, making a total of nine different synthetic populations being evaluated. The evaluation statistic used was classification error (CE); this is the total absolute error/ 2.

Table 2 shows that only simulated annealing has successfully recreated all of the constraint attributes at all three spatial scales with zero misclassification. The conditional probabilities algorithm produces a reasonable fit for all of the constraints over each scale. However, the classification error almost doubles for each constraint as the geographical scale becomes finer. The deterministic reweighting method produced the worst fit. With the exception of Highest Qualification (which shows a slight decrease in CE, but overall this constraint has a very poor fit to the observed data) all of the constraints show a slight increase in CE as geographical scale becomes finer.

Constraint	DR		CP		SA	
	CE	% CE	CE	% CE	CE	% CE
Middle Layer Super Output Area						
Gender	29,510	4.12	102	0.01	0	0.00
Ethnic Group	14,897	2.08	2,290	0.32	0	0.00
Age	128,999	18.03	144	0.02	0	0.00
Marital Status	95,335	13.33	478	0.07	0	0.00
NSSEC	84,731	11.84	4,378	0.61	0	0.00
Highest Qualification	229,407	32.07	2,569	0.36	0	0.00
Lower Layer Super Output Area						
Gender	30,297	4.23	176	0.02	0	0.00
Ethnic Group	15,631	2.18	4,010	0.56	0	0.00
Age	131,230	18.34	245	0.03	0	0.00

Marital Status	96,453	13.48	842	0.12	0	0.00
NSSEC	88,282	12.34	9,659	1.35	0	0.00
Highest Qualification	228,425	31.93	5,219	0.73	0	0.00
Output Area						
Gender	33,430	4.67	245	0.03	0	0.00
Ethnic Group	16,707	2.34	5,292	0.74	0	0.00
Age	135,673	18.96	418	0.06	0	0.00
Marital Status	98,696	13.80	1,828	0.26	0	0.00
NSSEC	95,117	13.30	21,939	3.07	0	0.00
Highest Qualification	227,720	31.83	11,385	1.59	0	0.00

DR = deterministic reweighting, CP = conditional probabilities, SA = simulated annealing

Table 2. Representation of the model constraints in the synthesised populations.

To investigate the poor fit of the deterministic reweighting algorithm, the number of misclassified people per zone is plotted for the Ethnic Group, Gender and Marital Status constraints at the MSA geography (fig. 1 - 3). The Ethnic Group scatter plot (fig. 1) shows that, despite having almost 15,000 classification errors, the spread of error tracks the line of perfect fit (where each point would reside if the synthesised population matched the observed population exactly). Only small discrepancies exist, but the discrepancies are evident in many geographical zones.

Fig. 2 shows a scatter plot of gender classification errors which are grouped very tightly together. The lack of spread along the line of perfect fit is a reflection that most geographical zones have a relatively balanced population between male and female and do not display the extremes that can be observed in other constraint attributes. Despite the relatively ubiquitous nature of the attribute, many of the geographical zones are some distance away from the perfect fit line; this is reflected in the 29,510 classification errors observed at the MSA geography. This high level of error may be due to the constraint being last in the processing order and the attempt of the algorithm to smooth towards the global mean.

The marital status constraint (fig. 3) is particularly poorly fit by the deterministic reweighting routine. Although this constraint does not have the highest level of associated classification error, it does display a distinct pattern. Most MSA zones have the married category over represented and the single category underrepresented in the synthetic population. This suggests that the algorithm is smoothing towards the distribution of the sample population rather than preserving the distribution observed in the constraint information for each geographical area.

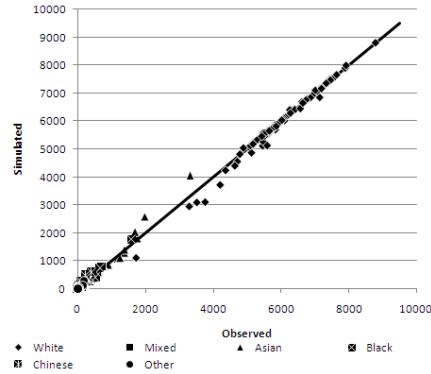


Figure 1. Deterministic reweighting - Ethnic Group misclassification error at MLSOA geography.

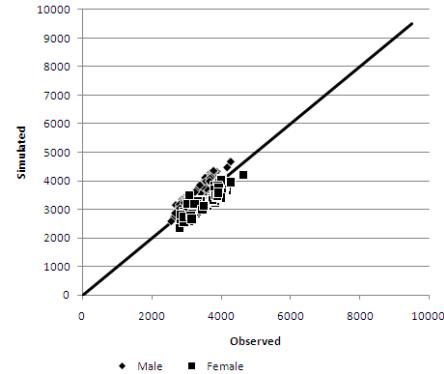


Figure 2. Deterministic reweighting - Gender misclassification error at MLSOA geography.

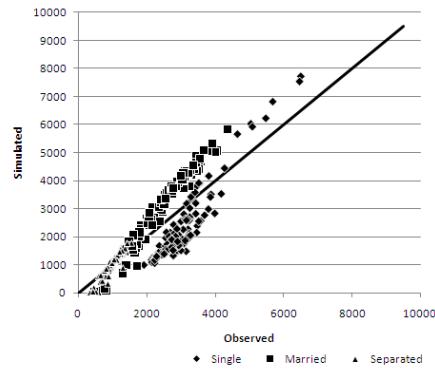


Figure 3. Deterministic reweighting - Marital Status misclassification error at MLSOA geography.

5. Conclusion

The work in this paper has briefly presented selected results of deterministic reweighting, conditional probabilities and simulated annealing spatial microsimulation methods for representing constraint variables at varying spatial scales. Of the three methods assessed, simulated annealing was found to consistently produce the best outcome when fitting constraints. Further conclusions and analysis drawn from the other experiments will be presented at the conference.

6. Acknowledgements

This work forms part of the ESRC funded Modelling Individual Consumer Behaviour project (RES-061-25-0030). Part of this work was funded by a Royal Geographical Society small grants award (SRG 04/09).

7. References

- Anderson B, 2007, *Creating small-area Income Estimates: spatial microsimulation modelling*, Department for Communities and Local Government, Communities and Local Government Publications, London
- Ballas D, Clarke G, Dorling D, Eyre H, Thomas B, Rossiter D, 2005, "SimBritain: a spatial microsimulation approach to population dynamics" *Population, Space and Place*, 11 13-34
- Beckman R J, Baggerly K A and McKay M D, 1996 Creating synthetic baseline populations. *Transportation Research* 30 (6), 415-429
- Birkin M, Clarke M, 1988, "SYNTHESIS - a synthetic spatial information system for urban and regional analysis: methods and examples" *Environment and Planning A* 20 1645 -1671
- Birkin M, Clarke M, 1989, "The generation of individual and household incomes at the small area level using synthesis" *Regional Studies* 23 535 - 548
- Brown L, Harding A, 2002, "Social modelling and public policy: Application of microsimulation modelling in Australia." *Jasss-the Journal of Artificial Societies and Social Simulation* 5(4)
- Heppenstall AJ, Evans AJ, Birkin MH, 2006, "Application of Multi-Agent Systems to Modelling a Dynamic, Locally Interacting Retail Market" *Jasss-the Journal of Artificial Societies and Social Simulation*. 9(3)
- Malleson NS, Heppenstall AJ, See LM, "Simulating Burglary with an Agent-Based Model". *Computers, Environment and Urban Systems*. In review
- McFadden D, Cosslett S, Duguay G and Jung W, 1977 *Demographic Data for Policy Analysis*. Urban Travel Demand Forecasting Project, Final Report Series, Vol VIII. Institute of Transportation Studies, University of California, Berkeley and Irvine
- Morrissey K, Clarke G, Ballas D, Hynes S, O'Donoghue C, 2008 "Examining access to GP services in rural Ireland using microsimulation analysis" *Area*, 40(3) 354-364
- Openshaw S, Rao L, 1995 "Algorithms for reengineering 1991 Census geography" *Environment and Planning A* 27 425-446
- Smith DM, Clarke GP, Harland K, 2009, Improving the synthetic data generation process in spatial microsimulation models *Environment and Planning A* 41 1251 – 1268
- Tomintz MN, GP Clarke, 2008 "The geography of smoking in Leeds: estimating individual smoking rates and the implications for the location of stop smoking services." *Area* 40(3): 341-353
- Voas D, Williamson P, 2000, "An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata" *International Journal of Population Geography* 6 349 - 366
- Voas D, Williamson P, 2001, "Evaluating goodness-of-fit measures for synthetic microdata" *Geographical and Environmental Modelling* 5 177 - 200
- Williamson P, Birkin M, Rees P, 1998, "The estimation of population microdata by using data from small area statistics and samples of anonymised records" *Environment and Planning A* 30 785 – 816
- Williamson P, Clarke GP, 1996, Estimating small-area demands for water with the use of microsimulation. *Microsimulation for urban and regional policy analysis*. Ed: Clarke, GP. London, Pion 117-148

Building Geodemographics on Parallel Graphics Processing Unit Architecture

1. Introduction

Geodemographic classification categorise small geographic areas into a series of discrete categories that aim to represent the multidimensional characteristics of individuals living within these neighbourhoods. Real-time geodemographic classification is the vision for an online and automated web based system that enables users to build, visualise and test a bespoke classification within a short time period (probably in minutes). There have been a number of technological advances which are enabling us to develop online systems for the creation of real-time classifications. This paper presents a summary of our research to date in this area, and cumulates in a pilot real-time geodemographic information system for specification, estimation and testing of classifications on the fly.

There are numerous methodologies for creating geodemographic classifications which differ based on the datasets used, the normalisation technique applied, the method of aggregation and finally, the visualisation techniques used. Geodemographic classifications are created by a clustering algorithm searching the attribute space of a matrix of standardised input data comprising a row for each small area (however defined) and a column for each attribute measure. For example, Vickers and Rees (2007) used k -means clustering for the creation of the National Statistics Output Area Classification (OAC) with data derived entirely from the 2001 Census of the Population. The k -means algorithm is a commonly used method for the geocomputation of geodemographic classification (Harris et al, 2005), however, in its original form, k -means is unstable and relatively sensitive to outlier values within the input data matrix. Because of this instability the algorithm requires multiple runs in order to ensure a robust result. For example, Singleton and Longley (2008) created a geodemographic classification using k -means with approximately 10,000 runs.

The geodemographic classification system described in this paper uses a parallel implementation of k -means (see Adnan et al, 2010) build upon NVIDIA's Computer Unified Device Architecture (CUDA)ⁱ. CUDA allows different processes to run in parallel on the Graphical Processing Units (GPUs) of NVIDIA's graphics cards enabling greater computational power than standard non parallel k -means clustering.

2. Clustering by parallel k -means

The K -means clustering algorithm has remained the core algorithm used in the creation of geodemographic classifications. K -means seeks to find a set of cluster centroids that minimises expression (3) below.

$$V = \sum_{x=1}^n \sum_{y=1}^n (z_x - \mu_y)^2 \quad (1)$$

Where n is the number of clusters, μ_y is the mean centroid of all the points z_x in cluster y . The k -means algorithm assigns a set of n seeds within the data set and then proceeds by assigning each data point to its nearest seed. Cluster centroids are then created for each cluster, and the data points are

assigned to the nearest centroid. The algorithm, then, re-calculates the cluster centroids and repeats these steps until a convergence criterion is met (usually when the switching of data points no longer takes place between the clusters).

This paper presents a parallel implementation of the *k-means* algorithm using CUDA. CUDA is a general-purpose parallel computing architecture that uses the GPUs of NVIDIA graphics cards to solve complex computational problems. A typical CUDA enabled NVIDIA graphics card has a number of GPUs and a set of memory capable of storing a reasonably large amounts of data. For example, “GeForce 8400M GT” graphics card has 16 GPUs and 512MB of internal memory. CUDA requires that the computational problem to be programmed in the C language for parallel processing.

Our proposed parallel *k*-means algorithm via CUDA works as follows:

Total number of runs is specified by N .

- a) Central Processing Unit (CPU) prepares the data points and counts the number of GPUs available on the NVIDIA graphics card. Afterwards the CPU uploads the data points and code instructing one *k*-means run to each GPU.
- b) GPU performs *k*-means clustering on the data points by minimizing expression (1). When an optimal solution is achieved, GPU returns the result to CPU and claims the next *k*-means run from CPU if there are any.
- c) CPU stores the results returned by GPUs in a local data structure contained in Random Access Memory (RAM). CPU keeps on delegating requests to GPUs until number of runs are less than N .
- d) If number of runs is equal to N , CPU compares the “within sum of squares distance” optimisation criteria of all the runs.
- e) The optimal solution is the one that has minimum “within sum of squares distance”.

In order to compare the “computational time” of *k*-means and parallel *k*-means, we ran *k*-means and parallel *k*-means for ($k=2-30$) cluster solutions at Output Area level using the London datasets, and then compared the time taken for each algorithm to converge on a specified number of clusters. For each value of k , each algorithm was run 100 times and the results are shown in Figure 1. “Computational time” represents the time an algorithm takes to complete 100 iterations for each value of k . The hardware used for this evaluation comprised an “Intel Core2 Duo 2.10GHz” CPU, 4GB RAM, and “GeForce 8600M GS” NVIDIA graphics card. The graphics card has 16 GPUs and 512 MB of RAM.

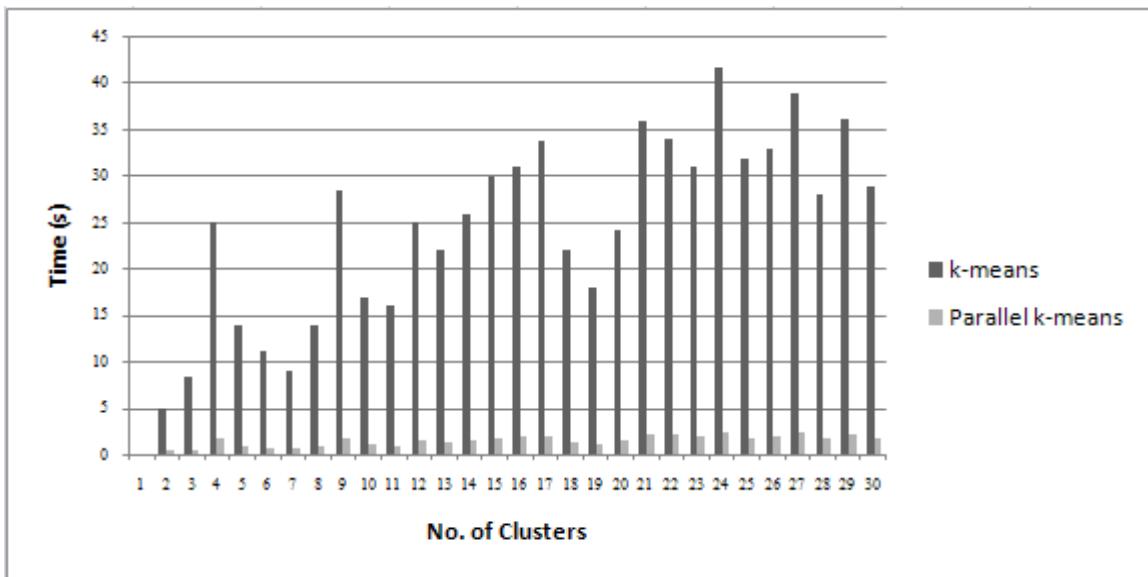


Figure 1: Output Area (OA) level results for the two clustering algorithms

Figure 1 indicates that parallel k -means is a lot faster than k -means clustering algorithm, and thus is the best choice for an online geodemographic system.

3. Creating a bespoke real-time geodemographic classification

A real-time geodemographic information system produces a classification in four steps which are Specification, Normalisation, clustering by Parallel k -means, and Visualisation. In the first step, user selects variables and their weightings. Weighting describes the importance of variables in the classification. User also specifies the number of Geodemographic Classes. In the second step, information system normalises the data using one of the normalisation techniques e.g. Z-scores, Range Standardisation, or Principal Component Analysis. In the third step, the system clusters the data using Parallel k -means clustering algorithm. In the final step, the information system shows the result in the form of maps and statistics.

We can represent the real-time geodemographic information system as a block diagram with different components communicating with each other. Following Figure 2 illustrates this.

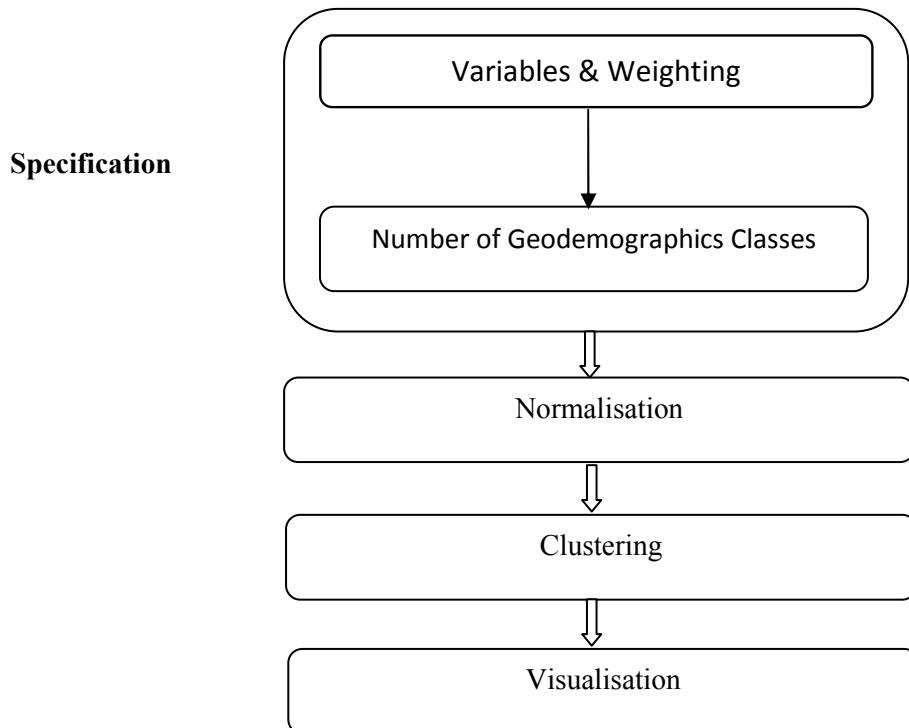


Figure 2: Block diagram of a real-time Geodemographics information System

The remainder of this paper outlines our beta real-time geodemographics information system. This uses the 2001 Census inputs to the National Statistics Output Area Classification (Vickers & Rees, 2007) aggregated at Output Area (OA). The normalisation technique incorporated into the system is z-score, and it uses parallel k -means to cluster the data.

3.1 Specification of Inputs

First step in creating a classification is the specification of input variables and an assignment of a weight of relative importance. This is shown in Figure 3 where the ‘Born outside the UK’ variable will have highest weight in the output classification.

OAC Variables	Selected Variabels	Weighting
Age 5-14	Age 0-4	 1
Age 45-64	Age 25-44	 2
Age 65+	Born Outside the UK	 3
Black african, Black Caribbean or Other Black	Indian, Pakistani or Bangladeshi	 1
Population Density		
Divorced		
Single person household (not pensioner)		
Single pensioner household		
Lone Parent household		
Two adults no children		
Households with non-dependant children		
Rent (Public)		

Figure 3: Specification of variables and their weight

After variables have been selected, the number of classes in the output classification can be specified. This is shown in Figure 4.

Select number of geodemographic classes : 

Figure 4: Specification of the number of geodemographic classes

3.2 Results

Based on the previous selected inputs, the system produced a classification for London. This is shown in Figure 5.

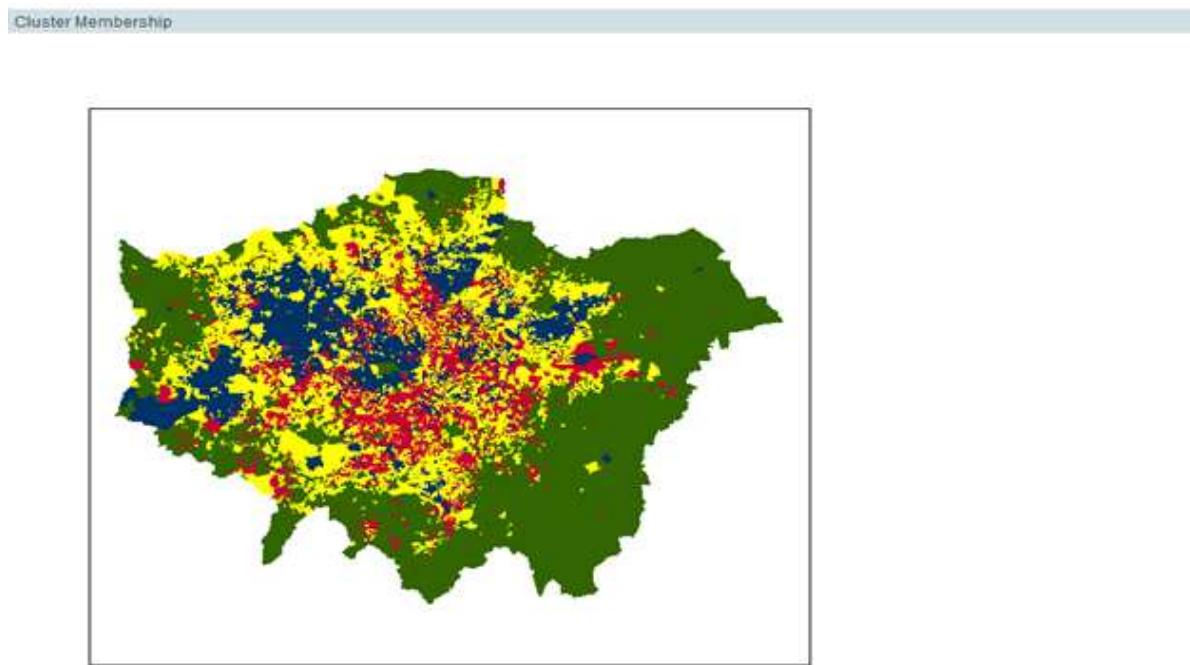


Figure 5: Classification produced for London based on selected variables

The system also gives considerable information about the sizes of clusters, which is important when the objective of building a classification is to produce clusters of reasonably equal sizes.

Within Sum of Squares	
Within sum of squares : 30311.46216817008	
Clusters	
Cluster No. Cluster Size	
1	4659.0
2	5285.0
3	6934.0
4	7262.0

Figure 6: Cluster Membership and Within Sum of Squares

4. Conclusion and Future Research

This paper has presented our pilot real-time geodemographic classification system based on CUDA parallel infrastructure. The system enables users to compile geodemographic classifications quickly (possibly within minutes) utilising the multiple processor architecture of graphics cards. Given that these technologies are now available as part of typical data centre and cloud architectures (e.g. Amazon EC2) we see this as a very scalable solution which could compile classifications based on inputs for more extensive geographies.

Future research aims to evolve the testing procedures used to produce the classifications. Also, alternate clustering algorithms could be incorporated into the system to allow users more flexibility when creating geodemographic classifications.

5. References

- Adnan, M., Longley, P.A., Singleton, A.D., Brunsdon, C. (2010) [Towards Real-Time Geodemographics: Clustering Algorithm Performance for Large Multidimensional Spatial Databases](#). Transactions in GIS, 14(3), 283 – 297.
- Harris, R., Sleight, P., Webber, R. (2005). Geodemographics, GIS and Neighbourhood Targeting. Wiley, London.
- Singleton, A.D., Longley, P.A (2008). Creating open source geodemographic classifications for Higher Education applications. Papers in Regional Science, 88(3), 643-666.
- Vickers, D.W. and Rees, P.H. (2007). Creating the National Statistics 2001 Output Area Classification. Journal of the Royal Statistical Society, Series A. 170(2), 379-403.

ⁱ For more information on CUDA see the Nvidia website: http://www.nvidia.com/object/cuda_home.html

The Use of Consensus Clustering in Geodemographics

J. A. Cheshire¹, M. Adnan¹, P.A. Longley

¹UCL Department of Geography,
Gower Street, London, WC1E 6BT.
james.cheshire@ucl.ac.uk

1. Introduction

Geodemographic classifications require clustering algorithms to partition the records of large multidimensional datasets into groups sharing similar characteristics. Many clustering algorithms have been developed but few have been as widely implemented as the "traditional" methods such as K-means or Ward's hierarchical clustering (Jain, 2010). No two methods create the same result, and multiple iterations of the same method may produce different clusters; it is left to the user to subjectively decide the best outcome. In addition most methods require an *a priori* impression of the number of groups in the data. This abstract outlines a new approach, known as consensus clustering, that utilises familiar clustering methods to produce more consistent results. The method offsets the weaknesses of one type of clustering with the strengths of another by establishing the consistent average outcome from multiple algorithms (Simpson et al. 2010). Consensus clustering has an additional advantage in that it provides a number of metrics that inform the researcher about the inherent groups within the data, and the robustness of the final cluster outcome. Still in its early stages of development, and largely applied in the fields of genetics and bioinformatics, the method has some performance issues when using large datasets but we are confident these can be overcome.

2. Consensus Clustering

Contemporary geodemographic classifications utilise clustering methods in isolation from one another; they do not combine their results in any way. Consensus clustering, proposed by Monti et al. (2003) and extended by Simpson et al. (2010), presents an alternative approach by representing the consensus across multiple runs of a clustering algorithm to determine the number of clusters in the data. This is especially useful when using methods that rely on random seeding to allocate the initial clusters (Monti et al., 2003). Confidence in the result will increase if the multiple clustering algorithms, or parameterisations of a single algorithm, produce comparable results. The output metrics from the Simpson et al. (2010) methodology inform the most appropriate clustering methodology in addition to indicating the optimum number of clusters.

Clustering was undertaken using the clusterCons package, developed by Simpson et al. (2010). A proportion of rows are sampled before clustering with the chosen algorithm and parameters. In this study we utilise the Ward's, K-Means and PAM algorithms. The sampling and clustering is repeated many times gauge the impact of feature removal. The results from each iteration, are stored in a consensus matrix which contains for each pair of items the proportion of the clustering runs in which they are clustered together. A merge matrix provides a way of combining the cluster

outcomes from multiple methods by weighted averaging of their respective consensus matrices. The weighting can be adjusted to increase/ decrease the influence of certain cluster methods. In this case all three are treated as equal. This process gives an indication of the cluster reliability because features consistently grouped together are more likely to be similar than those appearing in the same group less frequently. The merge matrix can then be clustered to yield the final outcome. The advantage of this approach is that it accounts for the different classification properties in each of the algorithms discussed above.

In addition to testing three algorithms, we group the data into a range of clusters. The optimal number in this case is defined by the criteria of Monti et al. (2003) who state that the true cluster number (k) can be estimated by finding the value of k at which there is the greatest change in cumulative density function (CDF) calculated from the consensus matrix across a range of possible values of k . By putting the unique elements into descending order it is possible to calculate a cumulative density function $CDF(c)$ defined over the range $c=[0,1]$ using the following equation.

$$CDF(c) = \frac{\sum_{i < j} 1\{M(i, j) \leq c\}}{N(N - 1)/2} \quad (3)$$

It is then possible to calculate the area under the curve, AUC as follows:

$$AUC = \sum [x_i - x_{i-1}] CDF(x_i) \quad (4)$$

where x_i is the current element of CDF and m is the number of elements. If every iteration from the consensus clustering identifies the same groups then the matrix elements will be either 0 or 1, thus producing an $AUC= 1$. This provides the benchmark against which to compare the different clustering results. By plotting the difference in AUC values it is possible to identify the appropriate cluster number as it exhibits the greatest reduction. Once the optimal number of clusters has been identified it is possible to re-cluster the merge matrix. The advantage of this approach is the stability in the results produced due to the removal of bias in the clustering structure unique to each clustering technique.

4. Data and Methods

For demonstration purposes we have taken a small dataset covering the London Borough of Southwark and the City of London. The boroughs represent a range of social characteristics. Their combined population is approximately 260,000 across 770 Output Areas (OAs). Each OA has the same 41 variables as the Output Area Classification (OAC) (see Vickers and Rees, 2007), standardised to z-scores. The data were consensus clustered over a range of k from 5 to 30. Figure 2 plots k against the change in AUC values. The greatest difference in AUC value occurs between 13 and 14 clusters, suggesting that 14 clusters will provide the optimal outcome. The resulting merge matrix was therefore clustered into 14 groups. In addition

conventional clustering without a final merge matrix was performed for comparison.

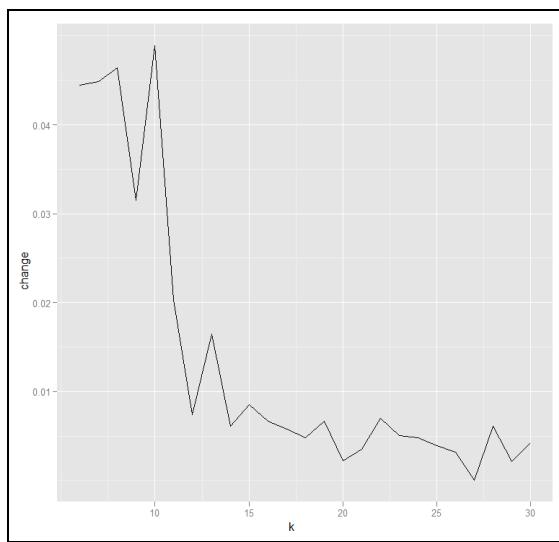


Figure 2: The change in AUC values at a range of k from 1 to 30.

5. Results and Discussion

Figure 3 demonstrates that the clustered merge matrix facilitates a more consistent outcome across all three methods with K-Means and PAM being almost identical. One of the most useful metrics from consensus clustering is the robustness measure mapped in Figure 4. The darker colours (signalling higher robustness values) are more prevalent when the merge matrix is clustered and there are significant improvements in the mean values when compared with the standard clustering approaches. In this case PAM produces the most robust cluster outcome that could be used as a basis for a final classification in this context.

Aside from the stability of its outcomes, one of the key advantages of the consensus clustering methodology is the metrics produced that can help inform the decision about the optimal number of clusters to use. In many contexts "optimal" can be defined quantitatively, but in geodemographics the outcomes are generally mapped, assigned group names and provide an important contextual basis for further research. For these reasons "optimal" in the quantitative sense, such as with the lowest within sum of squares value in the case of K-Means, may not be optimal in the practical sense. Consensus clustering does not circumvent these issues, but it does provide more information on which to base decisions. For example, in Figure 2 it is clear that a transition AUC values occurs between 13 and 14 clusters, partitioning the data further will clearly have less of an impact on the final classification (in terms of its robustness) than partitioning into fewer clusters.

A practical constraint to this methodology is its computational intensity. A national-level classification could not be produced at OA level on a standard desktop workstation, for example. It is our intention to integrate the approach with ongoing research into the creation of geodemographic classifications using NVIDIA's Computer Unified Device Architecture (see Adnan et al. (2010) for more information). This process would enable the consensus clustering to be undertaken many times faster and facilitate fine-scale classifications on a national level.

In conclusion, this abstract has sought to outline consensus clustering in a geodemographics context. The method has demonstrated a strong potential for developing stable classifications and overcomes several of the limitations associated with the conventional implementation of well-known clustering techniques. More work is required to decrease its computation time and also investigate the practical relevance of the results when building a geodemographic classification.

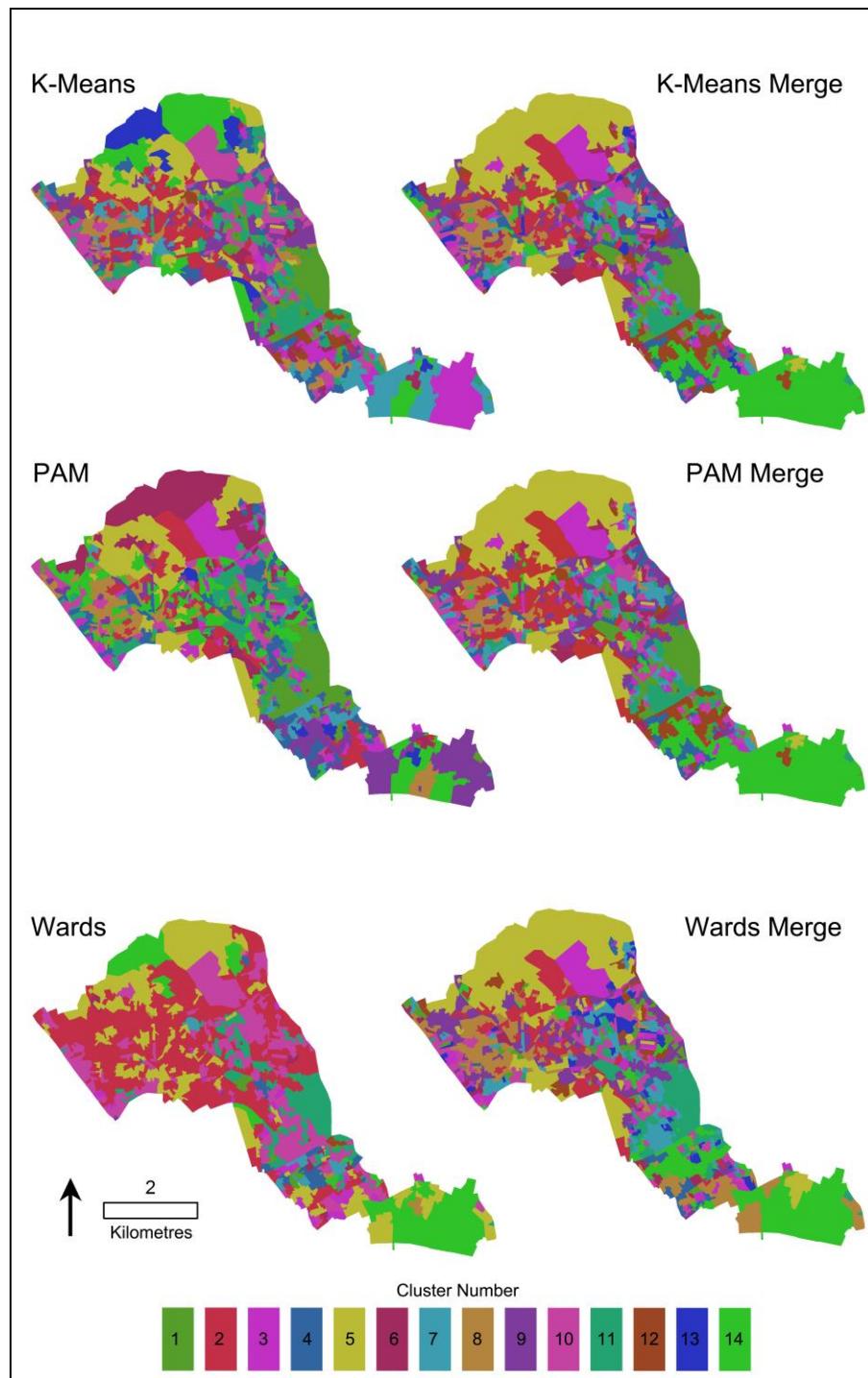


Figure 3: The mapped cluster outcomes from conventional clustering (on the right hand side) and merged consensus clustering (left hand side).

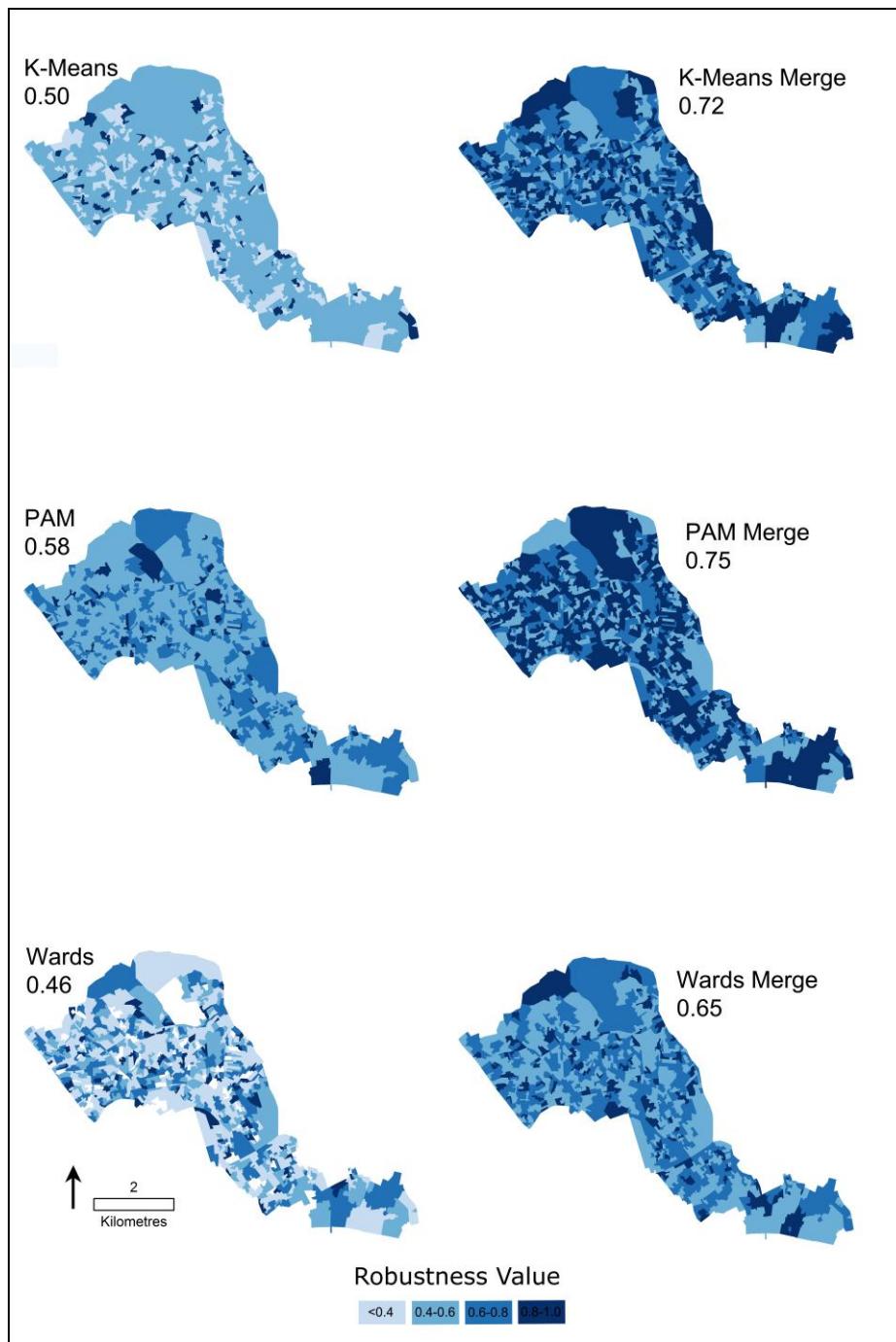


Figure 4: The mapped cluster robustness values outcomes from conventional clustering (on the right hand side) and merged consensus clustering (left hand side). Mean robustness values are also shown.

6. References

- Adnan, M., Singleton, S. Longley, P. 2010. *Parallel K-Means Clustering Using Graphical Processing Units for the Geocomputation of Real-Time Geodemographics*. Proceedings of the GIS Research UK 18th Annual Conference. University College London.
- Jain, A. 2010. Data Clustering: 50 years beyond K-Means. *Pattern Recognition Letters*. 31: 651-666.

Monti, S., Tamayo, P., Mesirov, J., Golub, T. 2003. Consensus Clustering: A Resampling Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*. 52: 91-118.

Simpson, I., Armstrong, D., Jarman, A. 2010. Merged Consensus Clustering to Assess and Improve Class Discovery with Microarray Data. *BMC Bioinformatics*, 11: 590.

West, M. 2002. Bayesian Factor Regression Models in the Large p , Small n Paradigm, *Bayesian Statistics*. 7.

Vickers, D.W., Rees, P.H. 2007. Creating the National Statistics 2001 Output Area Classification. *Journal of the Royal Statistical Society, Series A*. 170 (2), 379-403.

A Comparison of Genetic Algorithms and Reinforcement Learning for Optimising Sustainable Forest Management

Verena Rieser, Derek T. Robinson, Dave Murray-Rust, Mark Rounsevell

School of GeoSciences, University of Edinburgh, UK
Telephone: (+44) (0)131 650 2270
Fax: (+44) (0)131 650 2524
Email: verena.rieser@ed.ac.uk

1. Introduction

Sustainable forest management is defined as “*the stewardship and use of forests and forest lands in a way, and at a rate, that maintains their biodiversity, productivity, regeneration capacity, vitality and their potential to fulfil, now and in the future, relevant ecological, economic and social functions [...].*” (MCPFE, 1994). As such, forest management has to satisfy multiple and often conflicting goals. Furthermore, forest planning is characterised by the long-term horizon of its outcomes. Since long-term plans are made in the face of uncertain futures, long-term sustainable forest management should incorporate some measure of risk. Uncertainty emerges from a variety of sources, including irregular or unknown fluctuations in the demand for timber, or the occurrence of extreme events. In addition, forest management is dynamic in time and space, for example, different stands have different properties, and the likelihood of stochastic events may change over time. Forest planning may be suboptimal if it ignores these sources of uncertainty and risk.

Previous work on multi-objective optimisation in forest management has mainly used heuristic search methods. For example, Bettinger et al. (2002), Pukkala and Kurttila (2005) compare various heuristic optimisation techniques and conclude that Genetic Algorithms (GAs) perform well for more complex spatial problems. However, the studies did not investigate the algorithms' performance under uncertainty.

Reinforcement Learning (RL) is an alternative approach for optimal policy selection. RL is a Machine Learning approach frequently used with agent-based systems (Sutton and Barto, 1998). Contemporary research using RL in the context of forest management has shown that it can find robust optimal solutions to multi-objective forest management problems, e.g. (Bone and Dragicevic, 2009). To further explore the potentials that RL provides over heuristic optimisation approaches, we perform a systematic comparison between RL and GA for sustainable forest management for tasks with increasing uncertainty.

2. Problem Descriptions for Sustainable Forest Management

We present several different hypothetical task environments that are used to test the performance of GA and RL. The task descriptions are meant to provide a proof-of-

concept and are not striving to incorporate the multitude of complex factors in a real-world task environment. In particular, we investigate three aspects of the forest management problem with increasing levels of uncertainty: (1) multi-objective planning, (2) temporal planning with increasing uncertainty over time, (3) planning in environments, which are dynamic in time and space.

The overall task is to decide on a management option for a forest management unit (a “cell”), where the two management options available are to *preserve* or to *harvest* a cell. For task types (1) and (3) the optimisation task is to decide *how many* cells to harvest according some trade-off, reflected in the multi-objective goal. The forest is composed of 10 cells, where the decision for each of the cells is made sequentially. Task type (2) deals with temporal decision making, where the optimisation task is to decide *when* to harvest an individual cell over 10 time intervals.

2.1 Task 1: Multi-objective goal

The multi-objective goal implements the trade-off between economic return versus forest conservation: to satisfy the existing demand for timber while cutting as few forest cells as possible. Equation (1) formulates the objective as a weighted sum:

$$\text{objective} = w_f \times \text{forestCells} + (-w_d) \times \text{unsatisfiedDemand}; \quad (1)$$

We assume that the environment is static and behaves in a deterministic way, e.g. the demand can always be satisfied by harvesting five cells, and each cell has the same potential to satisfy demand.

2.2 Task 2: Increasing uncertainty over time

In Task (2) we explore uncertainty, which is introduced by the temporal nature of forest management. Within our modeling framework uncertainty increases over time, which is operationalised as an increasing probability of disturbance affecting a forest cell.

2.3 Task 3: Spatial Dynamics

In Task (3) we model the likelihood of forest disturbance as a function of tree age, similar to Bone and Dragicevic (2009). However, we extend the model to also include the spatial proximity to neighbouring cells and their average age. This implements the notion that forest disturbances tend to spread. The likelihood of forest disturbance is now a linear function of the cell's own age and the average age of its neighbouring cells, where we use a Moore neighbourhood. The cell's age is also positively related to the amount of demand it can satisfy: the older the cell, the more demand it can satisfy.

3. Problem Implementation

3.1 Problem Implementation in RL

RL addresses the problem of how a forest manager should take actions in an uncertain environment so as to maximise some notion of cumulative, long-term utility or “reward”. RL uses Markov Decision Processes (MDPs) as its underlying representation for decision making and learning. At each time step t the process is in some state s_t and the forest manager may choose any action $a(s)$, that is available in state s . The process responds at the next time step by moving into a new state s' according to the probability $P(s'|s,a)$, which is defined by the transition function $T_{ss'}$, and giving the decision maker a corresponding reward $R_{ss'}$ (see (Sutton and Barto, 1998) for further details). In our case, the reward corresponds to the multi-objective goal as formulated by Equation (1).

We use an implementation of the well-known SARSA algorithm. The state-action space of the MDP is defined as in fig.1.

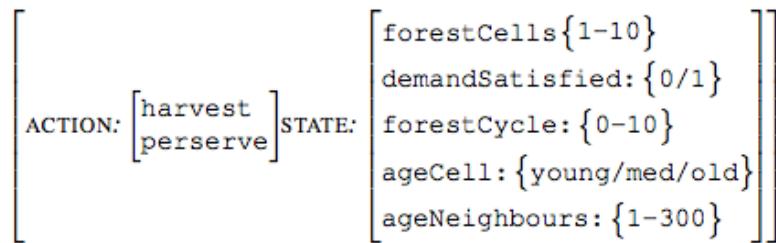


Figure 1. RL State-action space for the forest management problem

The state space keeps track of the number of preserved *forestCells* and whether the *demand* is satisfied or not. The feature *forestCycle* is only used for Task 2 to keep track of the temporal progression. The feature *ageCell* and *ageNeighbours* are only used for Task 3.

3.2 Problem Implementation in GA

Genetic algorithms (GAs) use mechanisms inspired by biological evolution: reproduction, mutation, recombination, and selection (see (Holland, 1975) for further details). We implement GA using binary encoding, as widely used in the forest modelling community, e.g. (Falcao and Borges, 2001; Pukkala, 2006). A gene represents a cell and an allele a binary forest management option. For Task (1) and Task (3), a chromosome represents the whole forest of 10 cells and the binary options represent preserve or harvest. For the temporal problem type in Task (3), a chromosome represents same cell over time. The fitness function corresponds to the multi-objective goal in Equation (1).

4. Results

RL outperforms GA with increasing significance the more uncertainty is introduced into the planning environment. We explain RL's superior performance by its ability to explicitly represent uncertainty in its transition function and to monitor dynamic changes in the environment in its state-space. Table 1 summarises the results and reports the average performance of RL and GA in terms of their average objective value (see Equation 1). We compare them for significant differences using a 2-tailed paired Student's T-test ($n=300$). Note that, subtasks (denoted by x.x) use different weights in their objective function. We will discuss and interpret the results in more detail in the full version of the paper.

Task	GA	RL
Task 1.1	95.00 (± 0.00)	95.00 (± 0.00)
Task 1.2	5.00 (± 0.00)	5.00 (± 0.00)
Task 2.1	-6.47 (± 9.03)	-4.63 (± 8.23) *
Task 2.2	12.27 (± 8.25)	14.13 (± 6.41) **
Task 3	-10.80 (± 31.85)	15.00 (± 13.09) ***

Table 1. Comparing mean performance of RL and GA for task types with increasing uncertainty, where *denotes $p<0.01$, ** denotes $p<0.005$, and *** $p<0.001$.

5. Discussion

Our implementation of GA follows a binary encoding as widely used in the forest modelling community. Unlike RL, this implementation of GA doesn't have an internal representation of the decision process, e.g. feature states, transition probabilities, or the expected return of taking an action in a state, as used by MDPs. In future work, we will investigate the performance of advanced evolutionary algorithms, such as Linear Classifier Systems (Holland, 1975). We will also test the algorithms with real data.

7. References

- P. Bettinger, D. Graetz, J. Sessions and W. Chung. Eight heuristic techniques applied to three increasingly difficult wildlife planning problems. *Silva Fennica* 36(2):561-584.
- C. Bone and S. Dragicevic. GIS and intelligent agents for multiobjective natural resource allocation: A Reinforcement Learning approach. *Transactions in GIS*, 13:253–272, 2009.
- A.O. Falcao and J.G. Borges. Designing an evolution program for solving integer forest management scheduling models: an application in Portugal. *Forest Science*, 47(2):158—168, 2001.
- J. H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.

MCPFE. Ministerial conference on the protection of forests in Europe. Documents, 16-17 June 1994.

T.Pukkala and M. Kurtila. Examining the performance of six heuristic optimisation techniques in different forest planning problems. *Silva Fennica*, 39(1), 2005.

T. Pukkala. The use of multi-criteria decision analysis and multi-objective optimisation in forest management. In Sustainable Forest Management: Growth Models of Europe. Springer Berlin / Heidelberg, 2006.

R. Sutton and A. Barto. Reinforcement Learning: An Introduction. MIT Press, 1998.

Evolving Simulation Modeling: Calibrating SLEUTH Using a Genetic Algorithm

M. D. Clarke-Lauer¹ and Keith. C. Clarke²

¹California State University, Sacramento, 625 Woodside Sierra #2, Sacramento, CA, 95825, USA
 Telephone: 1-626-437-1552
 Email: clarkelm@ecs.csus.edu

²Department of Geography, University of California, Santa Barbara, Santa Barbara CA 93106-4060, USA,
 Telephone: 1-805-456-2827
 Fax: 1-805-893-2578
 Email:kclarke@geog.ucsb.edu

1. Introduction

SLEUTH is a simulation model for urban growth and land use changes at geographic scales. The model couples two cellular automata, and uses input data to capture past behavior as parameters during calibration. Calibration uses brute force methods, requiring either long execution times, or parallel computing. We describe the implementation of a genetic algorithm (GA) that reduced calibration time and enhanced model accuracy. While the model has been successfully applied worldwide (Clarke et al. 2007), computation time remains an obstacle to effective calibration. By designing a GA to work in conjunction with SLEUTH, the computation time was reduced by 80%, while the accuracy was improved.

2. SLEUTH

SLEUTH uses two complex cellular automata operating on a geographic region represented by a two-dimensional cellular grid. Every cell can perform a transition to another state, directed by a transition function and the values in adjacent cells (Clarke, et al. 1997). Cellular automata models have revolutionized urban modeling (Torrens and O'Sullivan, 2001), and are used to simulate various natural and man-made phenomena.

SLEUTH simulates urban growth and land use dynamics when calibrated with a set of mapped data reflecting past patterns. A sequence of growth rules is applied to the cells, each controlled by a set of coefficients that encapsulate the dynamics of a region (diffusion, breed, spread, road gravity, and slope). These values are not initially known and require extensive calibration to determine. SLEUTH's code automates the calibration process, nevertheless the user is still required to guide the calibration phases (Silva and Clarke, 2005). The model provides thirteen metrics describing the fit of the calibration coefficients, with the best set being selected for forecasting. The Optimal Sleuth Metric, a product of eight of the metrics, is best for optimizing calibration (Dietzel and Clarke 2007).

Implementation requires calibration (determining the best coefficients) and predicting (modeling into the future). The calibration phase simulates historical change and compares it to known data to determine how accurately the model simulates growth (Jantz et al., 2010). SLEUTH repeatedly applies sets of the five coefficients to determine which yields the highest OSM. Coefficients consist of numbers between 0 and 100, the

entire search space constituting 10^5 coefficient sets. The brute force approach performs three passes through the search space, with each run the search granularity gets smaller. Two potential problems emerge, the first being the sheer computation time required. Each of the three calibration phases requires at least 2,000-10,000 iterations. Monte Carlo methods minimize within-run variability but further increase computation time. A recent application required over 6 months of CPU time.

The phased and stepped brute force approach may become unable to break free from a local optimum, since large areas of the search space are eliminated from the solution domain. Using a GA in the calibration addresses both problems. By allowing the values to be randomly, but evenly, distributed throughout the search space and by encouraging the best solutions to survive, both speed and accuracy can be improved. We applied the GA at the code level by replacing the source code that implements the brute force calibration. SLEUTH's modularity means that only the driver level function needed alteration, all of the model behavior modules remained unaffected.

3. GA Design

GAs simulate biological evolution and natural selection among a set of possible solutions, and can produce an optimal or near optimal solution. SLEUTH uses a bounded five dimensional search where the model metrics can direct the search. The five dimensions are the integer values of the five model behavior parameters, and the metrics reduce to the OSM.

The application of GA to SLEUTH was first achieved by Goldstein (2004) using both elitism and tournament selection, and combining gene competition strategies (stratified, partial random, and random). Crossover employed both uniform and self-crossover, and mutation used a 10% randomization. The approach was tested for Sioux Falls, South Dakota over 200 generations, with 18 chromosomes in each run, but for only one Monte Carlo iteration, with the calibration repeated 10 times. Results showed that 70% of the chromosomes outperformed brute force yet used one fifth as much CPU time, giving better goodness of fit measures. Nevertheless, there was evidence that the GA became stuck in local maxima, and some optimization ambiguity as the work predated the OSM, and so compared different metrics. While the GA was only simulated (separately generating the parameters, that were fed to independent runs across 10 computers), Goldstein did explore the consequences of sub-optimal calibrations for model forecasting, but not which gene selection, cross-over and mutation strategies worked best. Our approach first tested possible strategies, and then hard coded a single strategy into the SLEUTH source code driver module.

The GA for SLEUTH calibration was designed based on Goldstein's findings and prior GA research including choices on encoding, fitness evaluation, crossover, mutation, and survival selection (Eiben and Smith 2003). The model provides a natural encoding, each gene is represented as a set of five integer coefficients in the range {0,0,0,0,0} to {100, 100,100,100,100}. Each coefficient represents a separate piece of genetic material for a specific gene, with all five combined composing the entire composition of a gene. When running the model with the five coefficients, the OSM metric provided creates a natural fitness evaluation for an individual gene. Crossover, the process of combining existing genes to create new genes, takes a subset of the coefficients from one gene and

combines them with the opposite subset from the other, which was simpler than Goldstein's method (2004). This was performed by using a random number between 0 and 4 and using that value to decide how many elements from the first parent are used to create the offspring. The remaining elements were provided by the second parent. A second offspring was produced from the opposite elements that created the first offspring. Parents were selected using tournament selection, a random subset of the population is chosen and two selected using probabilities proportional to fitness. Mutation replaces coefficients within a gene with a random value at the mutation rate frequency to maintain diversity. The mutation rate was provided as an input to the genetic algorithm and can be tuned based on the performance of the model. Lastly, survival selection is the method for selecting a subset of the population and its offspring for the next generation. Each generation replaces the weakest genes in the old population with the strongest of the offspring, until at least half of the population is replaced and there are no old population genes that are weaker than any remaining offspring. Elitism prevents the fitness from regressing during the calibration.

The GA was first tested to determine population size and mutation rate. Testing used 2,000 iterations through the model per run of the GA. Using the Demo_City test data provided with SLEUTH showed that neither a low nor high mutation rate was ideal, but within the range 0.10 to 0.16 was satisfactory (Figure 1). Results showed that population sizes between 15 and 30 were good choices (Figure 2). A population size of 25 and a mutation rate of 0.16 were chosen. While 15 showed the strongest fitness, a population size must be sufficiently large to maintain genetic diversity.

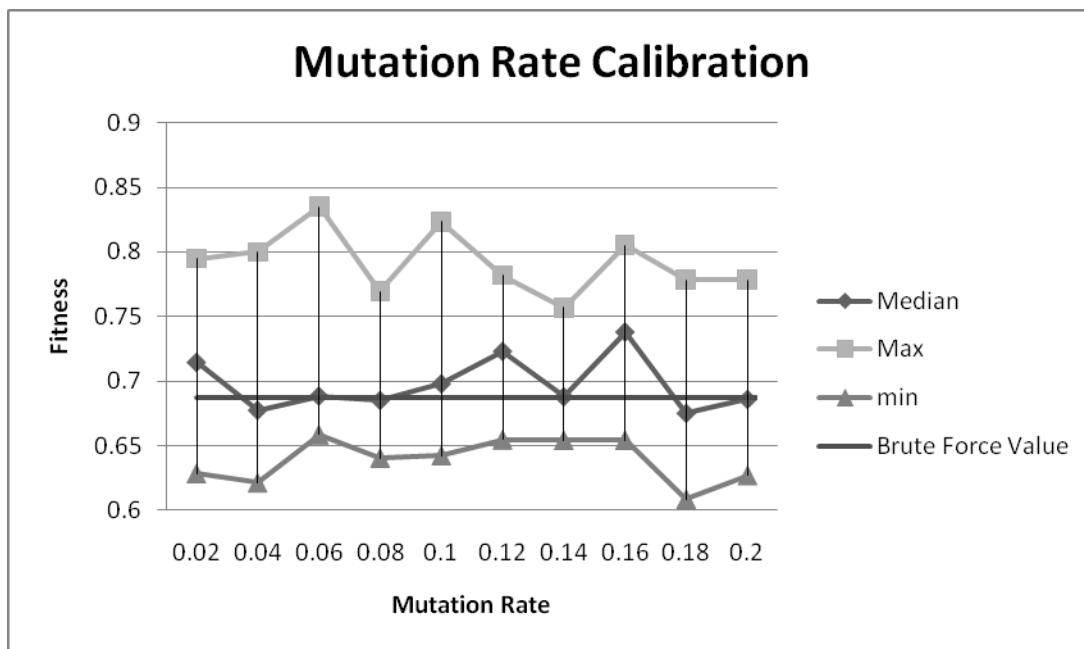


Figure 1. Results of GA test: Mutation Rate

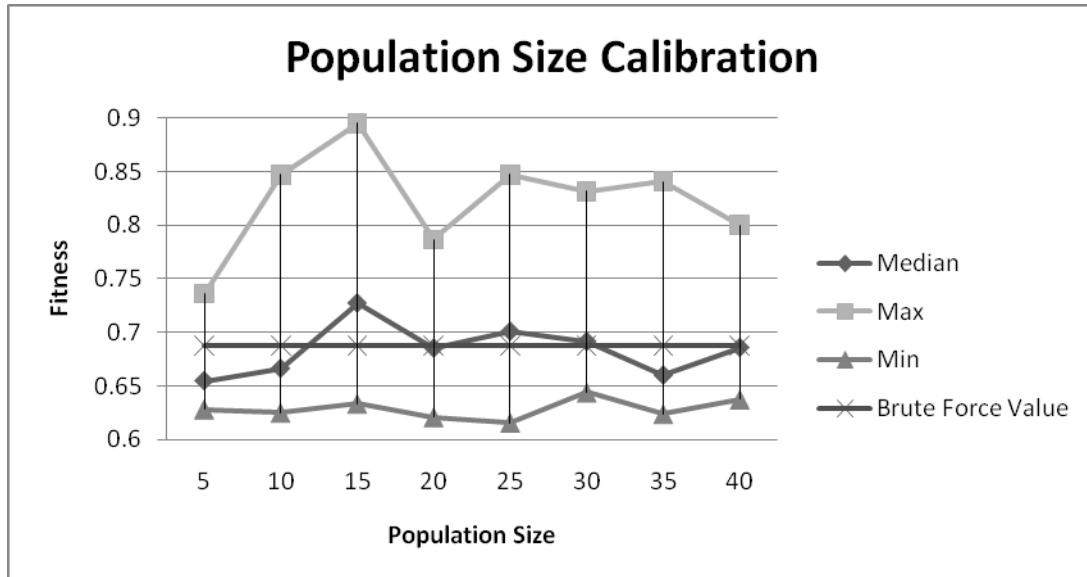


Figure 2. Results of GA test: Population Size

4. Results

The GA was used in SLEUTH and the results compared to the brute force method as applied to Demo_City. OSM values obtained were similar to those achieved in other SLEUTH applications (Table 1).

Statistic	Fitness (OSM)	% Improvement
Mean	0.705013	3%
Median	0.697704	2%
Standard Deviation	0.051764	--
Minimum	0.620926	-10%
Maximum	0.870902	27%
Brute Force Calibration	0.687381	--

Table 1: Genetic Algorithm Calibration Results

The GA on average performed slightly better than brute force. Due to the stochastic nature of a GA, there were rounds where it performed up to 10% worse or 27% better than brute force. As with Goldstein's test, the model was calibrated using a single Monte Carlo iteration to reduce computation time and allow for rapid evolution in the GA.

5. Conclusions

Results showed the GA can maintain or improve the fit of SLEUTH. While the median solution was a small improvement, performance boost varied from -10% to 27%. Yet the real value of GA is in reducing computation time, where it outperforms brute force

calibration by a factor of 5, without subjective input. This speed-up was also achieved by Goldstein (2004), and may be further improvable by experiment. We capped the GA at 2,000 generations, while the brute force required a minimum of ~10,000 iterations. On an Intel XEON 5570 CPU one run of the GA was completed in ~30 minutes and eight runs could be performed simultaneously per CPU without taxing the server. This would allow model calibration in hours, compared to weeks with brute force. Such a saving would permit calibration sensitivity tests not feasible otherwise. The SLEUTH code used in this research was posted to the SourceForge open source site (<https://sourceforge.net/projects/sleuth-ga/>).

Future improvements can be made to the GA through algorithm optimization and parallelization. These would increase the efficiency of the GA further improving speed, and reducing calibration to minutes. Such times would overcome one of the last remaining obstacles to SLEUTH's application in urban planning and land management (Clarke, 2008). Furthermore, it is a good example of geocomputation, where computer science optimization methods (GA) meet simulation modelling in geography.

6. References

- Clarke, K. C., Hoppen, S. and L. Gaydos. 1997. A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. *Environment and Planning B: Planning and Design*, vol. 24, pp. 247-261.
- Clarke, K. C., Gazulis, N., Dietzel, C. K. and Goldstein, N. C. 2007. A decade of SLEUTHing: Lessons learned from applications of a cellular automaton land use change model. Chapter 16 in Fisher, P. (ed) *Classics from IJGIS. Twenty Years of the International Journal of Geographical Information Systems and Science*. Taylor and Francis, CRC. Boca Raton, FL. pp. 413-425.
- Clarke, K.C. 2008. A decade of cellular urban modeling with SLEUTH: Unresolved issues and problems, Ch. 3 in *Planning Support Systems for Cities and Regions* (Ed. Brail, R. K., Lincoln Institute of Land Policy, Cambridge, MA, pp 47-60.
- Dietzel, C. and Clarke, K.C. 2007, Toward Optimal Calibration of the SLEUTH Land Use Change Model. *Transactions in GIS* 11(1): 29-45.
- Eiben, A. E. and Smith, J. E. 2003, *Introduction to Evolutionary Computing*, Springer-Verlag, Berlin.
- Goldstein, N. C. 2004. *Brains vs. Brawn: Comparative strategies for the calibration of a cellular automata-based urban growth model*. Chapter 18 in Atkinson, P., Foody, G., Darby, S., and Wu, F. (eds) *GeoDynamics*. Boca Raton, FL, CRC Press.
- Jantz, C. A., Goetz, S. J., Donato, D. and Claggett, P. 2010, Designing and implementing a regional urban modeling system using the SLEUTH cellular urban model. *Computers, Environment and Urban Systems*, 34, 1-16.
- Silva, E. A. and Clarke, K., (2005) Complexity, emergence and cellular urban models: Lessons learned from applying SLEUTH to two Portuguese metropolitan areas. *European Planning Studies*, vol. 13, no. 1, pp. 93-115.
- Torrens, P. M. & O' Sullivan, D. 2001. Cellular automata and urban simulation: where do we go from here? *Environment and Planning B*, 28, 163-168.
- US Geological Survey. 2007, Project Gigalopolis: Urban and Land Cover Modeling. <http://www.ncgia.ucsb.edu/projects/gig/index.html>

Calibration of a cellular automata model with the particle swarm algorithm

Nuno Norte Pinto¹, António Pais Antunes², Josep Roca Cladera³

^{1,2}Department of Civil Engineering, University of Coimbra
R. Luis Reis Santos, Pólo II da Universidade, 3030-788 Coimbra

Telephone: (+351)239797106
Fax: (+351)239797147
Email¹: npinto@dec.uc.pt
Email²: antunes@dec.uc.pt

³Center for Land Policy and Valuation, Technical University of Catalonia
Av. Diagonal, 649, 4^a planta, 08028 Barcelona, Spain
Telephone: (+34)934016396
Fax: (+34)933330960
Email: josep.roca@upd.edu

1. Introduction

Cellular automata (CA) models have long been applied to simulate the evolution of urban areas. The large majority of CA models reported in the literature make use of regular cells derived from remote sensed images to represent land use and the use of irregular cells is scarce (Moreno et al., 2008, Stevens and Dragicevic, 2007). However, regular cells are not directly connected to the information that underlies the drivers of land use change – population, employment, or built up area indicators. We proposed a CA model that operates over a cell structure derived from irregular cells obtained from census blocks, which hold reliable data and can be easily classified for their land use (Norte Pinto and Pais Antunes, 2010).

Calibration plays a critical role in modelling because it connects reality to model representation. CA model calibration has been a subject of different approaches using different types of procedures, from sensitivity analysis to optimization-based methods. SLEUTH (Silva and Clarke, 2002) is uses both visual calibration and a brute force computational procedure to compare model and reference data. Li and Yeh (2001) coupled a CA model with an artificial neural network to calibrate it. Barredo et al. (2003) used basic sensitivity analysis to calibrate the weighting parameters for the spatial interactions between land uses.

2. Cellular automata model

The CA model has a simple structure that derives from the classical formulation of CA with the consideration of constrained land use demand, following the concept introduced by White and Engelen (1993). The model operates over an irregular cellular fabric obtained from census blocks. Cell states are classified into a finite set of aggregated classes of land use. Land use interactions take place within a variable neighborhood which distance value is determined through model calibration. Transition rules intend to incorporate planning regulations and simulate land use change based on a composite transition potential that takes into account cell accessibility, land use suitability, and

neighborhood interactions within the cell neighborhood, calculated by the following expression

$$P_{i,s} = (v_P \times S_{i,s} + \chi_P \times A_i + \theta_P \times N_{i,s}) \times \xi, \forall i \in C, s \in S$$

where, for each cell i from the set of cells C , and for each state s from the set of states S , $P_{i,s}$ is the transition potential for state s of cell i , $S_{i,s}$ is the land use suitability value for state s of cell i , A_i is the accessibility value of cell i , $N_{i,s}$ is the neighborhood effect for state s of cell i considering its neighborhood V_i , v_P is the calibration parameter for land use suitability, χ_P is the calibration parameter for accessibility, θ_P is the calibration parameter for the neighborhood effect, and ξ is the stochastic parameter. The model has 30 more calibration parameters which define the linear relationships of neighborhood effect interactions between each pair of land uses, generically depicted in Figure 1(a) for attraction and Figure 1(b) for repulsion. The time step can be defined by the user. Land use demand is determined through the evolution of population and employment densities over time. The flowchart for the CA model is depicted in Figure 2. Further details on the structure of the model can be found in Norte Pinto and Pais Antunes (2010).

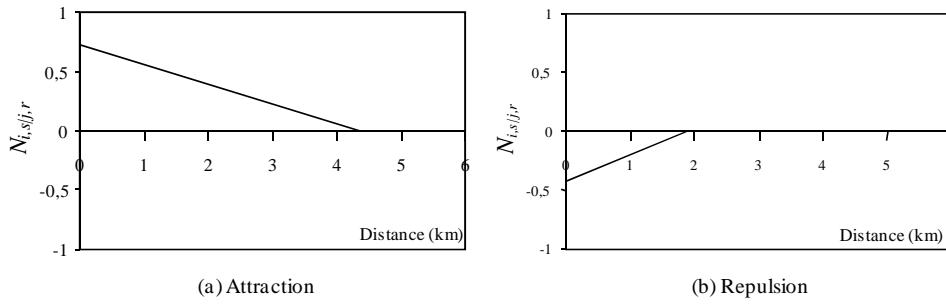


Figure 1. Generic neighborhood effect relationships

3. Calibration with particle swarm

The high number of calibration parameters indicates the use of an optimization procedure to ensure a good search of the solution space. The calibration of the CA model is processed through an optimization procedure that uses a fitness measure based on *kappa* index from contingency matrixes (Couto, 2003). We used a modified version of the traditional *kappa* (named k_{Mod}) to avoid the distortion that would have been produced if states that cannot take part in the urban dynamics – for example, agricultural or ecological reserve land – were considered. The inclusion of cells in this state would be misleading by producing a larger – though meaningless – agreement between simulation and reference maps.

The optimization algorithm chosen was the particle swarm (PS), which roots are in the simulation of social behaviors, in the study of the synchronized movement of bird flocks and fish schools (for further details please see Kennedy, 1997, and Parsopoulos and Vrahatis, 2002). This algorithm is suitable for dealing with a high number of dimensions (our calibration parameters) because it has a simple formulation which ensures that the complex interdependences between the parameters are taken into account in the calibration process. The algorithm makes use of a swarm of p particles (from a few to

traditionally up to 120, but with no upper limit) will fly through the search space during n iterations. The larger the swarm is, the better the search space is searched. Each particle has D dimensions: in our CA model each calibration parameter is represented by a PS dimension. Hence, there will be 48 dimensions for each particle. The algorithm retains the position and the velocity of each particle in every iteration, calculating their new values considering the group leader and their individual best positions. The flowchart for the PS algorithm is depicted in Figure 2. Note that CA are an embedded process that is called as many times as the number of PS iterations multiplied by the number of particles.

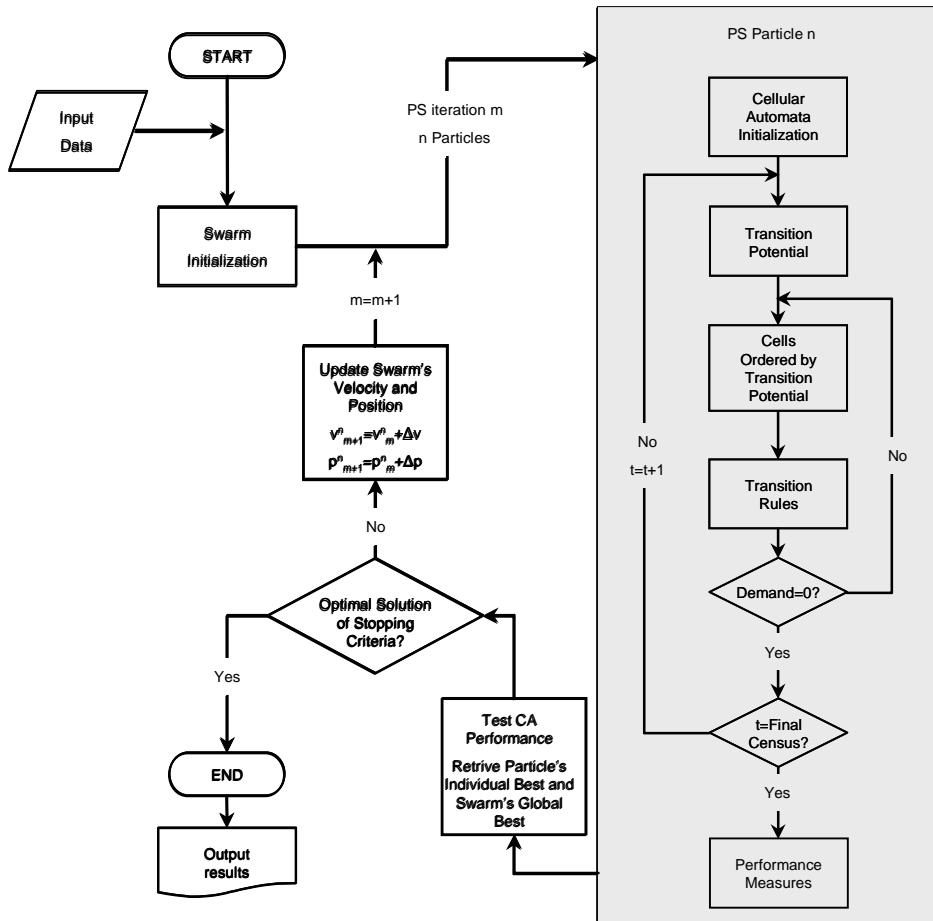


Figure 2. CA model (grey) and PS algorithm flowchart

4. Model results

The model was tested using a set of twenty test instances generated to simulate plausible spatial structures. These test instances have two reference land use maps (initial and final) for two moments in time, comprising information about population, employment and accessibility considering a road network. Three examples are depicted in Figure 3. Land use was classified with a set of aggregate cell states: urban low density (UL) and urban high density (UH), non-urbanized urban areas (XU); industry (I), non-urbanized industrial areas (XI); and areas where construction is highly restricted (R).

Global k_{Mod} results for the entire set of problems are depicted in Figure 4. These results can be considered good for a simulation process: 50 percent of the problems achieved a

k_{Mod} around 0.800 or higher and 75 percent of them exceeded 0.750. Figure 4 also presents the variation of the absolute κ measure for the set of test problems. For 65 percent of the problems, the agreement exceeded 0.900 and 95 percent exceeded 0.850. Overall accuracy for the k_{Mod} measure also exceeded 0.850 for 75 percent of the cases. These values are commonly accepted as very good agreement between modeled and reference situations (Barredo et al., 2003).

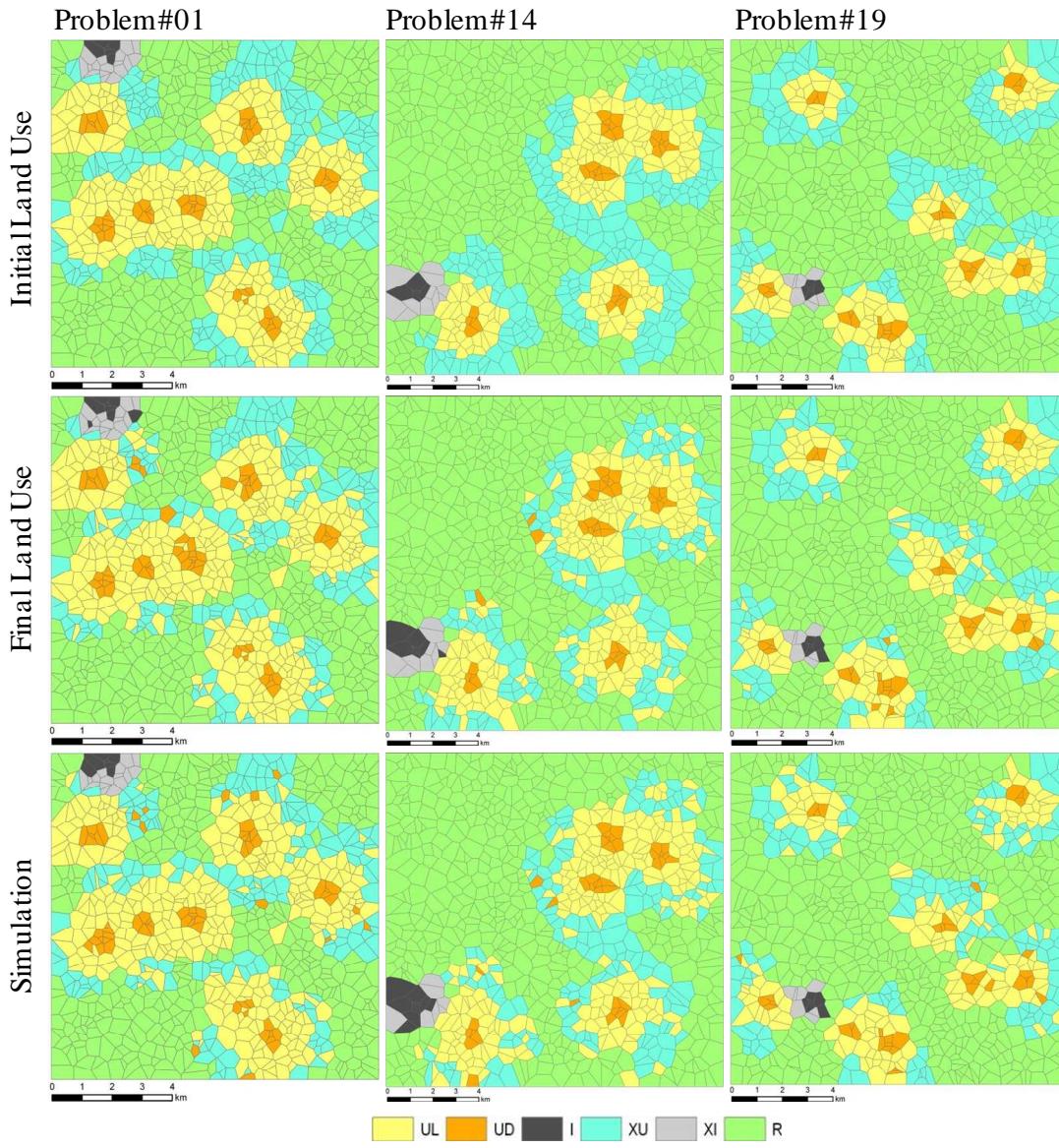


Figure 3. Three examples of test instances.

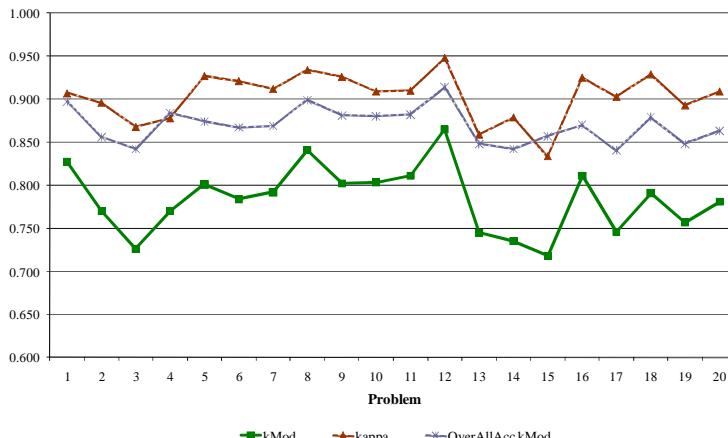


Figure 4. Global k_{Mod} and κ results for the set of twenty test problems.

5. Concluding remarks

The results obtained for the set of test instances show that the use of the PS algorithm ensures an efficient search of good sets of calibration parameters for the CA model. The average value of the fitness measure k_{Mod} is high and is equal or higher than the values founded in the literature for other CA models. Current developments of our CA models – focusing on a multi-scale approach – also use the PS optimization for model calibration.

6. Acknowledgments

Nuno Pinto wishes to acknowledge the support received from Fundação para a Ciência e a Tecnologia under grant SFRH/BD/37465/2007.

7. References

- Barredo, J., Kasanko, M., McCormick, N. & Lavalle, C. (2003) Modelling dynamic spatial processes: simulation of urban future scenarios through cellular automata. *Landscape and Urban Planning*, 64(3), 145-160.
- Couto, P. (2003) Assessing the accuracy of spatial simulation models. *Ecological Modelling*, 167(1-2), 181–198.
- Kennedy, J. (1997) The particle swarm: Social adaptation of knowledge. *Proceedings of the 1997 IEEE International Conference on Evolutionary Computation (ICEC '97)*, Indianapolis, IN, 303-308.
- Li, X. & Yeh, A. G. O. (2001) Calibration of cellular automata by using neural networks for the simulation of complex urban systems. *Environment and Planning A*, 33(8), 1445-1462.
- Moreno, N., Ménard, A. & Marceau, D. J. (2008) VecGCA: a vector-based geographic cellular automata model allowing geometric transformations of objects. *Environment and Planning B: Planning and Design*, 35(4), 647-665.
- Norte Pinto, N. & Pais Antunes, A. (2010) A cellular automata model based on irregular cells: application to small urban areas. *Environment and Planning B: Planning and Design*, 37(6), 1095-1114.
- Parsopoulos, K. E. & Vrahatis, M. N. (2002) Recent approaches to global optimization problems through Particle Swarm Optimization. *Natural Computing*, 1, 235–306.
- Silva, E. & Clarke, K. C. (2002) Calibration of the SLEUTH urban growth model for Lisbon and Porto, Portugal. *Computers, Environment and Urban Systems*, 26(6), 525-552.
- Stevens, D. & Dragicevic, S. (2007) A GIS-based irregular cellular automata model of land-use change. *Environment and Planning B: Planning and Design*, 34(4), 708-724.
- White, R. & Engelen, G. (1993) Cellular automata and fractal urban form: a cellular modelling approach to the evolution of urban land-use patterns. *Environment and Planning A*, 25(8), 1175-1199.

A macroscale cellular automata model for simulating urban change in regional urban systems

Nuno Norte Pinto¹, António Pais Antunes², Josep Roca Cladera³

^{1,2}Department of Civil Engineering, University of Coimbra
R. Luis Reis Santos, Pólo II da Universidade, 3030-788 Coimbra, Portugal

Telephone: (+351)239797106
Fax: (+351)239797147
Email¹: npinto@dec.uc.pt
Email²: antunes@dec.uc.pt

³Center for Land Policy and Valuation, Technical University of Catalonia
Av. Diagonal, 649, 4^a planta, 08028 Barcelona, Spain
Telephone: (+34)934016396
Fax: (+34)933330960
Email: josep.roca@upd.edu

1. Introduction

Cellular Automata (CA) models are among the most popular models for simulating spatial change and they have been developed and applied intensively during the past two decades. Two main features made CA interesting for urban studies, ever since they were introduced by Waldo Tobler in the late 1970s (Tobler, 1979): first, their inherent spatiality which suits the simulation of a wide range of geographic phenomena; second, the possibility of simulating complex patterns of, for example, land use starting from a simple conceptual framework that includes the definition of a cell space (form), a neighborhood (interaction), and a finite set of transition rules (behaviors) applied to a finite set of cell states (land uses). This conjugation of form and function make CA models suitable for capturing the contribution of different phenomena to the complex processes of urban change.

These models are commonly used to simulate land use change at a regional or metropolitan level considering land use dynamics at a local level (Barredo and Demicheli, 2003, Silva and Clarke, 2005). They consider increasingly smaller cells, making use of the high resolution of today's remotely sensed images to capture many interactions that occur at a very large scale. Regular cells are used at the local scale (pixels) and at a regional scale, as aggregations of smaller cells (Van Vliet et al., 2009).

We address these issues of scale and cell form by proposing a macroscale CA model that tries to capture aggregated land use change at a regional level. We use administrative units – municipalities or similar units, varying with the national context – as irregular cells to simulate land use change considering population and employment growth and accessibility measures at a regional scale. The use of irregular cells, regardless of the scale, is scarce in the literature (Stevens and Dragicevic, 2007, Moreno et al., 2008). It ensures a good link between form and reliable data, an approach that has been successfully applied at the local scale (Norte Pinto and Pais Antunes, 2010).

Scale has been debated over the years. The evolution of computation allowed researchers to downscale from the typical large scale models of the 1950s and 1960s to the high resolution models of our decade. The debate over modeling scale started with the

famous *Requiem for large-scale models* (Lee, 1973), and continued over the years, with a new moment in the mid 1990s when again the issue was brought to the agenda (Lee, 1994, Klosterman, 1994). Recently, there is again a new interest on scale, focusing also on CA models (Ménard and Marceau, 2005, Benesson, 2007, White, 2007, Briassoulis, 2008, Verburg et al., 2008).

2. Macroscale CA model

The model uses municipalities (or similar administrative units) as cells. Cell states are classified into a finite set of artificial land area, accounted as a percentage of the total cell area. Land use interactions take place within a variable neighborhood which distance value is determined through model calibration. Transition rules intend to simulate spatial interaction based on a transition potential functional that depends on the population, the employment, and a function of distance over the road network, calculated by the following expression:

$$V_i = \frac{\alpha_p \times P_i \times E_j}{d_{ij}^\beta}, \forall i \in C, j \in C \quad (1)$$

where, for each cell i from the set of cells C , V_i is the transition potential for cell i , P_i is the number of residents in cell i , E_i is the number of registered employees in cell i , d_{ij} is the distance between cells i and j (from the set of cells C) measured by the road network, α_p is a calibration parameter and β is the accessibility calibration parameter. In each time step, cells are selected by the model for urbanization though a measure of its relative probability (taking into consideration all cells) regarding the transition potential value, calculated through an application of the *logit* model as follows:

$$U_i = \frac{e^{\alpha_L \times V_i}}{\sum_j (e^{\alpha_L \times V_j})}, \forall i \in C, j \in C \quad (2)$$

where, for each cell i from the set of cells C , U_i is the relative probability value of cell i , V_i is the transition potential for cell i , and α_L is the calibration parameter of the *logit* model.

The model is calibrated through an optimization procedure based on the particle swarm (PS) algorithm that uses as fitness measure the *kappa* index for contingency matrixes. PS makes use of a swarm of p particles that will fly through the solution space during n iterations. Each particle has D dimensions: in our CA model each calibration parameter is represented by a PS dimension. The algorithm retains the position and the velocity of each particle in every iteration, calculating their new values considering the group leader and their individual best positions. Note that CA are an embedded process that is called as many times as the number of PS iterations multiplied by the number of particles.

3. Application to the Metropolitan Area of Barcelona

The Metropolitan Area of Barcelona (MAB) is composed by 164 municipalities which vary considerably in area, population, and employment. The city of Barcelona heads a complex set of mid-size and small urban systems which group urban areas and their hinterlands with their own functional relationships.

The model was applied to MAB in order to simulate the allocation of urbanized land over the municipalities, considering an aggregate value of population and employment density as limits for land demand. The model was calibrated using data from the censuses of 1991 and 2001 for population and employment and using aggregated land use information derived from Corine Land Cover for the same years. The model reached a value of *kappa* of 0.427 which represents a moderate agreement.

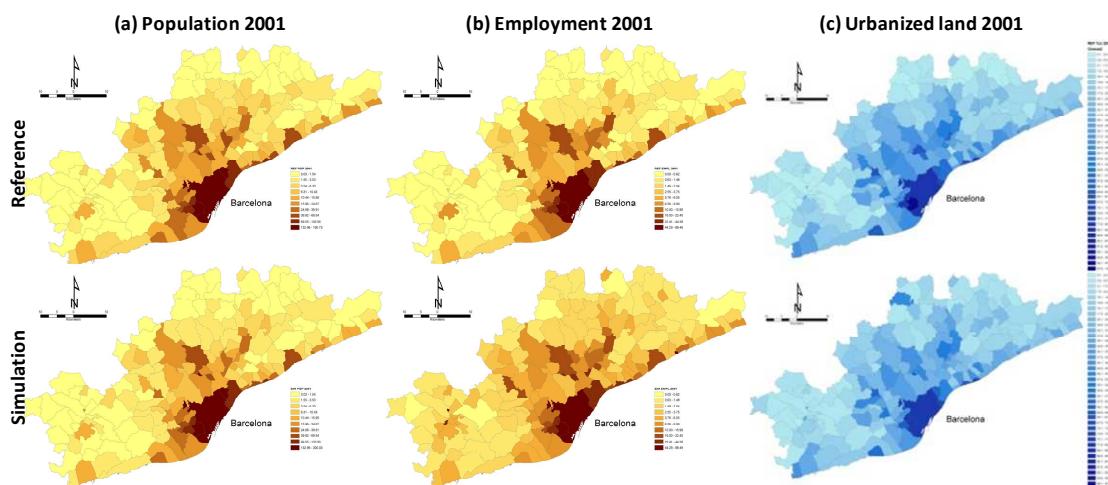


Figure 1. Model results for the MAB for population, employment, and urbanized areas.

4. Concluding remarks

This macroscale CA model is part of an integrated multiscale CA model that aims to capture different phenomena that occur at different spatial and time scales. The macroscale model aims to simulate the evolution of land use demand by modeling the areas of urbanized land at the municipality level as a function of the location of population and employment, considering accessibility. The values of urbanized land will be considered as land use demand at the microscale, and will be used as a constraint to a more traditional, local scale CA model.

5. Acknowledgments

Nuno Pinto acknowledges the support received from Fundação para a Ciência e a Tecnologia under grant SFRH/BD/37465/2007.

6. References

- Barredo, J. I. & Demicheli, L. (2003) Urban sustainability in developing countries' megacities: modelling and predicting future urban growth in Lagos. *Cities*, 20(5), 297-310.

- Benenson, I. (2007) Warning! The scale of land-use CA is changing! *Computers, Environment and Urban Systems*, 31(2), 107-113.
- Briassoulis, H. (2008) Land-use policy and planning, theorizing, and modeling: lost in translation, found in complexity? *Environment and Planning B: Planning and Design*, 35(1), 16-33.
- Klosterman, R. (1994) Large-Scale urban models - Retrospect and prospect. *Journal of the American Planning Association*, 60(1), 3-6.
- Lee, D. (1973) Requiem for large-scale models. *Journal of the American Planning Association*, 39(3), 163-178.
- Lee, D. (1994) Retrospective on large-scale urban models. *Journal of the American Planning Association*, 60(1), 35-40.
- Ménard, A. & Marceau, D. J. (2005) Exploration of spatial scale sensitivity in geographical cellular automata. *Environment and Planning B: Planning and Design*, 32(5), 693-714.
- Moreno, N., Ménard, A. & Marceau, D. J. (2008) VecGCA: a vector-based geographic cellular automata model allowing geometric transformations of objects. *Environment and Planning B: Planning and Design*, 35(4), 647-665.
- Norte Pinto, N. & Pais Antunes, A. (2010) A cellular automata model based on irregular cells: application to small urban areas. *Environment and Planning B: Planning and Design*, 37(6), 1095-1114.
- Silva, E. & Clarke, K. C. (2005) Complexity, emergence and cellular urban models: lessons learned from applying SLEUTH to two Portuguese metropolitan areas. *European Planning Studies*, 13(1), 93-116.
- Stevens, D. & Dragicevic, S. (2007) A GIS-based irregular cellular automata model of land-use change. *Environment and Planning B: Planning and Design*, 34(4), 708-724.
- Tobler, W. (1979) Cellular geography. IN GALE, S. & OLSSON, G. (Eds.) *Philosophy in Geography*. Boston, D. Reidel.
- van Vliet, J., White, R. & Dragicevic, S. (2009) Modeling urban growth using a variable grid cellular automaton. *Computers Environment and Urban Systems*, 33(1), 35-43.
- Verburg, P., Eickhout, B. & van Meijl, H. (2008) A multi-scale, multi-model approach for analyzing the future dynamics of European land use. *The Annals of Regional Science*, 42(1), 57-77.
- White, R. (2007) Multi-scale modelling with variable grid CA. IN BAVAUD, F. & MAGER, C. (Eds.) 15th European Conference on Theoretical Quantitative Geography. Montreux, Switzerland, Institute of Geography of the University of Lausanne.

Simulation of Cholera Diffusion to compare transmission mechanisms

Ellen-Wien Augustijn-Beckers¹, Juliana Useya¹, Raul Zurita-Milla¹, Frank Osei²

¹University of Twente, Faculty of Geo-Information Science and Earth Observation (ITC), Hengelosestraat 99, 7500 AA Enschede, The Netherlands

Telephone: +31 (0)53 4874 414

Fax: +31 (0)53 4874335

Email: augustijn@itc.nl

²Kwame Nkrumah University of Science and Technology (KNUST), Department of Geomatic Engineering, Kumasi, Ghana

1. Introduction

Since the initial transmission mechanism of Cholera was revealed by John Snow in 1854, the cause and spread of this disease has been under continuous research. Snow's study showed how disease incidences can be linked to a source based on the spatial distribution of the patients. However, Snow's work did not address the question of diffusion mechanisms. The predominant transmission mechanism of Cholera is via the fecal-oral route but in recent years several scientists have pointed toward a number of other transmission mechanisms that might contribute to the prevalence of the disease.

Cholera risk factors vary and stem from multiple transmissions including interactions between human hosts, pathogen and environment leading to person to person transmission (secondary transmission) and transmission via the environment (primary transmission) (Hartley et al., 2006). It is possible for the toxigenic *V. cholerae* to survive in surface water for up to several years (Codeco, 2001). Especially oceans and brackish water seem to function as a long term biotic reservoir for cholera (Emch et al., 2008). Driving factors for the ecology of *V. cholerae* are meteorological and climate variation. Where environmental forcing plays an important role at the macro level, secondary transmission is more related to local environmental variation.

Examples of local variation that influences the spread of cholera are water sources for household activities (pipe – well water), food control (seafood, fish and contamination during preparation) (Said, 2006) and sanitation (Emch et al., 2008). Fotedar (2001) provided evidence that houseflies (*Musca domestica*) are able to carry *V. cholerae*. Osei & Duker (2008) related cholera transmission to the mechanisms of filth breeding flies and flood water contamination. Spatial dependency of cholera infection on the proximity to and density of refuse dumps was shown by Osei et al. (2010) indicating that runoff from dump sites carry fecal materials to local rivers, creating a pathway for fecal contamination of surface water. Hartley et al. (2006) investigated the relative importance of the transmission factors and found a dependency on sanitation, population density and hygiene. We continue this work by investigating the relative importance of micro level transmission mechanisms by means of an agent-based simulation model.

There are relatively few mathematical Cholera models, perhaps because of the complex transmission mechanisms. A model was developed by Codeco (2001), who extended an existing model by Capasso for an Italian cholera outbreak. This model allows for long term dynamics incorporating an environmental reservoir of *V. cholerae*. A line of spatially explicit mathematical models was developed based on hydrology-driven cholera spreading (Bertuzzo et al., 2008, Bertuzzo et al., 2009, Righetto et al., 2010), and an age structured model was developed by Agheksanterian & Gobbert (2007).

2. Cholera model

The model presented in this research is a geographically explicit agent-based Cholera simulation. It is a micro scale, hydrology-driven model that differs from already existing ones in that it:

- Includes the spread of Cholera from dumpsites by the housefly (*M. domestica*)
- Includes runoff from dumpsites as a pathway of bacteria and feces to rivers
- Includes human to human transmission of cholera
- Is based on a synthetic population representing age categories, income levels and other population dynamics like hygiene levels and access to pipe water.

The proposed model consists of four different sub-models: (i) a hydrological model for the transport of the *V. cholerae* pathogen (ii) an epidemic model (iii) a house fly model for modeling flies as disease carriers (iv) a human interaction model.

2.1 Hydrological sub-model

The hydrological model consists of three elements, an elevation raster, line elements representing the river branches and rainfall particles. Rainfall particles will flow downhill according to the elevation surface and can be transferred into carriers of feces or carriers of feces with pathogen. The model assumes constant flow of water along the river branches. Changes in river water volume and speed of flow are not taken into account. Growth rate of free-living bacteria (in water) is normally negative (Bertuzzo et al., 2008). Because of the small area included in the simulation no “bacterial mortality” is implemented.

2.2 Epidemic sub-model

The model is based on the cholera transmission model from Hartley (2006) including hyper-infectivity. Hyper-infectivity is the fact that *V. cholerae* when passed through the gastrointestinal tract (via a human being) transfers into a short-lived hyper-infectious state (Hartley et al., 2006). The existence of this hyper-infectivity is associated with the explosive nature of some cholera outbreaks. Chance of developing the disease after exposure to hyper infective *V. cholerae* is very high. After recovery people become immune and this immunity lasts for at least two years (Koelle et al., 2005). For the time span of this simulation the immunity is permanent and no waning is considered. Currently the severity of the disease is not modeled although some cases are known to be asymptomatic.

2.3 House Fly sub-model

House Fly density maps were generated around the locations known to be exposed to human excreta (refuse dumps and rivers). For this model, density layers were regarded to be static. Assumption is made that flies are able to carry *V. cholerae* from open dumpsites to surrounding areas within a critical buffer distance of 500 meters (Osei and Duker, 2008). Within this buffer distance transmission can occur.

2.4 Human interaction sub-model

Agent unit is the individual person. Individuals are grouped into families. Prior to the start of the simulation families are distributed over houses, with multiple families living in a single building. Important attributes of families include the income level and the level of hygiene. Composition of families is based on Census data. Individuals are age specific and their behavior is based on age grouping. Individual behavior includes all activities that can lead to disease exposure including drinking/eating, caring for diseased relatives, dumping of feces and playing at dumpsites.

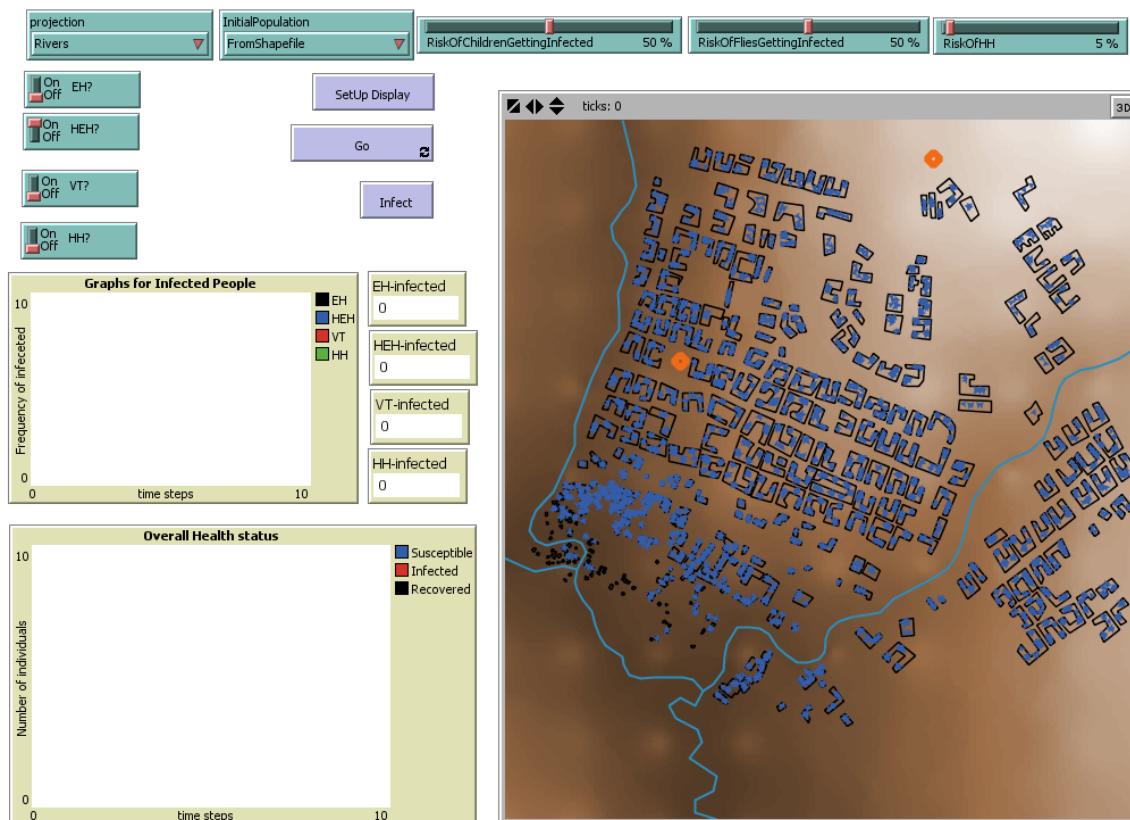


Figure 1. The Interface of the model

3. Case study

Study area is Kumasi, capital city of Ashanti Region located in south central Ghana. Kumasi has a population of approximately 3.5 million. The study area is located in the north-eastern part of the city. A severe outbreak of cholera occurred in this area in 2005 (data: Kumasi Metropolitan Disease Control Unit).

In the paper we will present the conceptual design and the initial findings of the model. Findings include the comparison of different transmission mechanisms. Importance of an agent-based cholera model is that the heterogeneity of the population is accounted for and that experiments can be conducted with intervention and changes in behavior of population. In future, we will continue to include behavioral changes of agents into this model.

4. Acknowledgements

Base data for this project was partially obtained from data obtained via the Planet Action Program (Spot Image).

5. References

- Agheksanterian, A. & Gobbert, M. K. 2007. Modeling the spread of epidemic cholera: an age-structured model. Department of Mathematics and statistics, University of Maryland, Baltimore County.
- Bertuzzo, E., Azaele, S., Maritan, A., Gatto, M., Rodriguez-Iturbe, I. & Rinaldo, A. 2008, On the space-time evolution of a cholera epidemic. *Water Resources Research*, 44.
- Bertuzzo, E., Casagrandi, R., Gatto, M., Rodriguez-Iturbe, I. & Rinaldo, A. 2009, On spatially explicit models of Cholera epidemics. *Journal of the Royal Society Interface*, 7, 321-333.
- Codeco, C. 2001, Endemic and epidemic dynamics of cholera: the role of the aquatic reservoir. *BMC Infectious Diseases*, 1, 1.
- Emch, M., Feldacker, C., Islam, M. S. & Ali, M. 2008, Seasonality of cholera from 1974 to 2005: a review of global patterns. *International Journal of Health Geographics*, 7, 1-13.
- Fotedar, R. 2001, Vector potential of houseflies (*Musca domestica*) in the transmission of *Vibrio cholerae* in India. *Acta Tropica*, 78, 31-34.
- Hartley, D. M., Morris, J. G. & Smits, D. L. 2006, Hyperinfectivity: A Critical Element in the Ability of *V. cholerae* to cause epidemics? *PLoS Med*, 3, 63-69.
- Koelle, K., Rodo, X., Pascual, M., Yunus, M. & Mostafa, G. 2005, Refractory periods and climate forcing in cholera dynamics. *Nature*, 436, 696-700.
- Osei, F. & Duker, A. 2008, Spatial dependency of *V. cholera* prevalence on open space refuse dumps in Kumasi, Ghana: a spatial statistical modelling. *International Journal of Health Geographics*, 7, 62.
- Osei, F. B., Duker, A. A., Augustijn, E.-W. & Stein, A. 2010, Spatial dependency of cholera prevalence on potential cholera reservoirs in an urban area, Kumasi, Ghana. *International Journal of Applied Earth Observation and Geoinformation*, 12 331-339.
- Righetto, L., Bertuzzo, E., Casagrandi, R., Gatto, M., Rodriguez-Iturbe, I. & Rinaldo, A. 2010, Modelling human movement in cholera spreading along fluvial systems. *Ecohydrology*.
- Said, M. D. 2006, *Epidemic cholera in KwaZulu-Natal: The role of the natural and social environment*. PhD, University of Pretoria.

Modified Navigation Algorithms in Agent-Based Modelling for Fire Evacuation Simulation

Tyng-Rong Roan¹, Muki Haklay¹, Claire Ellul¹

¹Department of Civil, Environmental and Geomatic Engineering, University College London
Gower Street, London WC1E 6BT, United Kingdom
Email: t.roan@ucl.ac.uk

1. Introduction

Hazardous events which threaten people's lives force an immediate movement of people wanting to escape from a dangerous area. In their review of man-made and natural disasters, Wolshon *et al.* (2005) listed a number of hazards requiring evacuation and pointed out that some evacuations could only be carried out after disasters occur. Therefore, people need to run through evacuation drills to learn evacuation skills and to ensure they are familiar with the environment. However, evacuation drills cannot realistically represent a real emergency situation and people may be injured during the practices. To overcome these issues, evacuation models are useful for simulating these hazardous situations. Models remove the risk to human safety that may be present during drills, and generate efficient evacuation routes for emergency plans. One of the common modelling approaches is agent-based modelling; an agent-based model is a computational model using virtual agents to simulate independent actions, social interactions, adaptive processes, and goal-directed navigations. This type of evacuation model was presented to study inter-relationships between individuals and groups' behaviours (Musse and Thalmann 1997), steering behaviour (Reynolds 1999), and the behaviour of individuals with disabilities (Christensen and Sasaki 2008).

The most common hazardous events in the built environment are related to fire (Federal Emergency Management Agency 2010) and this research considers agent-based modelling in the context of fire evacuation. Specifically, two aspects are examined: human evacuation behaviour based on fire investigation reports (for more details, see Roan *et al.* 2011) and navigation algorithms, which are described here.

Two approaches to such modelling can be identified – continuous space and grid-based. Simulating a high density of occupants moving around in continuous space models such as Social Force (Helbing and Molnár 1995) can cause issues as agents are restricted to moving around to avoid pedestrians and obstacles rather than being allowed to overlap. In reality, according to fire reports, occupants sometimes step over others and cause serious stampedes in real fire situations. Therefore, our model divides the space into regular grid cells and allows agents to overlap in extreme situations.

This paper focuses on modified navigation approaches to address one of the challenges in grid-based models – route selection, simulating pedestrian movement using multiple path selections rather than a fixed route in order to model behaviour in a more realistic manner.

2. Modified Navigation Algorithms

The shortest path search approach and potential field approach are commonly used for navigation in agent-based models (Overmars *et al.* 2008, Bennewitz *et al.* 2002). The *shortest path search approach* is used for finding a path between two nodes based on a weighted graph (Foudil 2009), and one of the typical algorithms – *A* algorithm*, which is a generalisation of Dijkstra's algorithm (Dijkstra 1959, Hart *et al.* 1968), uses a distance-plus-cost heuristic function to determine a list of nodes for an optimal route. The *potential field approach* uses potential distance calculated between coordinates and predefined waypoints (Pelechano *et al.* 2007, Koh *et al.* 2008). An example of this - the *priority queue flood fill algorithm* calculates distance costs by selecting the lowest distance cost as a prioritised node.

Existing models, such as EXODUS¹ model (Galea 1998), simulate interaction between pedestrians and environment in cell-based models but result in unrealistic movement with agents moving at 45 degrees as their first movement and changing directions until they meet an obstacle (Pelechano and Malkawi 2008). Figure 1 shows the effect of the movement from STEPS² software (Mott MacDonald 2009).

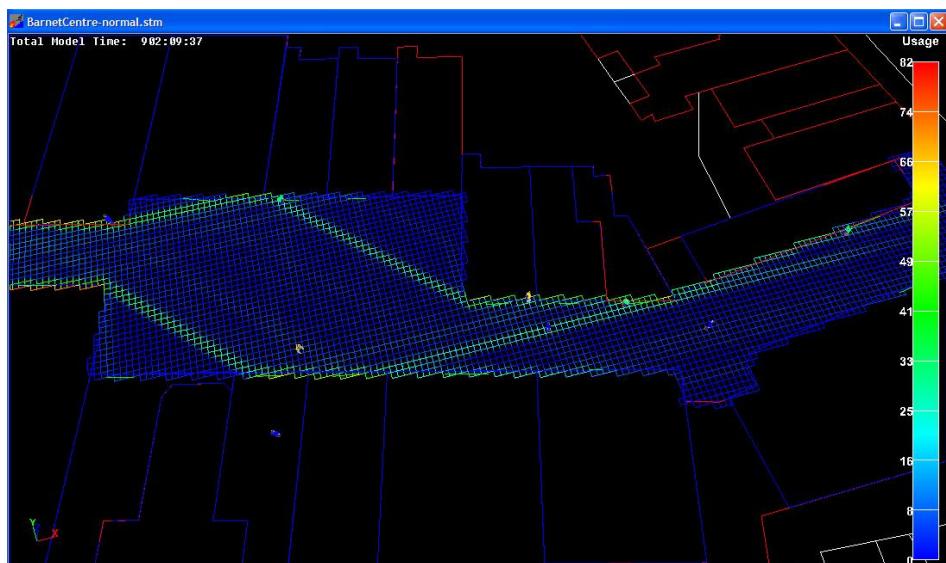


Figure 1. A frequency of grid usage based on a potential map shows the trajectory of pedestrian movement.

In general, the priority queue flood fill algorithm requires that agents always move to an adjacent cell with the lowest distance cost, and the A* algorithm selects the nearest node to final target if it calculates more than one node with the same lowest cost. Therefore, both methods return a fixed route selection that force agents to move towards

¹ EXODUS is developed by the Fire Safety Engineering Group at the University of Greenwich. The model is based on a set of sub models for evacuation simulations and pedestrian dynamics/circulation analysis.

² STEPS is a simulation tool developed by Mott MacDonald, UK. It is used to simulate pedestrian movement under a normal or emergency condition.

the same grid location (Foudil 2009, Overmars *et al.* 2008). Our evacuation model addresses these issues to simulate a more realistic pedestrian movement in a cell-based environment. We propose a modified algorithm which includes additional steps and directions when calculating distance cost, so pedestrian movement is determined by step numbers and directions instead of the calculated costs.

To validate the adapted versions of the A* and priority queue flood fill algorithms, a test scenario was developed (Figure 2). The potential field approach now calculates distance costs from an exit to every cell and creates a potential table, and the shortest path search approach calculates costs from each person's location to the destination. In both cases, after calculating a full list of costs, a path is identified in terms of step numbers and directions from an exit to the occupant location to ensure pedestrian can reach the final target. Unlike the standard approach (Figure 2-a), multiple start-to-finish routes may be considered (Figure 2-b) - the result on the test scenario in Figure 2-c shows 8 potential routes for the yellow agent, 3 potential routes for the brown agent, and the red agent has 34 potential routes from the starting point to the exit. These paths are more flexible compared to one fixed route from the standard calculations.

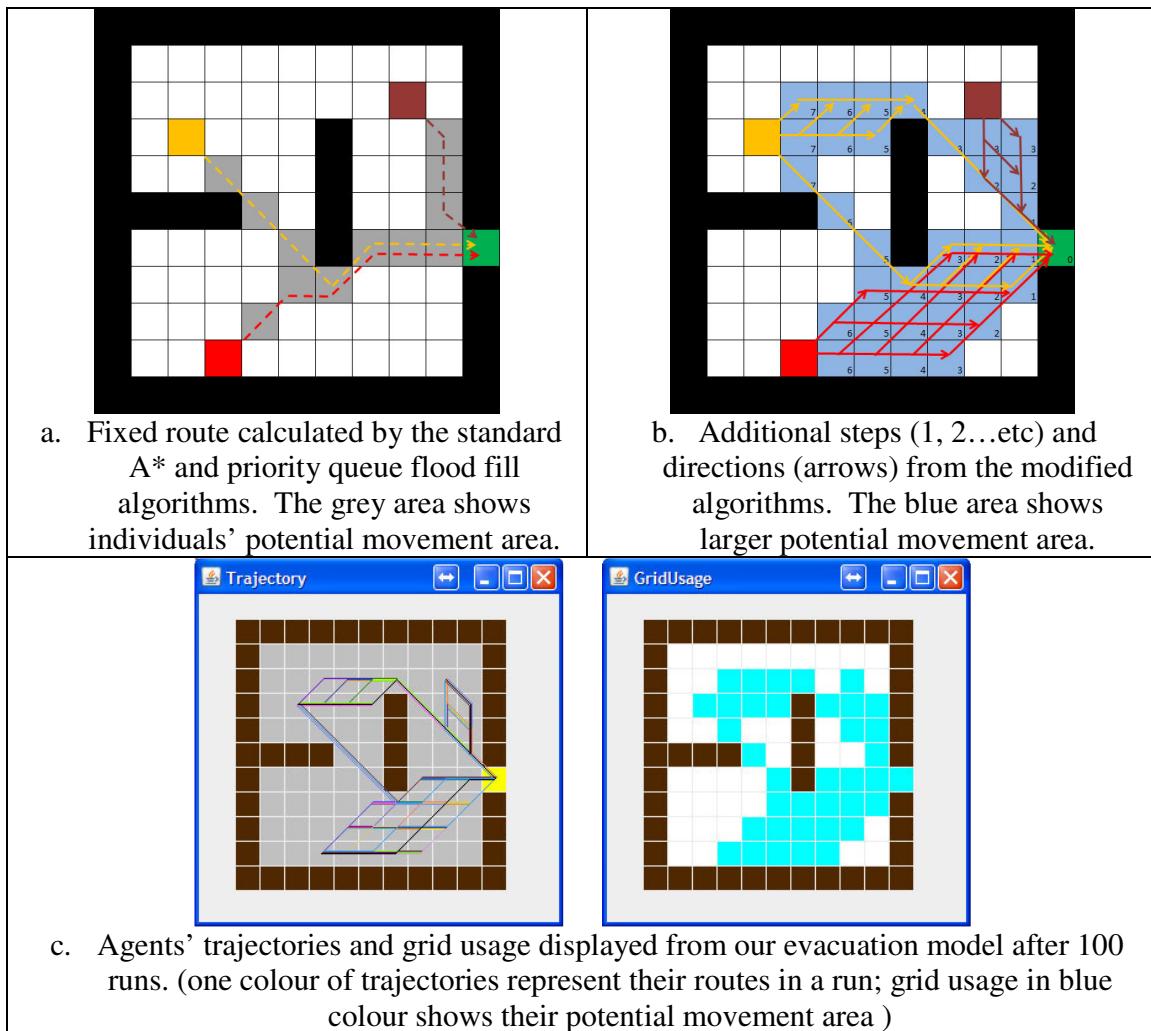


Figure 2. Standard and modified navigation algorithms in a test scenario.

3. Implementation and Results

A key limitation of the majority of existing building evacuation models on fire, such as the Gothenburg Disco fire simulation (Jiang *et al.* 2001), is that they only simulate agents evacuating from standard building exits. However, as evidenced by our review of fire investigation reports (Best and Swartz 1978, Yates 1991), people sometimes hide in a room or stand at windows waiting for rescue. Evacuation behaviour has been implemented in our model on the basis of a review of twenty fire investigation reports to represent a more complete and realistic test.

A 71x21 grid scenario (grid size: 0.5m^2) was built based on Comeau and Duval's report (2000) using JAVA programming with an agent-based toolkit, Repast Simphony (North *et al.* 2007). This report recorded a fire incident which resulted in 63 deaths and 180 injuries in a nightclub in Gothenburg, Sweden on 28 October 1998. The officials estimated that there were more than 400 occupants in the dance hall, whereas the building was only permitted 150 people at that time. There were two main exits which could allow people out of the dance hall, but the fire started at the southeast stairway and thus this exit was not able to be used during the evacuation. In addition, security bars were installed across the south side windows and three rooms were locked to avoid occupants entering during the party.

The simulation starts from a fire alarm that forces *pedestrian agents* to evacuate towards the main entrance where they entered. At this stage, they move in an orderly manner and form a queue at the exit. When the first *pedestrian* discovers the smoke, *pedestrians* communicate to warn each other of this hazard situation, and then their behaviour changes to panic. Therefore, *pedestrians* recalculate their routes according to individual own decisions, for example, rushing to an exit nearby, evacuating through alternative exits, seeking shelter in a room, or calling for helps from a window. In addition, other types of agents also influence pedestrian movement – *pedestrians* will not move towards the *fire*; *pedestrians* will faint after they inhale a specific amount of *smoke*, and they either die or are rescued by fire-fighters later; a *pedestrian* who discovers another *pedestrian* lying on the floor will chose to go around or step over the body; *exit agents* control pedestrian flow so that *pedestrians* move more slowly if too many agents rush to one exit. Figure 3 display grid usages of 400 agents' movement calculated by the two modified algorithms after 100 runs, and it shows the difference in tracking tendencies – agents tend to move in a diagonal direction and walk along the wall using the potential field approach, whereas in the shortest path search approach agents move straight toward the exit.

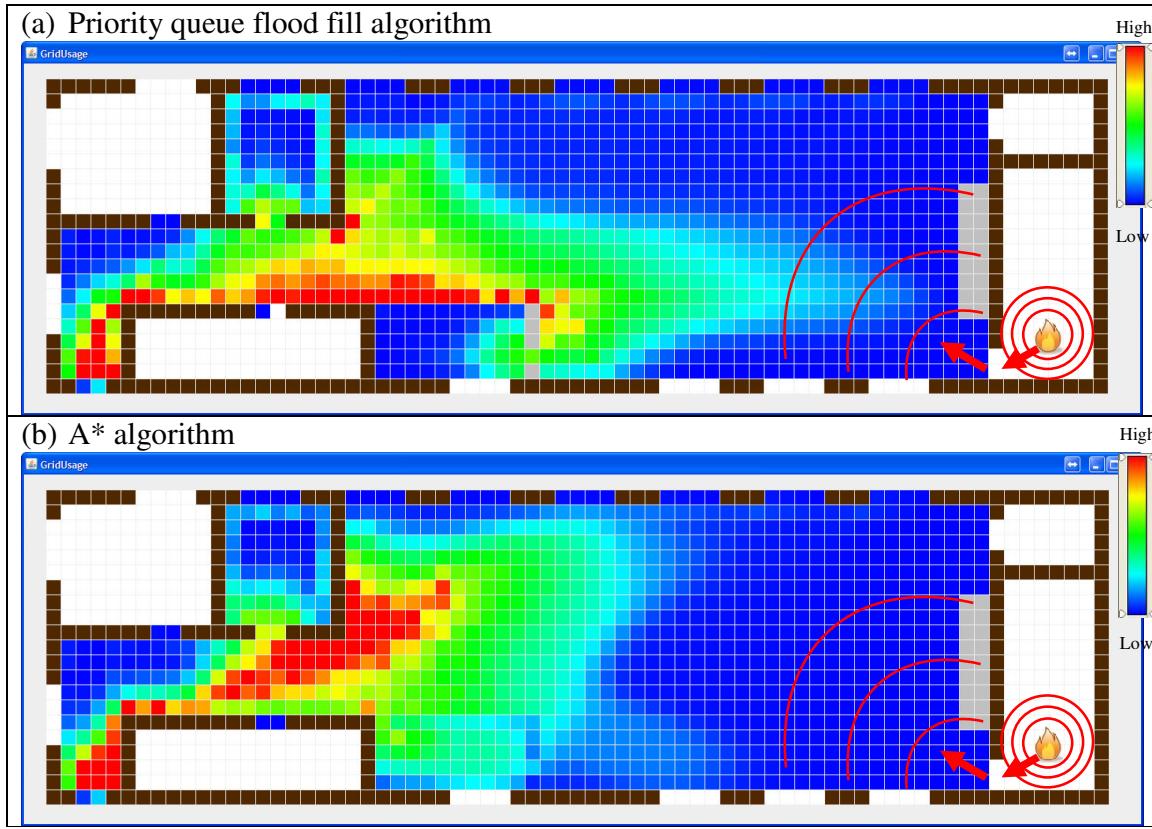


Figure 3. Grid usage maps of 400 agents' movement after 100 runs. The main entrance is at the south-west corner; some rooms and windows were locked, so occupants could not access them during the evacuation (white cells)

4. Conclusion and Further Work

Pedestrian movement, which is determined as a combination of modified navigation algorithms and pedestrian behaviour, influences overall evacuation time during the simulation. This paper presents modified algorithms to overcome issues with existing agent-based evacuation models in which agents are often routed to the same destination cell. Additionally, our evacuation behaviour, which is based on fire investigation reports, simulates a more realistic representation of egress selection. With the improvement of navigation calculation and behaviour determination, the model results an increasing accuracy of total evacuation time.

However, additional factors (such as individual height, gender, education level, group behaviour, pre-evacuation activities and location), which might influence individuals selecting egress routes, are not included in this stage of the model. Furthermore, this model simplifies smoke spread as having regular speed and movement which also influences the result. As shown in Figure 4, the model does not always results the correct location of deaths compared to the records in the selected fire report – in the fire report 43 bodies appeared around the main entrance and other 20 were found in the shelter room. Therefore, additional research into fire/smoke behaviour, how fire/smoke spreads through

the space, how furniture influences the burning of fire and pedestrian movement, and how people inhale smoke should also take into consideration.

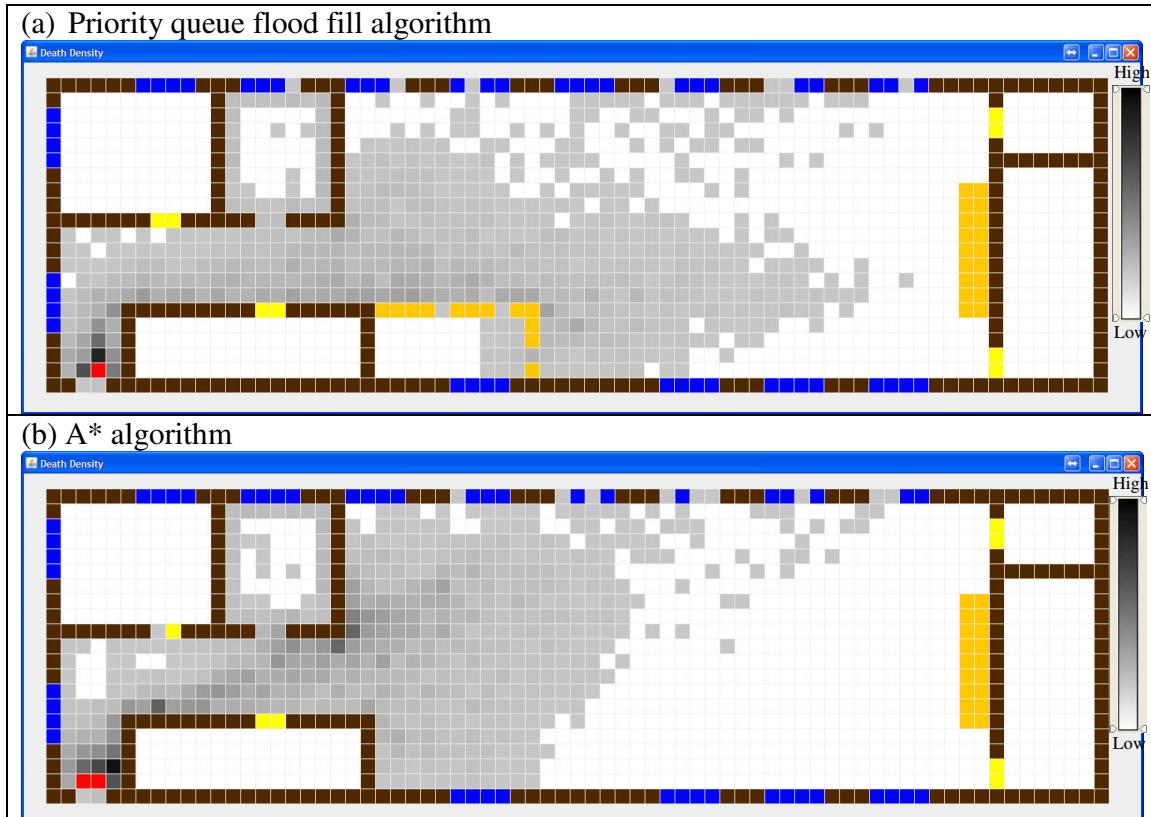


Figure 4. Percentage of death location in 100 simulations (red: very high possibilities where casualties appear).

5. References

- Bennewitz, M., Burgard, W. and Thrun, S., 2002. Finding and Optimizing Solvable Priority Schemes for Decoupled Path Planning Techniques for Teams of Mobile Robots. *Robotics and Autonomous Systems*, 41(2-3), pp.89-99.
- Best, R.L. and Swartz, J.A., NFPA, 1978. *Beverly Hills Supper Club Fire, Southgate, KY (May 28, 1977)*, National Fire Protection Association.
- Christensen, K. and Sasaki, Y., 2008. Agent-Based Emergency Evacuation Simulation with Individuals with Disabilities in the Population. *Journal of Artificial Societies and Social Simulation*, 11(3).
- Comeau, E. and Duval, R.F., NFPA, 2000. *Dance Hall Fire, Gothenburg, Sweden (October 28, 1998)*, National Fire Protection Association.
- Dijkstra, E.W., 1959. A Note on Two Problems in Connexion with Graphs. *Numerische Mathematik*, 1(1), pp.269-271.
- Federal Emergency Management Agency, 2010. FEMA: Hazards: Fire. Available at: <http://www.fema.gov/business/guide/section3a.shtm> [Accessed 15:17:24].
- Foudil, C., 2009. Path Finding and Collision Avoidance in Crowd Simulation. *Journal of Computing and Information Technology*.
- Galea, E.R., 1998. A General Approach to Validating Evacuation Models with an Application to EXODUS. *Journal of Fire Sciences*, 16(6), pp.414 -436.

- Hart, P.E., Nilsson, N.J. and Raphael, B., 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2), p.100–107.
- Helbing, D. and Molnár, P., 1995. Social Force Model for Pedestrian Dynamics. *Physical Review E*, 51(5), p.4282.
- Jiang, H. *et al.*, 2001. The Use of Evacuation Simulation, Fire Simulation and Experimental Fire Data in Forensic Fire Analysis. *Psychological Review*, 105(3), p.530–557.
- Koh, W.L., Lin, L. and Zhou, S., 2008. Modelling and Simulation of Pedestrian Behaviours. In *Proceedings of the 22nd Workshop on Principles of Advanced and Distributed Simulation*. IEEE Computer Society, pp. 43-50.
- Mott MacDonald, 2009. *STEPS-Simulation of Transient Evacuation and Pedestrian Movements On-line User Manual*.
- Musse, S.R. and Thalmann, D., 1997. A Model of Human Crowd Behavior: Group Inter-Relationship and Collision Detection Analysis. In *Proceedings of the Eurographics Workshop on Computer Animation and Simulation'97*. Budapest, Hungary, p. 39.
- North, M.J. *et al.*, 2007. *Visual Agent-Based Model Development with Repast Simphony*, Argonne National Laboratory (ANL).
- Overmars, M., Karamouzas, I. and Geraerts, R., 2008. Flexible Path Planning Using Corridor Maps. In *Algorithms - ESA 2008*. pp. 1-12.
- Pelechano, N., Allbeck, J.M. and Badler, N.I., 2007. Controlling Individual Agents in High-Density Crowd Simulation. In *Proceedings of the 2007 ACM SIGGRAPH/Eurographics symposium on Computer animation*. San Diego, California: Eurographics Association, pp. 99-108.
- Pelechano, N. and Malkawi, A., 2008. Evacuation Simulation Models: Challenges in Modeling High Rise Building Evacuation with Cellular Automata Approaches. *Automation in Construction*, 17(4), pp.377-385.
- Reynolds, C.W., 1999. Steering Behaviors For Autonomous Characters. In *Proceedings of Game Developers Conference*. San Jose, California.
- Roan, T.-R., Haklay, M. and Ellul, C., 2011. Modelling Fire Evacuation Behaviour Based on Fire Investigation Reports. In *Proceedings of the GIS Research UK 19th Annual Conference GISRUK 2011*. Portsmouth, UK.
- Wolshon, B. *et al.*, 2005. Review of Policies and Practices for Hurricane Evacuation. I: Transportation Planning, Preparedness, and Response. *Natural Hazards Review*, 6(3), pp.129-142.
- Yates, J., USFA, 1991. *Chicken Processing Plant Fires Hamlet, North Carolina (September 3, 1991) and North Little Rock, Arkansas (June 7, 1991)*, Federal Emergency Management Agency, United States Fire Administration, National Fire Data Center.

Integrating an Agent-Based Model and a Population Microsimulation to Explore Crime Patterns

Nick Malleson and Mark Birkin

January 28, 2011

1 Introduction

Crime is an extremely complex phenomenon. In order to understand and to predict crime patterns it is necessary to examine the behaviour of the offender(s), the physical attributes of the surrounding environment and the behaviour of other people who might be able to influence the event, such as the victims or passers-by. To further complicate matters, each of these elements are inherently local in nature; research that spatially aggregate these features will disregard important information [1] and are not able to truly capture the dynamics of systems that are non-linear and involve feedback [2] – such as the crime system.

For these reasons, the agent-based modelling methodology has started to be used in quantitative crime research to better understand and predict crime. The methodology involves simulating the individual components of a system directly (such as individual “offenders”) and hence constructing an artificial geographic system that closely replicates the real system under study. To take advantage of the benefits offered by agent-based modelling, *BurgdSim* [4] is an agent-based model that creates a realistic representation of an urban environment and simulates the spatio-temporal behaviour of individual offenders to predict occurrences of residential burglary. Although the model contains individual houses, roads and burglar agents, other people who might influence the system (e.g. residents, passers-by) are included at an aggregate level due to a lack of individual-level data. This is a considerable drawback to the simulation because criminologists suggest that victim behaviour is an important determinant of crime risk.

Fortunately, although there is insufficient *primary* individual-level data to include in the simulation it is possible to use the technique of microsimulation to synthesise a population of individuals from aggregate-level data sources. This paper will present ongoing research into generating a spatially-explicit population of synthetic individuals from census data and using this as an input into an agent-based burglary model. Although still in early stages, preliminary results show that a lot of information about the demographics of potential burglary victims can be gained by aligning microsimulation and agent-based models.

2 Existing Tools

2.1 The Burglary Model

BurgdSim is an advanced agent-based model of crime that simulates the behaviour of intelligent “burglar” agents and predicts occurrences of the crime of residential burglary. The model has been implemented in Java using the Repast Simphony library (<http://repast.sourceforge.net/>) (as illustrated by Figure 1) and has also been adapted to run on the grid using National Grid Service (NGS) compute resources. The model uses real GIS data to create a virtual environment that consists of the following layers:

- **Household layer.** The household layer includes a representation of individual houses in the study area. As the geometry of each building is available, they have been analysed in

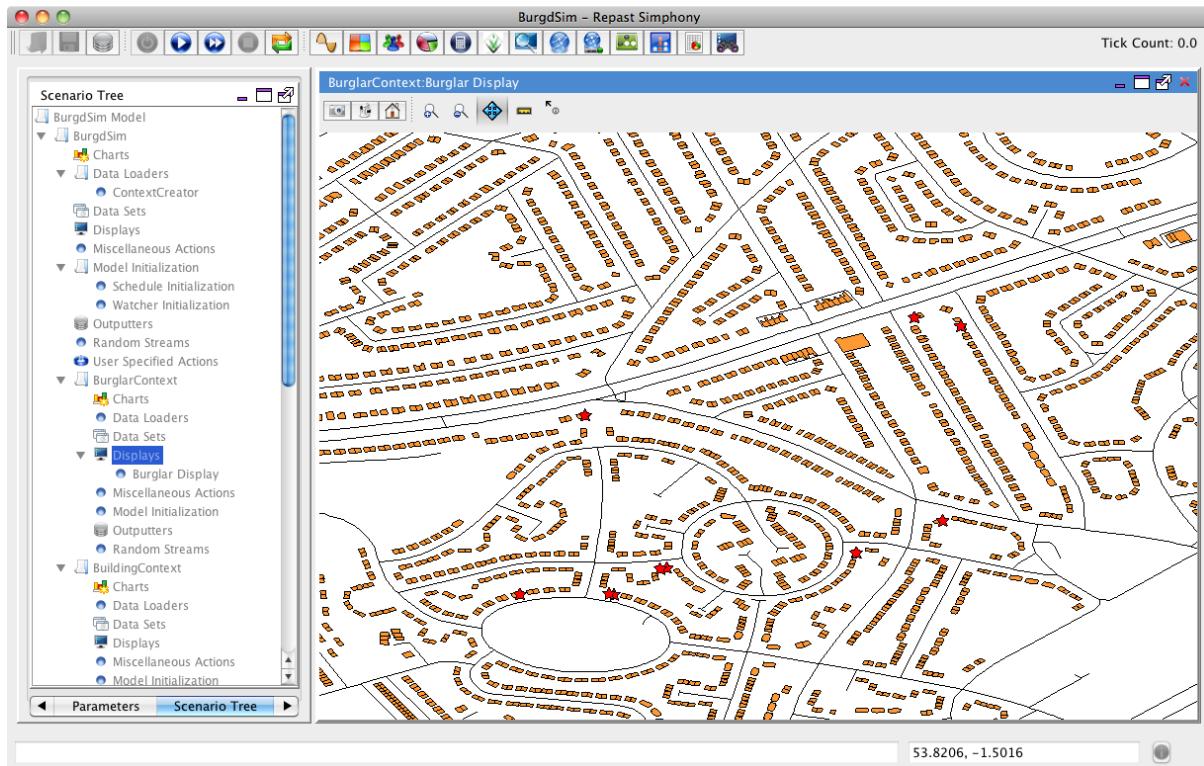


Figure 1: The *BurgdSim* user interface depicting a small number of burglar agents in the virtual environment – part of Leeds in this case.

a GIS to estimate features that might influence their burglary risk, such as visibility to neighbours or passers-by and the building structure (terraced, detached or semi-detached).

- **Road network.** As with the household layer, the road network is built up from real GIS data to create a model of the network that agents can use to traverse the study area. Different types of roads have been included (such as motorways, alleyways, minor roads etc) and these affect the speed of travel across a road as well as the types of vehicle that are permitted.
- **Communities layer.** The purpose of the communities layer is to simulate the effects that *other people* (i.e. non-burglars) might have on occurrences of burglary. The layer uses UK census data to estimate at what times a house in the community might be unoccupied, how wealthy the residents are and how cohesive the community is.

The burglar agents themselves have been implemented using the PECS cognitive architecture to simulate intelligence [5]. They all have a home and exhibit dynamic behaviour that changes depending on their current circumstances. The agents require money for certain behaviours (such as drug-use and socialising) and, in these cases, must attempt to commit a burglary first. Therefore, although the model is able to accurately represent individual burglar agents and the houses that they might attempt to burgle, the victims of burglary are represented in the communities layer and are therefore homogeneous across all houses in the community. This is a drawback for a model that is otherwise able to simulate at the level of the individual. However, the problem is purely a result of data availability; demographic data obtained from the 2001 UK census have been aggregated. As the following section will discuss, there is software available that can be used to disaggregate the census and, hence, produce an estimate of the individual people/families who live in each house.

2.2 NeISS and the Population Reconstruction Model (PRM)

The National e-Infrastructure for Social Simulation (NeISS: <http://www.neiss.org.uk>) is a multi-disciplinary project that aims to develop new tools and services for social scientists and planners. The tools will enable users to run their own simulations and visualise/analyse results as well as share them for future discovery and reuse. A tool of particular relevance is the Population Reconstruction Model (PRM) as this can be used to disaggregate the UK census.

The PRM is a *microsimulation* technique which uses a combination of Small Area Statistics and anonymised individual records to provide a synthetic population of individual people and families for any region which has available census data. Although it is not possible to validate the resulting synthetic population directly (as data comparable to that which the procedure generates are not available), re-aggregation of the population show an extremely close match to the distributions from which they are derived [3], adding confidence to the accuracy of the results. The individual level data that are generated can be extremely valuable for subsequent applications, such as *BurgdSim*, as the following section will discuss.

3 Preliminary Results

In order to improve the representation of crime victims in the *BurgdSim* model, the microsimulation was used to generate individual-level demographic data from the 2001 UK census and this was subsequently used as an additional input into the agent-based model. Therefore the model could be adapted so that the wealth levels of potential victims and their occupancy behaviour (the times that they leave their houses unoccupied as estimated from their employment) were no longer homogeneous across all houses in a community, but were unique to individual houses. Therefore, when burglar agents decide which houses they will target for burglary they take individual-level victim characteristics into account, rather than assuming that all people in a neighbourhood are identical.

Although the procedures used to combine the synthetic population data with the agent-based model are in their infancy, preliminary results suggest that already the improved model is able to offer additional insights into the simulated burglary victims. Figure 2 illustrates the demographic characteristics of the population once all people have been assigned to virtual houses in the model and compares these to the demographics of burglary victims after a simulation has been completed. With the exception of *social group*, none these attributes (*age*, *gender* and *ethnicity*) are taken into account by the burglar agents during their assessment of where to commit a burglary. Therefore these trends are a result of *where* in the city the individuals live as well as the types of houses and neighbourhoods they inhabit rather than an artefact of model rules. The *managerial* social group is a prime example of this. Although in the synthetic population there are a similar number of people with managerial employment to those with manual jobs or unemployed/students, people with managerial jobs are rarely predicted to be victims of burglary. This is most likely because they live in places that the burglar agents are unaware of or do not consider suitable for burglary.

Clearly further research is needed to clarify and confirm these findings, but the results suggest that the combination of agent-based and microsimulation techniques have a lot to offer in terms of geocomputational prediction and modelling.

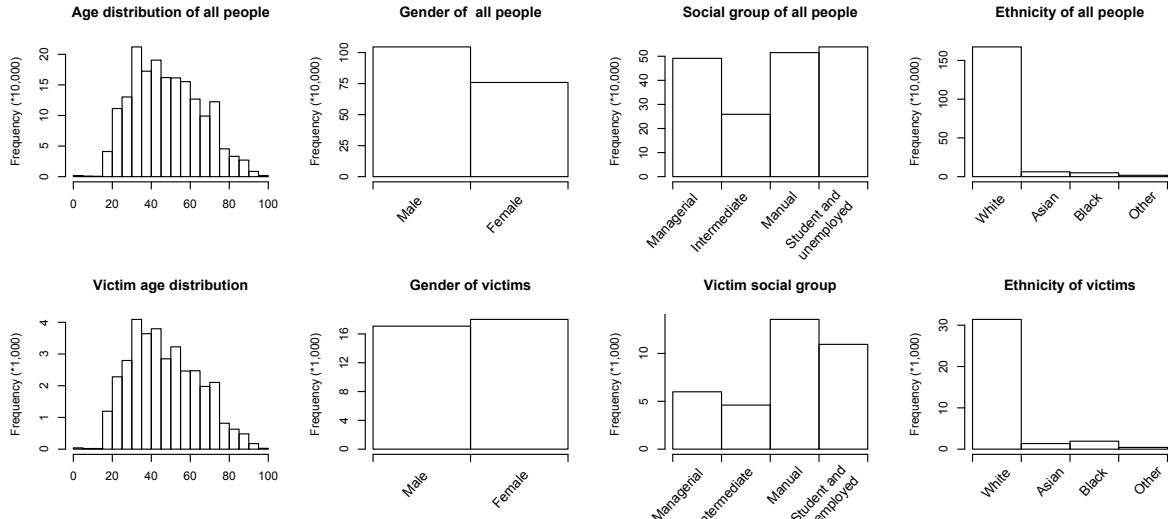


Figure 2: Demographics of the synthetic individual population generated using the PRM and the subset of individuals who were burgled in the agent-based model.

References

- [1] Martin A. Andresen and N. Malleson. Testing the stability of crime patterns: implications for theory and policy. *Journal of Research in Crime and Delinquency*, forthcoming, 2010. Under review.
- [2] John E. Eck and Lin Liu. Contrasting simulated and empirical experiments in crime prevention. *Journal of Experimental Criminology*, 4(3):195–213, 2008.
- [3] K. Harland, M. Birkin, A. Heppenstall, and D. Smith. Creating realistic synthetic populations at varying spatial scales: A comparative critique of microsimulation techniques. *Journal of Artificial Societies and Social Simulation*, 2010. in press.
- [4] Nick Malleson. *Agent-Based Modelling of Burglary*. PhD thesis, School of Geography, University of Leeds, LS6 3DT, UK, 2010.
- [5] Nick Malleson, Linda See, Andy Evans, and Alison Heppenstall. Implementing comprehensive offender behaviour in a realistic agent-based model of burglary. *Simulation: Transactions of the Society of Modeling and Simulation International*, 2010. in press, published online.

Understanding Route Choice using Agent-based Simulation

Ed Manley¹, Dr Tao Cheng¹, Andy Emmonds²

¹Department of Civil, Environmental and Geomatic Engineering, University College London, Gower Street, London, United Kingdom
Telephone: +44 (0)20 7679 2898
Email: Edward.Manley.09@ucl.ac.uk, Tao.Cheng@ucl.ac.uk

²Transport for London, 197 Blackfriars Road, London, United Kingdom
Telephone: +44 (0)203 054 0911
Email: Andy.Emmonds@tfl.gov.uk

1. Introduction

In traffic analysis and simulation, it is usually assumed that all individuals hold a complete knowledge of the road network and a homogenous preference in route choice, either via the shortest (in time or distance) or least cost path. These modelling assumptions do not, it is argued, truly represent human preference in relation to route choice. Rather, the shortest path strategy is viewed as one factor influencing choice. Golledge describes revealed use of first noticed path, fewest turns and shortest leg first (Golledge 1995). Conroy-Dalton (2003) also demonstrated how individuals primarily seek to minimise the number of turns as they proceed along their route. In a recent study, in investigating real path correlation with shortest path, Papinski and Scott (2011) demonstrated that movement does not follow shortest length or least time paths. It has also been found that the shortest path method performed worse than least angular change and least turns in predicting the movement of vehicles through four small test areas in London (Hillier and Iida, 2005).

This paper seeks to add to this growing literature on route choice methodology by testing these measures within an agent-based simulation environment. The model, described in Section 2, simulates the movement of multiple individual agents across the London road network between given origin and destinations. The movement patterns created by these agents will be compared to real movement data (described in Section 3), with initial results documented (Section 4) and discussed (Section 5) herein. This work represents an initial yet contributory step towards establishing a realistic route choice model for use in traffic simulation.

2. Model Development

An Agent-based Simulation was developed to simulate the movement of individuals around the complete London road network. The model is an extension of that described in Manley and Cheng (2011) – a Java-based application developed using the Repast framework – with inter-agent variation contained within the route choice mechanism applied in wayfinding. Between a given origin and destination (restricted to those selected for testing, described in Section 3), agents minimise their path cost according to one of four measures, these are as follows:

- Metric: The shortest length path between origin and destination.
- Angular Change: The least cumulative angular change between origin and destination, where deviation at each junction is accumulated.
- Turns: The least number of turns between origin and destination.
- Angular Choice: Minimising the ‘Angular Choice’ value associated with each segment. This measure is a betweenness value scored for each segment when it falls

on the shortest angular path between any origin and destination. This value is calculated for all possible origins and destinations (see Turner 2001).

The former three measures described here represent an extension of the work carried out by Hillier and Iida (2005), while Angular Choice has also been recognised as a possible predictor of route choice (Turner 2007). Agents proceed towards their destination at a given speed and coordinate at junctions according to a set of priority rules. Traffic regulations are implemented also to ensure a parallel with real data, with most-notably Oxford Street – a key road in central London – being closed to all through traffic. The resulting paths are then exported by the simulation into an ArcGIS shape file for comparison with movement data.

3. Test Data

The test dataset is drawn from a database of taxi driver traces provided by Addison Lee Taxi Company. This dataset contains the GPS traces of some 1.5 million trips between locations in London over a three month period spanning December 2010 to February 2011. For the purposes of this initial study, four test scenarios were extracted representing a range of routes within central London. The scenarios used were as follows:

- Scenario 1:** Knightsbridge (SW7) to Herne Hill (SE24) on 15th February 2011 between 18:03 and 18:43.
- Scenario 2:** Saville Row (W1) to Highbury and Islington (N1) on 16th February 2011 between 16:01 and 16:28.
- Scenario 3:** Islington (N1) to Chelsea Royal Hospital (SW3) on 15th February 2011 between 20:26 and 21:01.
- Scenario 4:** Abbey Road Studios (NW8) to Bermondsey Wall (SE1) on 16th February 2011 between 14:16 and 15:35.

For each scenario, the corresponding GPS traces were matched to the ITN road network. This process yielded polyline data that can be seen in the result maps below. The origin and destination points for each scenario are passed to the agent-based simulation for the production of test routes according to each agent's rules.

4. Results of Simulation

The simulation yields a total of 16 datasets exhibiting the movement of agents defined using each of the four routing mechanisms in each of the four scenarios. Maps of these results are presented below in Figure 1, with further analysis of route similarity presented below:

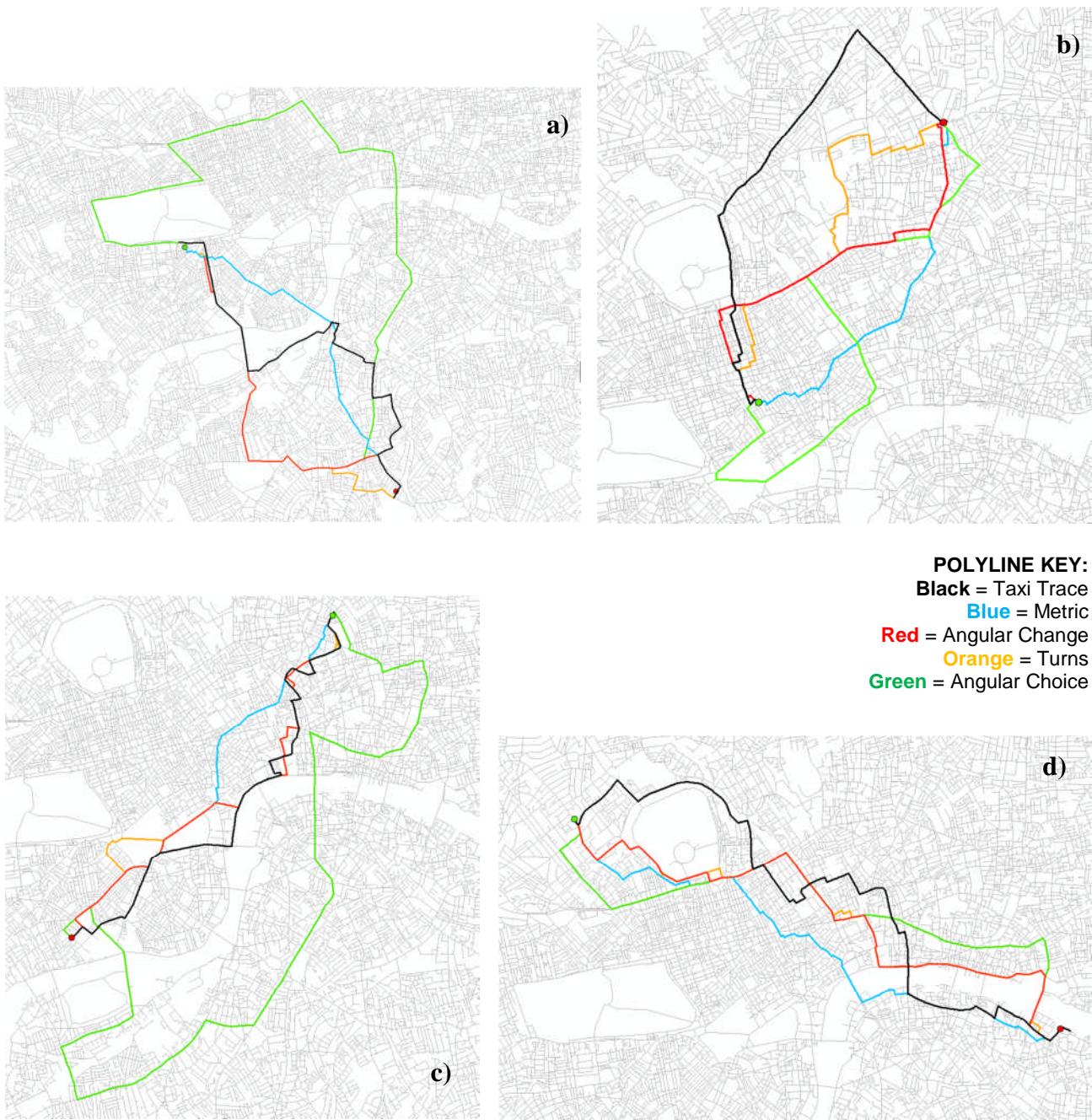


Figure 1:

- a) Scenario 1: Knightsbridge (Green point) to Herne Hill (Red point)
- b) Scenario 2: Saville Row (Green point) to Highbury and Islington (Red point)
- c) Scenario 3: Islington (Green point) to Chelsea Royal Hospital (Red point)
- d) Scenario 4: Abbey Road Studios (Green point) to Bermondsey Wall (Red point)

Using the route datasets yielded from the simulation, it is also possible to calculate the extent to which the real taxi driver route is predicted by the routes of each agent. These results are calculated on a segment by segment basis and are as follows:

Route Choice Measure	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	Segments Matched	%	Segments Matched	%	Segments Matched	%	Segments Matched	%
Length	17/136	12.5	1/123	0.8	21/141	14.9	28/169	16.6
Angular Change	25/136	18.4	10/123	8.1	41/141	29.1	14/169	8.3
Turns	20/136	14.7	8/123	6.5	30/141	21.3	7/169	4.1
Angular Choice	16/136	11.8	5/123	4.1	2/141	1.4	14/169	8.3

5. Discussion and Conclusions

The results generated from the simulation suggest that none of the four metrics employed to route agents between two locations provide a full answer to the route choice conundrum. However, as has been also noted by others, the results from these scenarios demonstrate a clear difference between reality and the shortest path algorithm. In three of the four scenarios, least cumulative angular change and least number of turns represented better models of movement than simply shortest path. Where the shortest path algorithm did score favourably (in scenario 4) this may be put down to the point at which the individual driver decided to cross the River Thames en route to the destination. The results indicate that, in extension to the work of Hillier and Iida (2005), angular change and number of turns are also employed as heuristics in guiding longer journeys within the urban environment. The selection of these measures, understood as ‘most direct’ (least angular) and ‘simplest’ (least turns) paths, align more with human preference than expressed by existing transport models.

The performance of Angular Choice as a predictor was demonstrated to be variable during these investigations. In the cases of scenarios 1 and 3, the agent appears to travel some considerable distance away from the target before converging upon it. Yet equally, in the case of scenario 4, its performance surpasses that might have been expected. The answer perhaps lies in the distribution of high scoring segments as defined by the Angular Choice measure. The location of these highly-attractive roads – albeit those which appear to correlate with high traffic flows – in relation to the origin and destination appears to influence the quality of these results. For instance, three of these higher scoring sections are Euston Road, Woburn Place/Southampton Row and Holborn Viaduct (all featuring within the top 5% of Angular Choice values in the London road network), all of which fall between the origin and destination of scenario 4.

There are, of course, a number of caveats that must be offered alongside these results. Firstly, the small sample size presented cannot be representative of the complete variation in route choice that may be observed. In the case of scenario 2, there is a vast difference between the actual route and those predicted by all four measures. For this piece of work no further investigation behind the dynamic influences upon route choice (such as congestion avoidance, knowledge of a road closure etc.) has been carried out. Furthermore, the extent to which local traffic regulations impact on these results is equally not fully incorporated, with only basic rules implemented at this stage. Finally, the influence of local knowledge should

not be discounted in assessing correlation. While taxi drivers may generally be expected to have a good knowledge of the road network, this is by no means confirmed in this situation. Scenario 2, for example, may represent a driver wishing to avoid the busy Upper Street road (chosen by the driver agents) yet not having knowledge of a more direct route to the final destination.

In conclusion, this work presents an opportunity for further investigation into the prevalence of such factors during the process of route choice. The drawbacks of this investigation should be tackled at the next iteration and the study extended to account for individual variation and traffic dynamics. Other measures, relating also to the city configuration, should also be investigated for correlation with these data. Of particular note may be that of travel time, an improvement upon shortest metric path and also widely employed within transportation modelling. By extending this work it may be possible to begin to draw clearer trends with regard to the most important measures employed by drivers and how the influence of these parameters vary with space and time.

Acknowledgements

This work is part of the STANDARD project – Spatio-Temporal Analysis of Network Data and Road Developments, supported by the UK Engineering and Physical Sciences Research Council (EP/G023212/1) and Transport for London (TfL). The authors would also like to thank Addison Lee Ltd and Space Syntax Ltd for the provision of data to support this work.

References

- Conroy-Dalton, R. A. 2003. The secret is to follow your nose: Route path selection and angularity. *Environment and Behaviour* 35: 107–131.
- Golledge, R.G. 1995. *Path Selection and Route Preference in Human Navigation: A Progress Report*. Berkeley, California: The University of California Transportation Center.
- Hillier, B. & Iida, S. 2005. Network and psychological effects in urban movement. In: Cohn, A. G., Mark, D. M. (Eds.) *Spatial Information Theory: COSIT 2005*, Lecture Notes in Computer Science number 3693, 475–490, Springer-Verlag, Berlin.
- Manley, E. & Cheng, T. 2011. Multi-Agent Simulation of Drivers Reactions to Unexpected Incidents on Urban Road Networks. *Proceedings of GISRUK Conference 2011*. Portsmouth, UK.
- Papinski, D. & Scott, D. 2011. A GIS-based toolkit for route choice analysis. *Journal of Transport Geography* 19: 434-442.
- Turner, A. 2001. Angular analysis, 3th International Space Syntax Symposium, Georgia Institute of Technology, Atlanta.
- Turner, A. 2007. From axial to road-centre lines: a new representation for space syntax and a new model of route choice for transport network analysis. *Environment and Planning B: Planning and Design* 34:539–555.

Merging Areal and Point Data in Medical Geography and Soil Mapping

P. Goovaerts

BioMedware, 3526 W Liberty, Suite 100, Ann Arbor, MI 48103
 Telephone: 001-734-913-1098 (ext. 202)
 Fax: 001-734-913-2201
 Email: Goovaerts@biomedware.com

1. Introduction

A common issue in spatial interpolation is the combination of data measured over different spatial supports. For example, in the field of medical geography (Goovaerts, 2009) information available for mapping disease risk typically includes point data (e.g. patients residence) and aggregated data (e.g. socio-demographic and economic data at the census tract level). Similarly, soil measurements recorded at discrete locations on the ground are often supplemented with choropleth maps (e.g. soil or geological maps) that model the spatial distribution of soil attributes as the juxtaposition of polygons (areas) with constant values (Goovaerts, 2011). This paper presents a coherent geostatistical approach to accommodate both areal and point data in the spatial interpolation of continuous attributes. The procedure is illustrated using two datasets: 1) geological map and heavy metal concentrations recorded in the topsoil of the Swiss Jura, and 2) incidence rates of late-stage breast cancer diagnosis per census tract and location of patient residences in Michigan for the period 1985-2002 (Figure 1).

2. Methodology

2.1 Area-and-Point Kriging

Consider the problem of estimating the value of a continuous attribute z at any location \mathbf{u} within a study area A . The information available consists of set of point data collected at n discrete locations \mathbf{u}_α $\{z(\mathbf{u}_\alpha); \alpha=1,\dots,n\}$, supplemented by a set of B areal data $\{z(v_k); k=1,\dots,B\}$ recorded for mapping units v_k of various size and shape. Both point and areal data can be simultaneously incorporated into the prediction using the Area-And-Point (AAP) kriging estimate defined as:

$$z_{AAP}^*(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha(\mathbf{u}) z(\mathbf{u}_\alpha) + \sum_{k=n(\mathbf{u})+1}^{n(\mathbf{u})+K} \lambda_k(\mathbf{u}) z(v_k) \quad (1)$$

where $n(\mathbf{u})$ and K are the number of surrounding point and areal data, respectively. Point observations are typically selected based on their distance to the interpolation node \mathbf{u} , while areal data are chosen according to adjacency rules; for example, all polygons adjacent to the polygon including \mathbf{u} are used in the estimation.

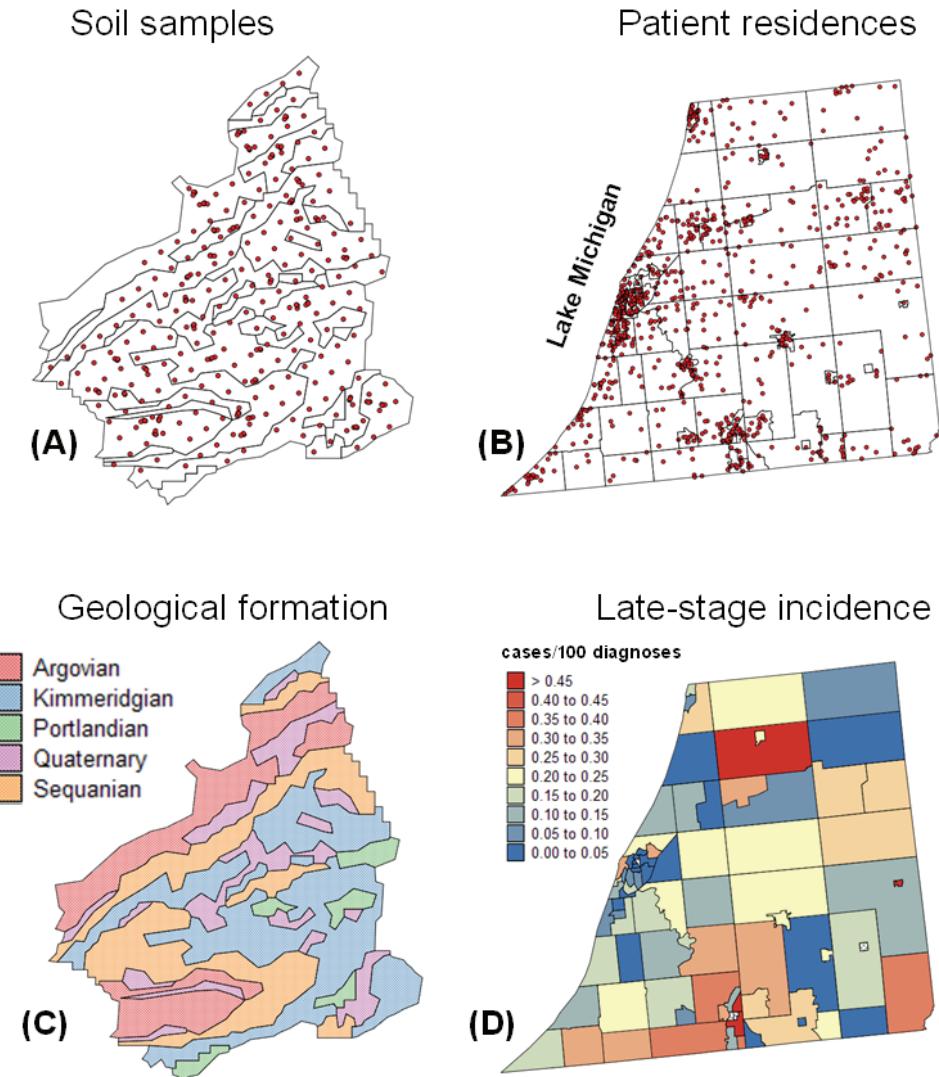


Figure 1. Information available for mapping topsoil heavy metal concentration and late-stage breast cancer incidence. (A) Soil field measurements. (C) Choropleth map of the main geological formations. (B) Location of 937 patient residences. (D) Choropleth map of late-stage breast cancer incidence rate in three Michigan counties, by census tract.

The kriging weights are the solution of the following ordinary kriging system:

$$\begin{aligned} \sum_{j=1}^{n(\mathbf{u})+K} \lambda_j(\mathbf{u}) \bar{C}(x_i, x_j) + \mu(\mathbf{u}) &= \bar{C}(x_i, \mathbf{u}) \quad i = 1, \dots, n(\mathbf{u}) + K \\ \sum_{j=1}^{n(\mathbf{u})+K} \lambda_j(\mathbf{u}) &= 1. \end{aligned} \quad (2)$$

where $\mu(\mathbf{u})$ is the Lagrange multiplier, and $x_i = \mathbf{u}_i$ if $i \leq n(\mathbf{u})$, and $x_i = v_i$ otherwise. The quantity $\bar{C}(x_i, x_j)$ is a point-to-point, point-to-block or block-to-block covariance depending on the indices i and j . Like in traditional block kriging, the block to-point covariances $\bar{C}(v_k, \mathbf{u})$ are approximated by the average of the point support covariance $C(\mathbf{h})$ computed between the location \mathbf{u} and a set of P_k points discretizing the block v_k . A

similar procedure is used for the block-to-block covariances $\bar{C}(v_k, v_{k'}) = \text{Cov}\{Z(v_k), Z(v_{k'})\}$ and involves averaging $C(\mathbf{h})$ computed between any two points discretizing the blocks v_k and $v_{k'}$. A major difference between AAP kriging and the related algorithms (area-to-area and area-to-point kriging) introduced recently in the geostatistical literature (Kyriakidis, 2004), is the availability of point data here. Thus, the point support semivariogram can be inferred directly from the observations without any need for a deconvolution of the areal semivariogram (Goovaerts, 2008).

2.2 Binomial Kriging

The application of AAP kriging to the medical geography case-study must account for the fact that the K areal data have varying degrees of reliability: these observations are incidence rates that tend to become unstable when the denominator (i.e. the number of cancer cases in this particular example) is small. On the other hand, point data can be viewed as an extreme case where the population size is one (individual-level data). The information about each cancer case, referenced geographically by its residence's spatial coordinates $\mathbf{u}_\alpha = (x_\alpha, y_\alpha)$, takes the form of an indicator of early/late stage diagnosis:

$$i(\mathbf{u}_\alpha) = \begin{cases} 1 & \text{if late - stage diagnosis} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The Area-And-Point (AAP) kriging estimate is now expressed as a linear combination of point indicator data and areal incidence rates:

$$z_{AAP}^*(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha(\mathbf{u}) i(\mathbf{u}_\alpha) + \sum_{k=n(\mathbf{u})+1}^{n(\mathbf{u})+K} \lambda_k(\mathbf{u}) z(v_k) \quad (4)$$

The kriging weights are the solution of the following system of linear equations (Webster *et al.*, 1994; Goovaerts, 2010):

$$\begin{aligned} \sum_{j=1}^{n(\mathbf{u})+K} \lambda_j(\mathbf{u}) \left[\bar{C}_I(x_i, x_j) + \delta_{ij} \frac{a}{n(v_i)} \right] + \mu(\mathbf{u}) &= \bar{C}_I(x_i, \mathbf{u}) \quad i = 1, \dots, n(\mathbf{u}) + K \\ \sum_{j=1}^{n(\mathbf{u})+K} \lambda_j(\mathbf{u}) &= 1. \end{aligned} \quad (5)$$

where $\delta_{ij}=1$ if $i=j$ and 0 otherwise, $a = m^*(1-m^*) - \bar{C}_I(v_i, v_i)$, $C_I(\mathbf{h})$ is an indicator covariance function, and m^* is the population-weighted mean of the N rates ($N=83$ census tracts here). The addition of the error variance term, $a/n(v_i)$, for a zero distance accounts for variability arising from population size, leading to smaller weights for less reliable incidence rates based on fewer cases.

3. Results and Discussion

Figure 2 (left column) shows the maps of chromium concentration estimated using alternative interpolation techniques. The reference approach is ordinary kriging (OK) that uses only field data (Fig. 2A). The other two maps incorporate areal data that take the form of average chromium concentration per geological mapping unit. These concentrations were used either as local means in residual kriging or directly incorporated

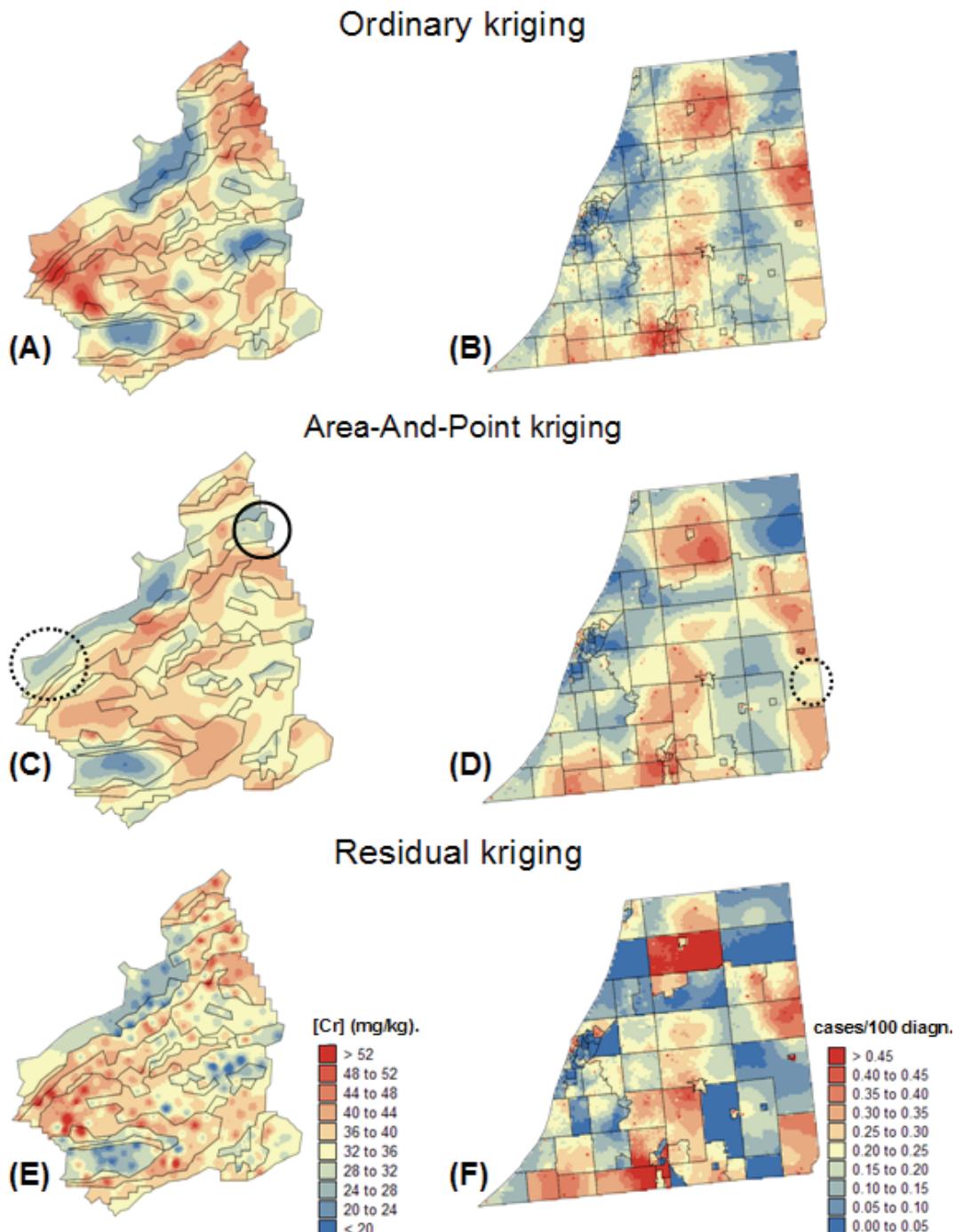


Figure 2. Maps of chromium concentration and late-stage breast cancer incidence rate created by alternative interpolation techniques. (A,B) Ordinary kriging. (C,D) Kriging that combines both areal and point data “AAP kriging”. (E,F) Residual kriging with a choropleth trend model. The same color scale is used for each series of three maps.

into the Area-And-Point estimator. In the later case, the average of kriged estimates equals the mapping units’ mean (coherence constraint). The residual semivariogram model has a short range, leading to “bull’s-eye” effect around sample points in the map

created by residual kriging (Fig. 2E). In contrast, the AAP map (Fig. 2C) is much smoother and clearly displays the lower concentrations expected on the Argovian formation. Differences between the three maps are the largest in sparsely sampled areas where the choice of a trend model becomes preponderant. In particular, incorporating the geological information leads to smaller estimates on the section of Argovian formation where no sample was collected (dashed circle in Fig. 2C) and in a small Argovian mapping unit that must satisfy the coherence constraint despite the presence of larger sampled concentrations (solid circle in Fig. 2C).

A similar analysis was conducted for the health outcome data in Figs. 1B-D. All incidence maps were created using the 32 closest point indicator data and, for AAP kriging, the rates recorded in census tracts that share a boundary or vertex with the tract including the interpolation node (1st order adjacency). Incorporating census-tract information through residual kriging adds more details to the map but generates discontinuities at the tract boundaries. On the other hand, accounting for adjacent areal data in AAP kriging leads to a map with more compact spatial features than the indicator kriging map.

The performance of the proposed approach, relatively to ordinary kriging or a traditional residual kriging with choropleth map trend model (e.g. constant value within each polygon), was assessed using jackknife. Performance criteria included the magnitude of prediction errors, the accuracy of the model of uncertainty, the smoothness of interpolated maps, and the ability to discriminate between early and late-stage cancer cases. Results (Goovaerts, 2010) demonstrated the overall better prediction performance of AAP kriging over ordinary kriging and residual kriging. In particular when sampling is sparse, incorporation of areal data improves the prediction accuracy while the exactitude property of areal data decreases the smoothness of interpolated surfaces.

4. Conclusions

The ability to combine data measured at various scales and over different spatial supports in kriging is becoming a pressing need, in particular as the field of geostatistical applications now encompasses social and health sciences. Whereas the first analytical developments of kriging clearly demonstrated its flexibility to accommodate different measurement and prediction supports, geostatistical analysis of a mixture of point data and irregular blocks has rarely been implemented in practice, mainly because of its lack of application in mining. Joint advances in GIS software and computational resources now allow the application of kriging to the complex geographies found in social and health sciences (Goovaerts, 2009). In addition, the recent development of binomial and Poisson kriging allows one to take into account both the spatial extent of the geographical unit and the size of the population under study within that unit (i.e. number of breast cancer cases) in the interpolation.

5. Acknowledgements

This research was funded by grants R43CA150496-01 and R44CA132347-02 from the National Cancer Institute. The views stated in this publication are those of the authors and do not necessarily represent the official views of the NCI.

6. References

- Kyriakidis P, 2004, A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, 36:259-289.
- Goovaerts P, 2008, Kriging and semivariogram deconvolution in presence of irregular geographical units. *Mathematical Geosciences*, 40:101-128.
- Goovaerts P, 2009, Medical geography: a promising field of application for geostatistics. *Mathematical Geosciences*, 41:243-264.
- Goovaerts P, 2010, Combining areal and point data in geostatistical interpolation: Applications to soil science and medical geography. *Mathematical Geosciences*, 42:535-554.
- Goovaerts P, 2011, A coherent geostatistical approach for combining choropleth map and field data in the spatial interpolation of soil properties. *European Journal of Soil Science*, in press.
- Webster R, Oliver MA, Muir KR and Mann JR, 1994, Kriging the local risk of a rare disease from a register of diagnoses. *Geographical Analysis*, 26:168-185.

Reducing aggregation error in spatial interaction models by location sampling

A. Hagen-Zanker, Y. Jin

Department of Architecture, University of Cambridge, 1-5 Scroope Terrace, Cambridge CB2 1PX, UK
 Telephone: +44(0)1223 330573 / 760112
 Email: ahh34@cam.ac.uk / yj242@cam.ac.uk

1. Introduction

Models of spatial interaction such as transport, migration, commuting and trade usually partition space into zones, to represent the receiving and sending end of the interaction. When zones encompass multiple locations, the partitioning causes an aggregation error (Hillsman and Rhoda 1978). The aggregation error increases with the size of zones. Aggregation errors can cause bias (Goodchild 1979; Openshaw 1984) and when zones are larger than a (generally unknown) threshold, models become invalid (Tobler 1989). It therefore seems obvious to make zones smaller whenever possible. In practice, however, zones often remain large for a number of reasons, including data availability, parsimony and computational complexity.

There are different aspects to the aggregation error; there is the information loss associated with averaging variables and the loss of spatial precision – typically by conceptually concentrating all of a zone in its centroid. Both types of error are amplified when non-linear functions are applied on the aggregated variables, which can lead to a further model bias. One domain where non-linear use of aggregated variables causes a risk of bias is Discrete Choice Modelling where the utility of an alternative is typically an exponential function of descriptive variables. It is therefore well-recognized that aggregation of alternatives must account for the effect of size and variability of those alternatives. However size and variability are often imperfectly understood and the analysis has to depend on judgment, experience and proxy variables (Ben-Akiva and Lerman 1985 p. 252-275). In recent years (micro)simulation has been established as a method for aggregation that circumvents many of the complications of analytical solutions (Train 2009). The location variation however, is not usually considered in simulation applications. For instance Train (2009 p. 55) suggests that alternatives with a geographical dimension require utility parameters specified in a log function to facilitate analytical aggregation. This paper intends to follow the simulation approach and extend it to the issue of geographical aggregation.

2. Method

The model that will be used to test the approach is a doubly-constrained model of commuting. The general doubly constrained model has the following form:

$$T_{ij} = a_i b_j P_{ij}, \quad (1)$$

where T_{ij} is interaction between origin zone i and destination j , in this case the number of commuting trips. P_{ij} is the prior distribution of interaction from i to j . a_i and b_j are balancing factors, whose values are determined by the constraints respectively at the origin and destination zone. Balancing factors a_i and b_j are chosen such that:

$$R_i = \sum_j T_{ij} \quad \text{and} \quad C_j = \sum_i T_{ij}, \quad (2)$$

where R_i is the constraint for the i -th row and C_j is the constraint for the j -th column, which also implies $\sum R_i = \sum C_j$. Balancing factors are typically found by iteratively applying the following equations (Fratar 1954):

$$a_i = \frac{R_i}{\sum_j b_j P_{ij}}, \quad b_j = \frac{C_j}{\sum_i a_i P_{ij}}. \quad (3)$$

The prior distribution expresses the ‘gravity’ nature of the model, it is defined as follows:

$$P_{ij} = O_i D_j e^{-\beta d_{ij}}, \quad (4)$$

where O_i is the size of origin zone i and D_j is the size of destination zone j . In the case of commuting, origin size is the working residents and destination size is the number of workplaces. d_{ij} is the distance between zones i and j and parameter β the sensitivity to distance.

The doubly constrained model is linear except for the exponential function of distance. The simulation approach will therefore focus on that function. In the traditional approach the prior distribution is calculated on the basis of mean distance between zones:

$$P_{ij}^{\text{traditional}} = O_i D_j e^{-\beta \bar{d}_{ij}}, \quad (5)$$

where mean distance is the distance between zone centroids, with the intrazonal distance being approximated by the ‘internal radius’:

$$\bar{d}_{ij} = \begin{cases} \|c_i - c_j\| & \text{if } i \neq j \\ \sqrt{A_i / \pi} & \text{if } i = j \end{cases}, \quad (6)$$

where c_i is the centroid of zone i and A_i is some measure of the land area of zone i .

This paper proposes the following alternative:

$$P_{ij} = O_i D_j \frac{1}{n} \sum_{i=1}^n e^{-\beta d_{ijn}}, \quad (7)$$

where d_{ijn} is the n -th random sample of distance between locations in zones i and j :

$$d_{ijn} = \|p_{in} - p_{jn}\|, \quad (8)$$

where p_{in} and p_{jn} are random locations within respectively zones i and j . The random locations are drawn from a uniform spatial distribution: a random location in a zone is found by a series of geometrical operations on the polygon that outlines the zones; First the polygon is decomposed into triangles using a dedicated triangulation library (Shewchuk 1996); Next one triangle is randomly selected using the area of each triangle as the weight; Finally a point is found within the selected triangle by applying the algorithm of Turk (1990).

3. Case study and results

The model is applied on commuting data of England as measured by the U.K. Census of 2001 at the level of Standard Table Wards (‘wards’ from here) as well as Local Authority Districts (‘districts’ from here). The data used is available from Centre for Interaction Data Estimation and Research (<http://cider.census.ac.uk>). Wards form the most detailed geography at which Census commuting data is made available. Districts present a more aggregated geographical level at which practical policy analysis is often carried out. There are 354 districts and 7932 wards in England. The digital boundaries (as polygons) of districts and wards come from UK Borders

(<http://edina.ac.uk/ukborders/>). The centroids of zones are calculated as their geometric centre. Fig. 1 presents the ward and district geographies.

The model has been calibrated twice, with both versions of priors (i.e. equations 5 and 7). A bracketing approach called Golden Section Search (Press 1992) was followed to find the value of β that minimized the following error:

$$\delta = \sum_{i,j} (T_{ij}^{model} - T_{ij}^{census})^2, \quad (9)$$

where δ is the discrepancy between modelled and actual (Census) commuting matrices.

Table 1 gives estimated values for β and the associated error δ . It shows that for the case of wards it makes little difference which approach is chosen, but for districts there is a marked difference in performance where the simulation based model performs 35% better than the traditional model. The graphs in fig. 2 depict the trip distribution as a function of distance and confirm the difference in performance.

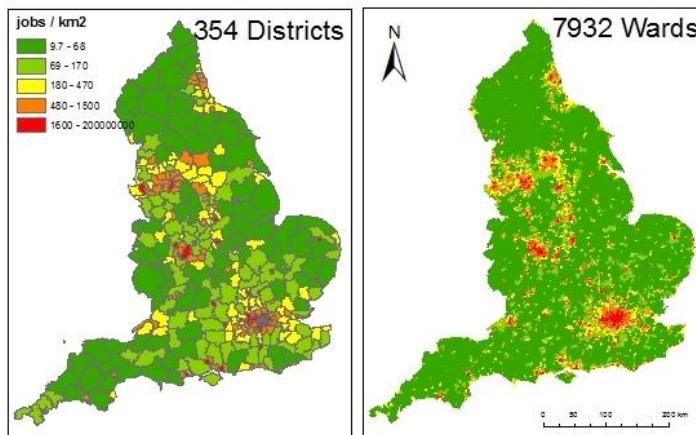


Figure 1. Study area England at district (left) and ward (right) levels of aggregation.

Geography	Model	β	$\delta(*10^9)$
Wards	Traditional	0.34	2.99
Wards	Simulation	0.36	2.87
Districts	Traditional	0.37	90
Districts	Simulation	0.31	58

Table 1. Calibration results and errors. Note that errors are only comparable between models applied at a common geography.

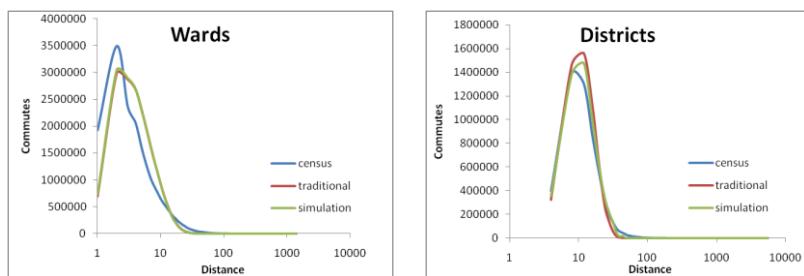


Figure 2. Census and modelled trip distributions. Note zone sizes distort distribution patterns particularly at the district level.

4. Conclusion

This paper follows up on the recommendation of Train (2009) and others to employ simulation when faced with discrete choice models for which analytical models are not feasible or too restrictive. The case study is carried out on the generic doubly-constrained model, which is readily generalisable to more sophisticated random utility models.

By comparing two cases that differ in the level of spatial aggregation it became clear that location sampling does significantly reduce the error caused by using average distances. At the fine scale of wards the effect of error reduction is small although still apparent. At the coarser scale of districts however, simulation would seem essential in future models to contain the aggregation error.

Simulation can be a mechanism for reliable modelling on the basis of coarse scale data when fine scale data is not available. An example of such data is the UK Census commuting data that only offers thematically refined data at coarse spatial scales, for instance commuting patterns specified by industry and socio-economic group which allow segmented modelling of commuter behaviour.

5. Acknowledgments

Geographical boundary data is provided through EDINA UKBORDERS with the support of the ESRC and JISC and uses boundary material which is Crown copyright. Census data (Special Workplace Statistics Levels 1 and 3) is Crown copyright and reproduced with permission of the Controller of HMSO and the Queen's Printer for Scotland. This work is part of the EPSRC Energy Efficient Cities Project.

6. References

- Ben-Akiva ME and Lerman SR, 1985, *Discrete choice analysis : theory and application to travel demand*. MIT Press, Cambridge.
- Fratar T, 1954, Vehicular trip distribution by successive approximation. *Traffic Quarterly*, 8(1):53-65.
- Goodchild M, 1979, The aggregation problem in location-allocation. *Geographical Analysis*, 11(3):240-255.
- Hillsman EL and Rhoda R, 1978, Errors in measuring distances from populations to service centers. *The Annals of Regional Science*, 12(3):74-88.
- Openshaw S, 1984, Ecological fallacies and the analysis of areal Census-data. *Environment and Planning A*, 16(1):17-31.
- Press WH, 1992, *Numerical recipes in C : the art of scientific computing*, 2nd ed. Cambridge University Press, Cambridge ; New York.
- Shewchuk J, 1996, Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator. In: Lin MC and Manocha D (eds), *Applied Computational Geometry Towards Geometric Engineering*. Springer, Heidelberg; Berlin, 203-222.
- Tobler WR, 1989, Frame independent spatial analysis. In: Goodchild MF and Gopal S (eds), *Accuracy of Spatial Databases*. Taylor & Francis, London ; New York, 115-122.
- Train K, 2009, *Discrete choice methods with simulation*, 2nd ed. Cambridge University Press, Cambridge ; New York.
- Turk G, 1990, Generating random points in triangles. In: Andrew SG (ed) *Graphics gems*. Academic Press Professional, 24-28.

Incorporating Environmental data into Poisson Kriging Approaches for Mapping Patterns of Herbivore Species Abundance in Kruger National Park, South Africa

Kerry, R.¹, Goovaerts, P.², Smit, I.³ and Ingram, B. R.⁴

¹Department of Geography, Brigham Young University, Provo, Utah, USA.
ruth_kerry@byu.edu

²Biomedware Inc., Ann Arbor, Michigan, USA.
goovaerts.pierre@gmail.com

³South African National Parks, Skukuza, South Africa.
izaks@sanparks.org

⁴Department of Computer Science, University of Talca, Talca, Chile.
ingrambr@googlemail.com

1. Introduction

Kruger National Park, South Africa, provides 19,485 km² of protected habitats for the unique species of the African savanna, several of which are endangered. For the last forty years annual aerial surveys to monitor large herbivore populations have been conducted. These have been used to understand population trends and the environmental factors and management actions that influence herbivore density and distribution patterns. From 1980-1993, the whole park was surveyed annually, but this was costly and time consuming. In 1998, the park-wide census approach was replaced by a sampling strategy whereby the number of animals is recorded along 800 m wide East-West transects, spaced at intervals of 2.5-5.6 km (Kruger et al. 2008). However, such strip transects leave “gaps” in the data spatially. The park currently use the Distance method (Thomas et al. 2004) but several assumptions of the method are not met especially for rare species or species that tend to be clustered in space.

Geostatistical methods at first glance might seem ideal for populating the gaps in survey data and for estimating the total numbers of each animal in the park in a given year. However, the histogram of animal count data for the park is usually highly positively skewed, especially for the rarer species or those that tend to cluster spatially. The histograms tend to approach the Poisson distribution. This hampers the estimation of the variogram by the traditional method of moments. Kerry et al. (2010a) compared an Auto-Indicator kriging approach (Goovaerts, 2009) and Poisson kriging (Monestiez et al. 2006) as potential methods for populating the data gaps between transects and to create continuous surfaces of species abundance. It was thought that an auto-indicator approach could be used to efficiently compute and model variograms for numerous thresholds representing each count. However, the study showed that variograms for the rare high counts were pure nugget and so the number of thresholds had to be reduced. This meant that the number of large counts of each animal was under-estimated. Also, the nature of the data meant that unless the data were preprocessed to migrate the data to a grid, there were no zero counts and this meant that low counts were over-estimated. In contrast to the Auto-Indicator approach, with and without pre-processed data, two Poisson approaches produced markedly smaller, and sometimes an order of magnitude smaller, mean absolute errors (MAEs) in cross-validation. An initial investigation showed that incorporating environmental data into a simple 0/1 Indicator approach reduced MAEs slightly. Here we illustrate a method of incorporating environmental data into the Poisson kriging approaches and compare the errors associated with this to the errors when no environmental data are included.

2. Methods

Poisson kriging of count data was performed using two types of denominator:

- (1) observational area (ratio = spatial density, Fig. 1a)
- (2) total number of animals in a given area (ratio = proportion, Fig. 1b).

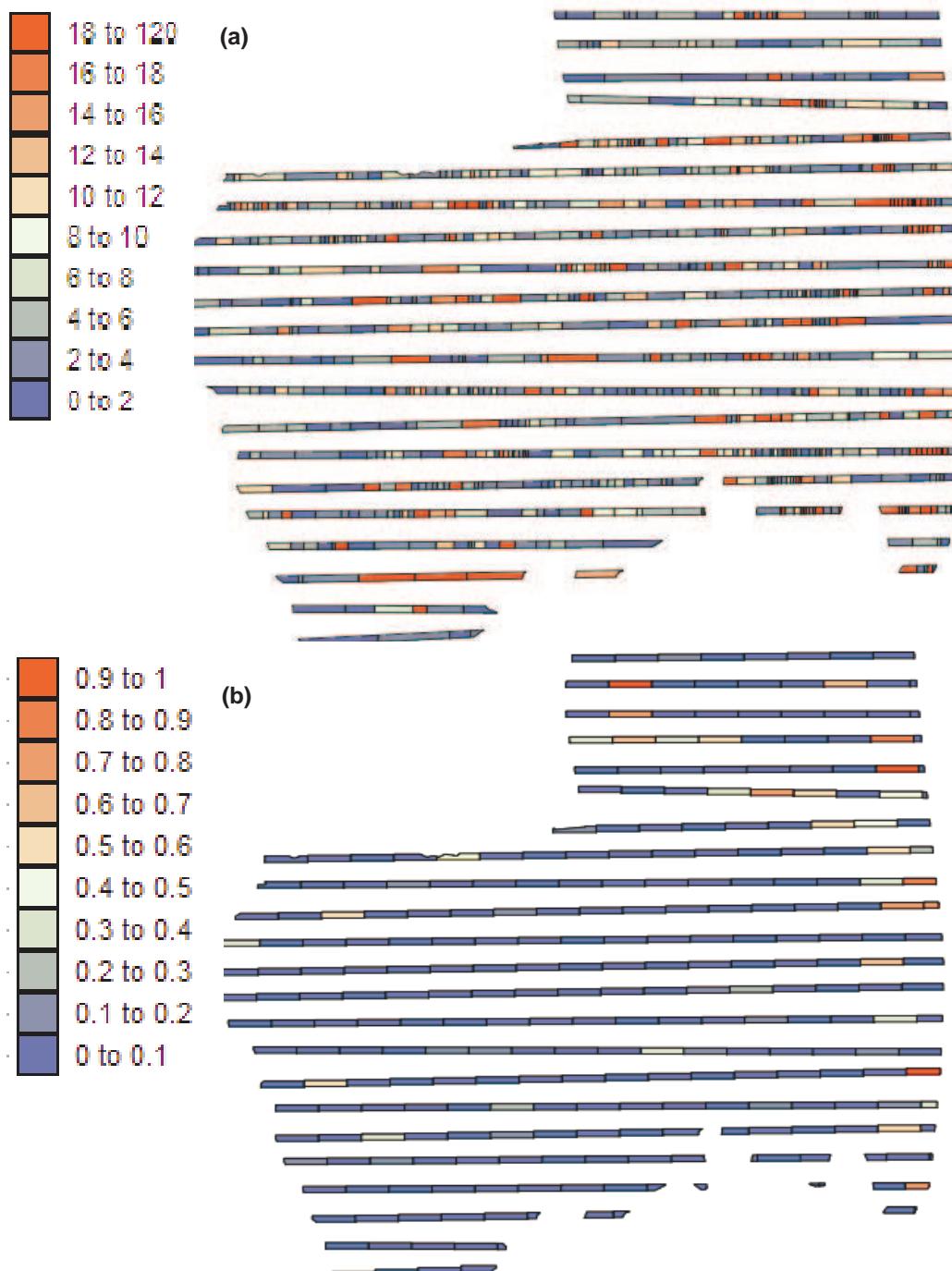


Figure 1. Calculation of (a) spatial density from 800 m wide transect data for Poisson approach (1) and (b) proportion of each animal from 5 km long blocks of the 800 m wide transect data (e.g. number of impala/total number of all animals in 5km by 800m block) for Poisson approach (2).

Both Poisson approaches result in sightings of rare animals in sparsely populated areas (i.e. small numbers) being down-weighted for variogram computation and kriging. However, Approach (2) is only suitable for accurately mapping the distribution of individual species in the park.

The following observational area/population-weighted estimator adjusts for the small number problem:

$$\hat{\gamma}_{Rv}(\mathbf{h}) = \frac{1}{2 \sum_{\alpha, \beta}^N \frac{n(v_\alpha)n(v_\beta)}{n(v_\alpha) + n(v_\beta)}} \sum_{\alpha, \beta}^N \left\{ \frac{n(v_\alpha)n(v_\beta)}{n(v_\alpha) + n(v_\beta)} [z(v_\alpha) - z(v_\beta)]^2 - m^* \right\}, \quad (1)$$

where $N(\mathbf{h})$ is the number of pairs of areas (v_α, v_β) whose observational area/population-weighted centroids are separated by the vector \mathbf{h} and m^* is the observational area/population-weighted mean of the N area rates. The usual squared differences, $[z(v_\alpha) - z(v_\beta)]^2$, are weighted by a function of their respective observational area/population sizes, $n(v_\alpha)n(v_\beta)/[n(v_\alpha) + n(v_\beta)]$, which gives more importance to more reliable data pairs based on large observational areas/large total counts of animals (Monestiez et al. 2006, see also Kerry et al. 2010b).

The animal density/proportion and the associated kriging variance for a location X are estimated as:

$$\hat{r}(X) = \sum_{i=1}^K \lambda_i z(v_i), \text{ and} \quad (2)$$

$$\sigma^2(X) = \bar{C}_R(X, X) - \sum_{i=1}^K \lambda_i \bar{C}_R(v_i, X) - \mu(X), \quad (3)$$

The kriging weights (λ_i) and the Lagrange parameter $\mu(X)$ are computed by solving the “Poisson kriging” system:

$$\begin{aligned} \sum_{j=1}^K \lambda_j \left[\bar{C}_R(v_i, v_j) + \delta_{ij} \frac{m^*}{n(v_i)} \right] + \mu(X) &= \bar{C}_R(v_i, X), \quad i = 1, \dots, K, \\ \sum_{j=1}^K \lambda_j &= 1, \end{aligned} \quad (4)$$

where $\delta_{ij}=1$ if $i=j$ and 0 otherwise. The covariances are estimated from the results of a deconvolution of the model fitted to variogram (1), see Goovaerts (2008). The “error variance” term, $m^*/n(v_i)$, leads to smaller weights for rates measured over smaller areas/populations.

Various environmental data (Figure 2) was incorporated into the mapping by kriging the residuals from a Poisson regression between environmental and animal data.

Leave-one-out (LOO) cross-validation was used to assess the relative performance of the different methods for estimating counts of all species, and of representatives of the key feeding groups grazers, browsers and mixed feeders such as giraffes, impala and zebra for the whole park. Rarer species or those species that tend to cluster in herds were also investigated.

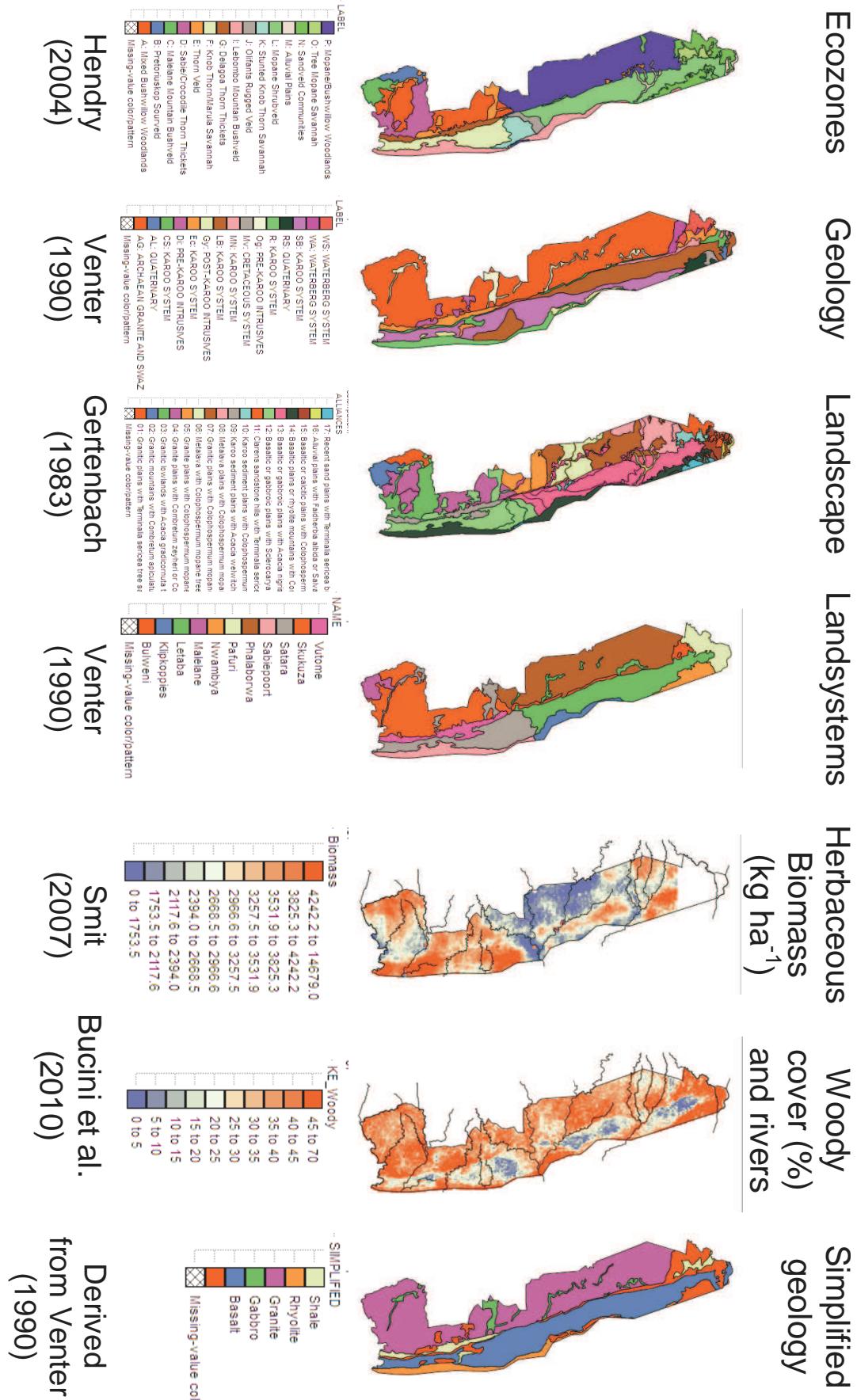


Figure 2. Environmental Data used in Poisson Regression

3. Results and Discussion

Table 1 shows that for estimating numbers of giraffe, impala and zebra, Poisson approach (1) yielded smaller errors. It also created patterns that are more sensible than those of Poisson approach 2 when compared to the observed counts (Figure 3). The MAEs indicate that approach (1) produces its best estimates when there are more animals i.e. looking at counts of all animals, or more abundant animals such as impala. Poisson approach (2), however, leads to best estimates for the rarest animals or those that tend to occur in isolated herds. The effects on MAEs of incorporating environmental data such as biomass, tree cover, geology and ecotypes into both Poisson approaches (results not shown here) will be discussed in the presentation.

Table 1. Mean Absolute Errors (MAEs) from Leave-One-Out Cross-validation for Poisson kriging using spatial density (approach 1) or proportion of animals (approach 2).

Data	MAE	
	Poisson approach (1)	Poisson approach (2)
All animals 1998	0.0528	*
All animals 2000	0.0401	*
All animals 2001	0.0463	*
All animals 2005	0.0448	*
<i>Key feeding groups</i>		
Giraffe 2000	0.1337	0.1481
Impala 2000	0.1668	1.3791
Zebra 2000	0.2630	0.3909
<i>Rarer species</i>		
Elephant 2000	0.4494	0.0996
Kudu 2000	0.6264	0.1142
Waterbuck 2000	0.7747	0.0516
Warthog 2000	0.8165	0.0460
Wildebeest 2000	0.3133	0.1290
White rhino 2000	0.5348	0.0850

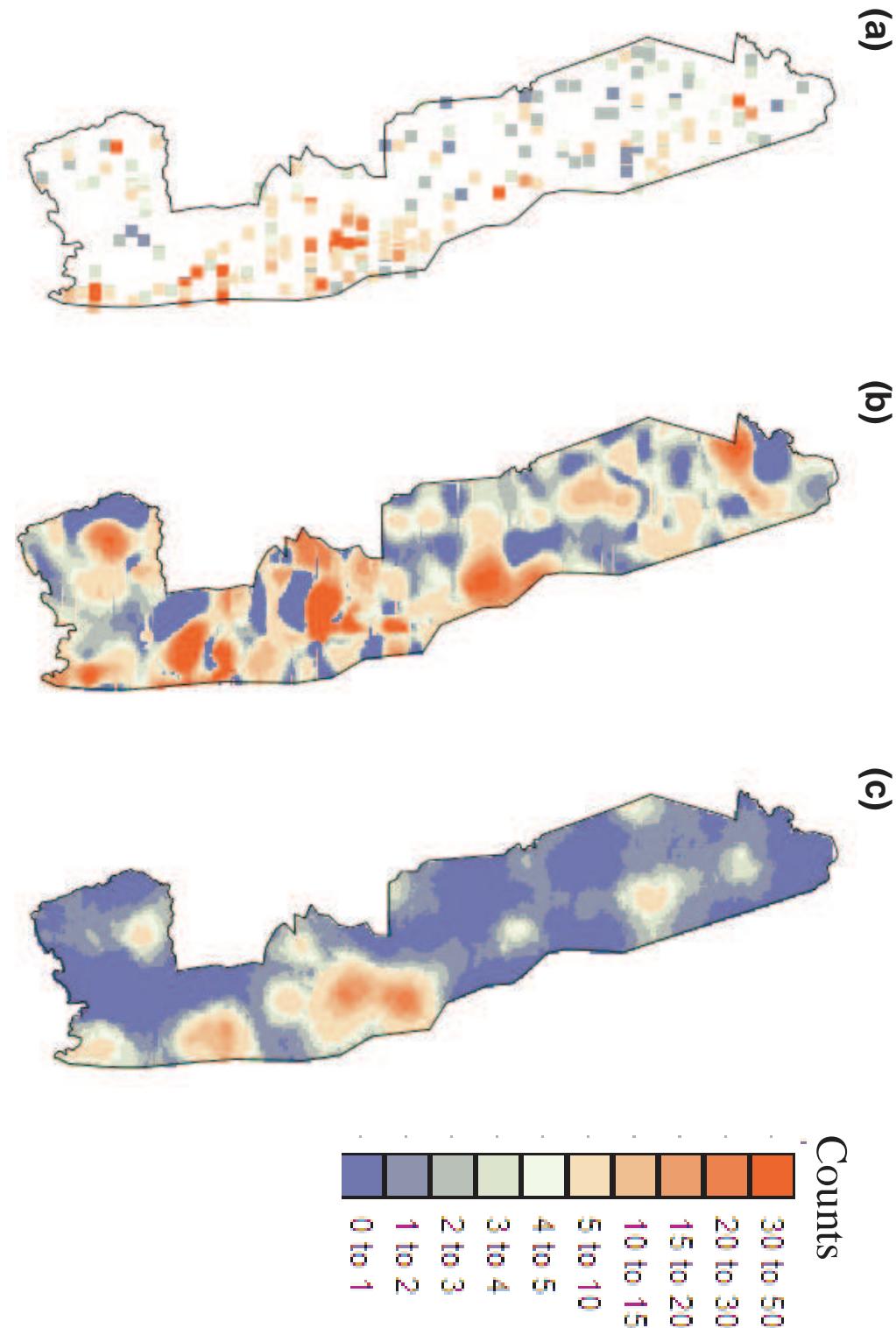


Figure 3. (a) Observed counts of zebras in 2000 and kriged maps of counts produced by (b) Poisson approach (1), and (c) Poisson approach (2).

5. References

- Bucini, G, Hanan, NP, Boone, RB, Smit, IPJ, Saatchi, S, Lefsky, MA & Asner, GP, 2010. Woody fractional cover in Kruger National Park, South Africa: remote-sensing-based maps and ecological insights. *Ecosystem function in savannas: measurement and modeling at landscape to global scales* (ed. Hill, M.J. & Hanan, N.P.), CRC/Taylor and Francis, pp 219-237.
- Gertenbach, WPD, 1983. Landscapes of the Kruger National Park. *Koedoe*, 26:9-121.
- Goovaerts P, 2009. AUTO-IK: a 2D indicator kriging program for the automated non-parametric modeling of local uncertainty in earth sciences. *Computers & Geosciences*, 35:1255-1270.
- Goovaerts, P, 2008. Kriging and semivariogram deconvolution in presence of irregular geographical units. *Mathematical Geosciences*, 40(1), 101-128.
- Hendry O. 2004. *Kruger Park Ecozone Map*. Jacana Media. 16p.
- Kerry R, Goovaerts P, Haining RP and Ceccato V 2010b. Geostatistical Analysis of Car Theft and Robbery in the Baltic States. *Geographical Analysis*. 42:53-77.
- Kerry R, Goovaerts P, Smit I and Ingram BR, 2010a. Comparing the Accuracy of Indicator and Poisson Kriging for Investigating Patterns of Herbivore Species Abundance in Kruger National Park, South Africa. *Proceedings of 9th Ninth International Symposium on Spatial Accuracy Assessment in Natural resources and Environmental Sciences, Leicester, United Kingdom, July 2010*.
- Kruger JM, Reilly, BK and Whyte IJ, 2008. Application of distance sampling to estimate population densities of large herbivores in Kruger National Park. *Wildlife Research*, 35:371–376.
- Monestiez P, Dubroca L, Bonnin E, Durbec JP and Guinet C, 2006. Geostatistical modelling of spatial distribution of *Balaenoptera physalus* in the Northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts. *Ecological Modelling*. 193:615–628.
- Smit IPJ, 2007. *Artificial surface-water provision in a semi-arid savanna: a spatio-temporal analysis of herbivore distribution patterns in relation to artificial waterholes under different habitat, rainfall and management scenarios in the Kruger National Park, South Africa*, Unpublished PhD thesis, University of Cambridge, Cambridge.
- Thomas L, Laake JL, Strindberg S, Marques FFC, Buckland ST, Borchers DL, Anderson, DR, Burnham KP, Hedley SL, Pollard JH and Bishop JRB 2004. *Distance 4.1. Release 2. (Research Unit for Wildlife Population Assessment*, University of St Andrews: St Andrews, UK.
- Venter FJ, 1990. *A classification of land for management planning in the Kruger National Park*. Unpublished PhD thesis, University of South Africa, Pretoria.

Evaluation of Geostatistical Analysis Capability in Wireless Signal Propagation Modeling

Samira Kolyaei^{1, 3}, Marjan Yaghoobi^{2, 3}

¹ MSc. Graduate, GIS division, Dept. of Surveying and Geomatics Eng., College of Eng., University of Tehran

Telephone: +98-9352000174

Email: samira.kolyaei@gmail.com, kolyaei@ut.ac.ir

² BSc. Graduate, Dept. of Surveying and Geomatics Eng., College of Eng., University of Tehran

Telephone: +98-9352000569

Email: m.yaghoobi@gmail.com

³ MTN Irancell, Network Group, Radio Division, GIS Team

1. Introduction

Radio signal path loss is a particularly important element in the design of any radio communication system or wireless system, and it is necessary to be able to determine the levels of the signal loss for a give radio path. The more accurate model the better decision making will occur for the network rollout, planning and optimization. The radio signal path loss could be estimated by many elements of the radio communications system in particular the transmitter power, and the antennas, especially their gain pattern, height, azimuth and also are highly influenced by clutters (land use), terrain height and morphology, spatially related parameters. Okumura-Hata empirical model, which has many parameters, is the most common model for prediction and estimation of this complicated phenomenon.

On the other hand, geostatistical techniques offer interpolation methods for describing the continuity of spatially/temporally variable data which is essential feature of many natural phenomena. Over the last decades, this theoretical framework has been successfully applied in other type of spatial problems (Konak, 2010). As it is mentioned above signal path loss is highly influenced by spatial parameters, therefore geostatistics has high potential to be implemented for such purpose (Arpee J. et al., 2000).

There are few researches conducted using geostatistical techniques for modelling wireless propagation models. In a recent paper, Konak (2009) reports that ordinary kriging is competitive with radial basis ANNs to estimate the signal-to-noise ratio in cellular wireless networks. Konak in 2010 extends the ordinary kriging approach proposed in 2009 by considering path loss due to obstacles and other factors in indoor environments. In this paper we have compared the result of different spatial and non-spatial interpolation techniques with Okumura-Hata empirical model to evaluate geostatistical analysis capability to find a way of having more accurate mobile coverage models. As the result, this research will help us to have RF path loss trend estimation which is extracted from sampled data to describe data variability across the entire cellular system areas.

The area of interest is a city in west of Iran. Okumura-Hata, Inverse Distance Weighting (IDW- a non-spatial interpolation technique) with different power and number of neighbour and different type of kriging (ordinary & universal) with different semi-variogram model and

number of neighbours are used to model RF propagation in study area. Finally, the result of all interpolation methods are compared using the check points of real data.

2. Research background

In this section, a brief description of Okumura-Hata empirical model and also geostatistical interpolation technique, kriging, are presented.

2.1. Okumura-Hata empirical model

Path Loss (L) is a measure of the reduction in power density of an electromagnetic wave as it propagates through space (Konak, 2010). The method analyzes raw RF power data that is collected by drive testing a sample of roads in a cellular system. Radio Propagation predictions are mainly used to demonstrate how the mobile signals are scattered in the environment as well as how strong the signals are in different places. however, there are various factors influencing the radio propagation prediction accuracy such as reflection, diffraction, scattering, transmission, refraction, etc.

This method is used for cellular system planning and management, the one which is used is the standard macrocell model which is based upon the Okumura-Hata empirical model with a number of additional features to enhance its flexibility. The model has a number of features that enhance its flexibility and accuracy such as the inclusion of clutter offsets and heights and the use of diffraction. The Okumura-Hata model has the following validity range:

- The distance from the site between 500 m and 30 Km
- Antennas height in the range of 15-200m
- Receiver heights in the range of 1-10m
- Frequency: 150...1000 MHz and 1500...2000 MHz
- MS height: 1 m...10 m

The data needed for computation using this data are in two categories, (a) data related to the antenna such as antenna gain, pattern, height, azimuth, power, etc (b) mapping data including terrain DTM and terrain clutter (land use).

The model has a large number of parameters and options which may be selected or calibrated by the user in order to obtain a close representation to measured propagation data. Since network simulations are very time consuming, the choice of the macrocell propagation model is a trade off between prediction accuracy and computational efficiency. For this reason the standard Okumura-Hata macrocell model has been chosen. The basic equation used in the path loss calculation is given as follows (Equation 1, Table 1) (Aircom Tech Doc, 1999):

$$PL(d) = K1 + K2 \cdot \log(d) + K3 \cdot Hms + K4 \cdot \log(Hms) + K5 \cdot \log(Heff) + K6 \cdot \log(Heff) \cdot \log(d) + K7 \cdot Ldiff_Losses + Lclutter_Losses \quad (\text{Equation 1})$$

Where:

k1 & k2	Intercept and Slope. These factors correspond to a constant offset (in dBm) and a multiplying factor for the log of the distance between the base station and mobile.
k3	Mobile Antenna Height Factor. Correction factor used to take into account the effective mobile antenna height.
k4	Okumura-Hata multiplying factor for Hms.

k5	Effective Antenna Height Gain. This is the multiplying factor for the log of the effective antenna height.
k6	Log (Heff)Log(d). This is the Okumura-Hata type multiplying factor for log(Heff)log(d).
k7	Diffraction. This is a multiplying factor for diffraction calculations. A choice of diffraction methods is available.
P_L	the path-loss in dB
d	the distance between the BS and the MS in meters
h_{MS}	the height of the MS in meters
h_{eff}	the BS effective antenna height in meters
Diff_Losses	the diffraction losses and Clutter_Losses are the losses associated with the clutter types.

Table 1 - Model Parameters

2.2. Geostatistics

Kriging is a geostatistical technique to interpolate the values of a random spatially/temporarily field value in an unobserved location from observed nearby locations. It was developed by Krige (1951) and Matheron (1963) to accurately predict ore reserves from the samples taken over a mining field. There are different types of kriging such as ordinary, universal, indicator, disjunctive kriging [Krivoruchko K., 2001]. In Kriging, the prediction is based on semivariogram which is function of distance and is highly dependent of researcher's experience (Yoo S., 1994). There are different models which are used for modeling semivariogram such as spherical, gaussian, exponential and circular (Zimmerman DL et al., 1991)

3. Method

This paper is a part of an ongoing research to test and use the capability of geostatistical analysis for coverage predicting, simulating and tuning. The goal of this research is to test different interpolation techniques in coverage prediction. Method presented here, analyzes raw data that is collected by drive testing a sample of roads in study area (Figure 1). First this data were processed to eliminate gross errors and duplicate values. Here 58029 point data are collected.

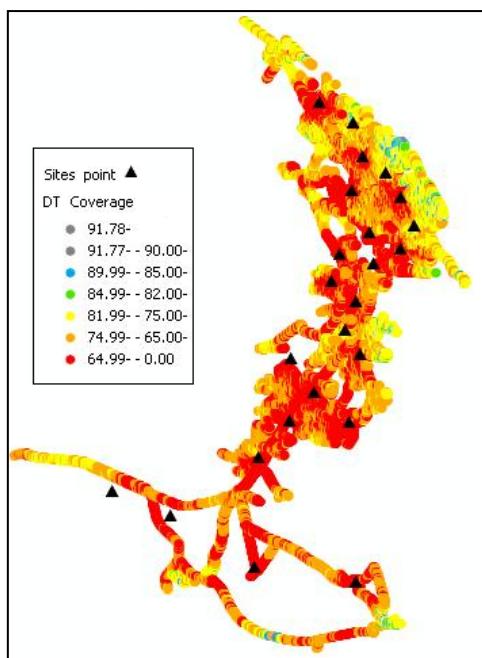


Figure 1- raw data that is collected by drive testing a sample of roads in study area

Different continuous surface are created using different interpolation methods. To compare the interpolation methods used here, we have selected 15417 as check points, to be able to compare the predicting values of different type of techniques. Although kriging, could be evaluated by error prediction analysis which is one of the advantage of this technique but we used check points to be able to compare the result of this technique with non-spatial interpolation technique. To have check points in different locations, they are selected by creating a grid network (Figure 2). So 42612 points are used for prediction and 15417 points are used as check points.

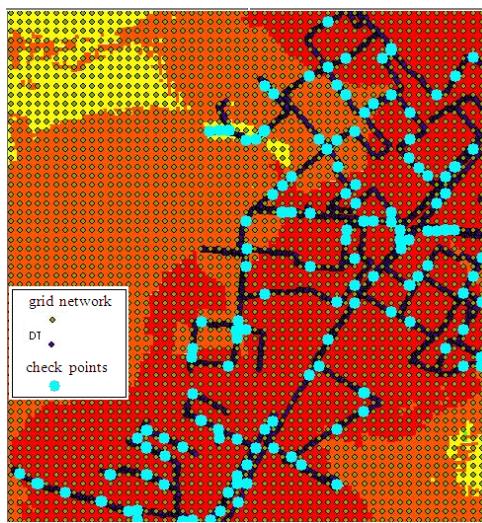


Figure 2- Check points selection using grid network

After the steps of data preparation, different surface using Okumura-Hata model with optimum factors and different interpolation methods are created. The interpolation methods used here are Inverse Distance Weighting (IDW) with different power and number of neighbours, ordinary kriging and also universal kriging with gaussian, spherical, circular and exponential models for semivariogram with differnt number of neighbors. In following figure (Figure 3) surfaces created by some of used methods are depicted.

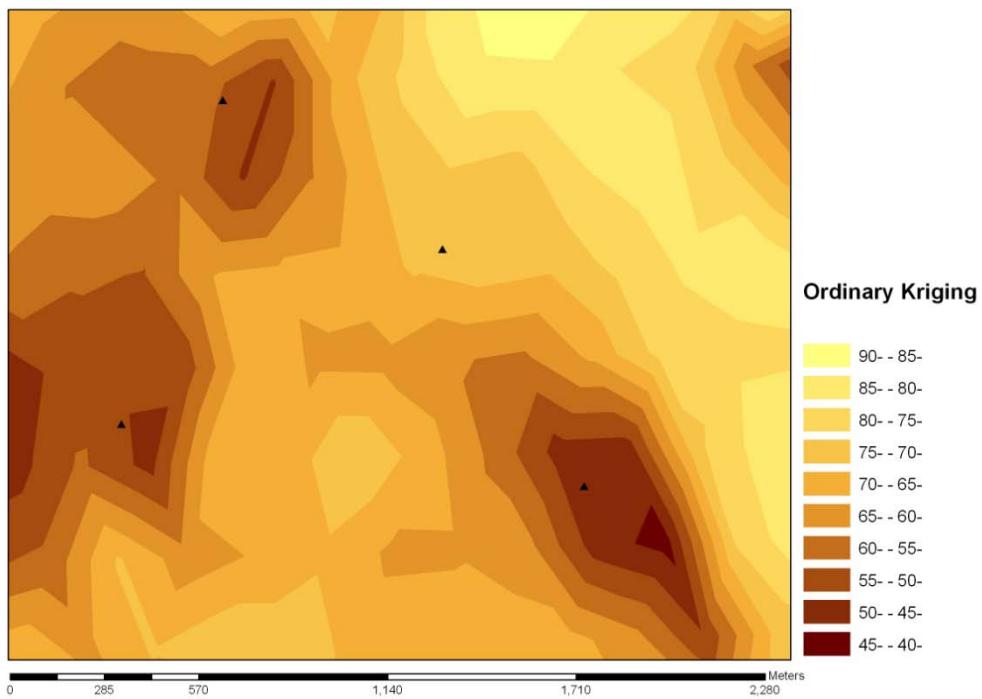


Figure 3a- Surface Created by Ordinary Kriging

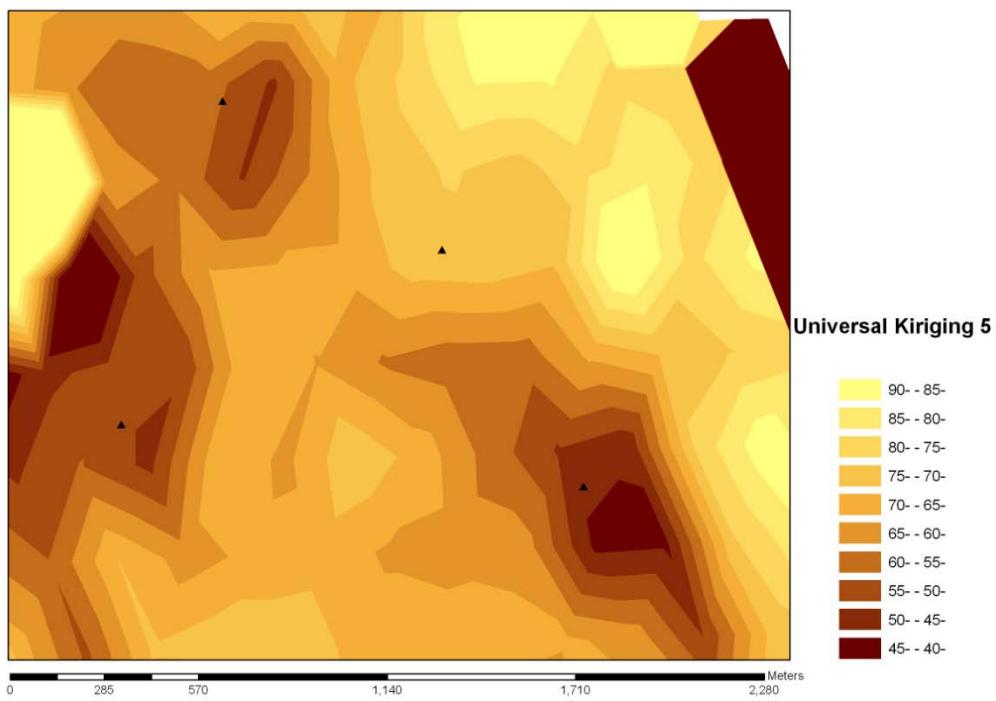


Figure 3b- Surface Created by Universal Kriging

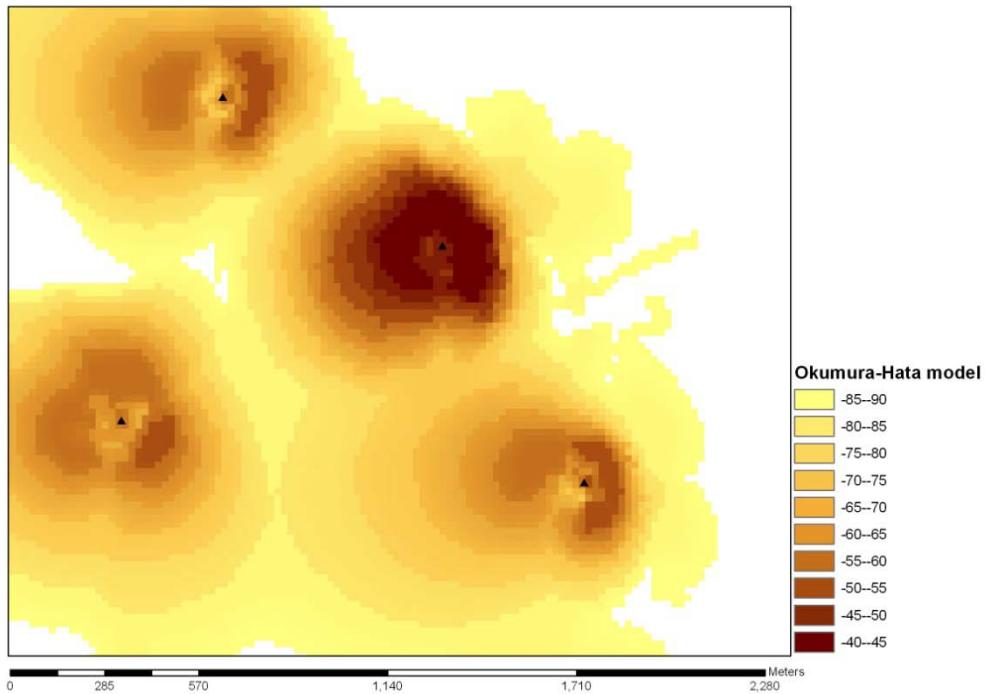


Figure 3c- Surface Created by Okumura-Hata Model

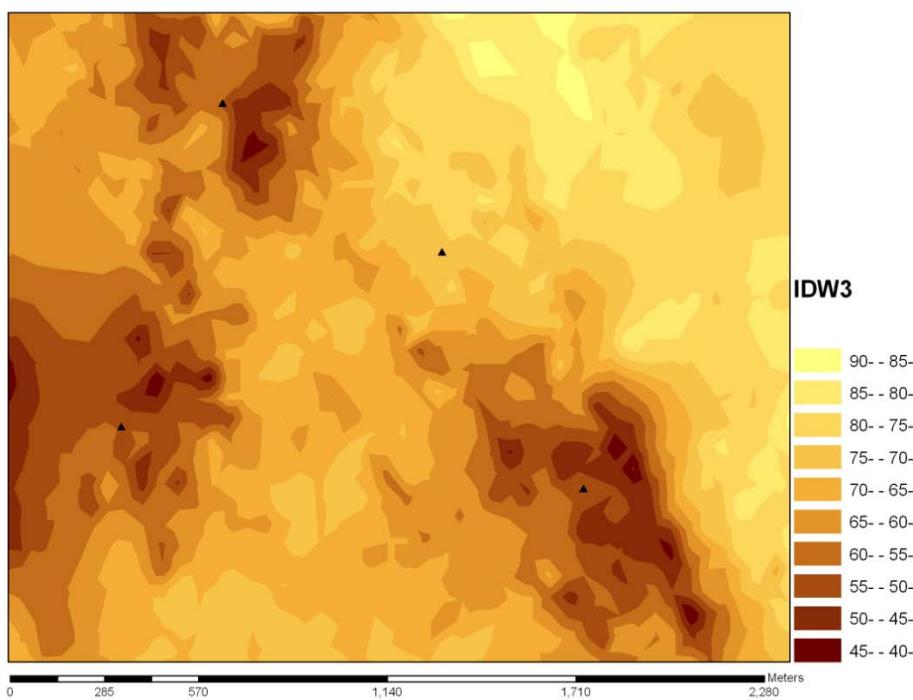


Figure 3d- Surface Created by IDW

For the result comparison, we have used check points. Prediction values for check points are extracted from each surface and the difference with actual value are computed. The root mean square is used to compare the methods (Table¹).

Table¹ - numerical result for comparison of best fitted surfaces of each method

Surface	Model	Detail	σ
S2	idw_3	Power3	4.684433536
S9		Okumura-Hata	4.450834117
S6	Univ_Kriging3	Gaussian	4.0973997
S4	Univ_Kriging5	Circular	3.706218192
S7	Univ_Kriging2	Spherical	3.668627883
S5	Univ_Kriging4	Exponential	3.351953058
S8	Ordinary_Kriging	Exponential	3.307196376

4. Result

The outlines of the results of this study are listed as below:

- Kriging methods predict coverage having acceptable error and are even more accurate than Okumura-Hata with much less than input and computation
- Okumura-Hata methods are more accurate than IDW.
- Surfaces created by universal kriging demonstrate, exponential model for semi-variogram is best fit for coverage prediction among the tested ones. Results show the suitability of exponential and spherical for semi-variogram are almost same and place after exponential model and the last one is gaussian model
- Ordinary kriging is a little more accurate than universal kriging with exponential semi-variogram but the computational time is much more than universal kriging
- Among tested models, considering tradeoff between accuracy and computational time, universal kriging having exponential semi-variogram is best one.

REFERENCES:

AIRCOM Technical Documents, 1999, "Asset Standard Macrocell Model Calibration document", AIRCOM International Ltd

John Arpee, Herndon, Stan Gutowski, Mustafa Touati, 2000, "Apparatus and method for geostatistical analysis of wireless signal propagation", US patent 6,711,404,2004, Technical Aspect of Geostatistics, pp. 2-1 to 2-17

Krige, D. G. 1951. A statistical approach to some basic mine valuation problems on the Witwatersrand. Journal of the Chemical, Metallurgical and Mining Society of South Africa 52: 119_139.

Konak, A. 2009. "A kriging approach to predicting coverage in wireless networks", International Journal of Mobile Network Design and Innovation 3: 64-70.

Konak, A., 2010, "Estimation path loss in wireless local area using ordinary", Proceedings of the 2010 Winter Simulation Conference, B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

Krivoruchko K., 2001., “Using linear and non-linear kriging interpolators to produce probability maps”, Annual conference of the international association for mathematical Geology, Cancun, Mexico, September, IAMG 2001

Matheron, G. 1963., Principles of geostatistics. Economic Geology 58: 1246-1266.

Yoo S., 1994, “A spatial prediction theory for long-term fading in mobile radio communications”, ETRI Journal, Vol 15, No 3/4, Jan 1994

Zimmerman DL, Zimmerman M. B., 1991., “A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors”, Technometrics, Vol. 33, pp. 77-91

Where were you? A time-geographic approach to activity destination re-construction

M. W. Horner¹, J. A. Downs²

¹Florida State University, Tallahassee, FL
Email: mhorner@fsu.edu

²University of South Florida, Tampa, FL
Email: jdowns@usf.edu

1. Introduction

The trend towards spatially disaggregate modelling and computation has manifested itself across a host of scientific fields, particularly that of transportation (Shaw and Wang 2000). With the proliferation of individual-level datasets describing the detailed activities of households and their constituent members, it is possible to analyze human movements at a high degree of precision (Goodchild 2010). Moreover, these data are becoming better in quality as transportation surveys increasingly take advantage of digital technologies (e.g. GPS receivers) to effectively capture information on people's daily activities at spatially dispersed origin and destination locations (Stopher et al. 2007).

However, despite these advances, travel survey data are not without quality issues (Wolf et al. 2001). First, many daily activity travel surveys still use methods that require a sampled individual's reported origin or destination location to be physically geocoded. In other words, the survey data capture method (e.g. paper survey, telephone interview) does not directly collect the exact spatial coordinate of an activity origin and/or destination (e.g. 123 Pine Street, Sunnyville, FL, 31234). Potential exists for origins and destinations of reported trips not to be geo-referenced, perhaps due to mis-reported information, and/or inconsistencies in spatial address databases, which can limit the usefulness of the survey data. Given an account of trips made by an individual during a typical day, the analyst may be able to confirm that a survey respondent took a particular trip, but may be unable determine the exact origin and/or destination of the trip. From a transportation analysis standpoint, this is a serious problem because it in effect renders the record useless, especially in cases where analyzing the chain of activity locations is of interest (Hensher and Reyes 2000, Horner and O'Kelly 2007).

This paper explores a computational approach for recovering unlocateable activity locations from travel surveys. Derived from recent work in time geography, the method reconstructs the most probable location(s) of missing origins and destinations that were unable to be determined via georeferencing procedures. We adapt a recently developed probabilistic time-geographic approach (Downs 2010) that incorporates individuals' known origin and destination locations, and the time they spent at these (and the unmatched) locations.

2. Background and Method

Facilitated by improvements in computational power and geographic information systems (GIS) technology, there has been renewed interest and subsequently many recent developments in the field of time geography. Haagerstrand's pioneering work of the 1970's, which was revisited and extended in the 1990's by Miller and others, set the foundation for the sustained

stream of research that continues today (Haagerstrand 1970, Miller 1991, Kwan 1998, O'Sullivan et al. 2000). Time geographers concern themselves with examining and applying the classical constructs (e.g. space time prisms, cones, geo-ellipses), and research has also proceeded along several other related lines including addressing uncertainty and representations of space, including adapting metrics for use on networks (Kuijpers et al. 2010, Neutens et al. 2011).

One recent area of interest has been in the idea of developing a ‘probabilistic’ time geography (Downs 2010, Winter and Lin 2010, Winter and Yin 2010). In work by Downs (2010), the traditional geo-ellipse representation is improved to visualize likely area(s) a mobile object could have travelled given a time budget. Known as time geographic density estimation (TDGE), the method does not focus solely on the outer polygon depicting the maximum space that could be consumed (i.e., the geo-ellipse), but rather it incorporates a form of data smoothing to interpolate a surface within the polygon showing the most likely places that object could have been located. Of course, this likelihood is determined not only by the individual’s available travel budget but also by their other known points.

Per Downs (2010) the formulation for the TDGE estimator is

$$\hat{f}_t(x) = \frac{1}{(n-1)[(t_n - t_1)v]^2} \sum_{i=1}^{n-1} G\left(\frac{\|x - x_i\| + \|x_j - x\|}{(t_j - t_i)v}\right), \quad (1)$$

where $\hat{f}_t(x)$ is the time geographic density estimate at any point x in a map and G is a distance-weighting function of the geo-ellipse. The number of control points is indicated by n , with each point having a time stamp t . Sequencing of points is governed by the ordering of subscripts i to j . Effectively this formulation in equation 1 fits a distance-weighted geo-ellipse function to each consecutive pair of control points in a space-time path. In this paper we adapt this method to be used with travel survey data, where the interest is in identifying missing origin and destination locations, given other known spatial and temporal information about a respondent’s activity locations.

3. Research Structure

We provide a detailed review of time geography, including related developments with respect to uncertainty and probabilistic issues. We also discuss disaggregate travel methods in regards to transportation surveys and activity analysis. From there, we modify the TGDE approach to work with empirically observed travel survey data from a smaller Midwestern U.S. city. Several adaptations to TDGE are suggested, including incorporating a traditional transportation network structure into its estimation (Neutens et al. 2008). We also compare various approaches for re-creating missing survey points by experimenting with alternative weighting functions as well as exploring the whether using more than two known points does a better job of predicting intermediate unknown locations. To get at the idea of ‘better’ we can simulate missing data simply by dropping out a known intermediate point between two other known activity points, use the TDGE technique to ascertain how well we predict the known point. This will act as a calibration procedure to indicate the preferred combination of parameters for applying TDGE to recover missing destination locations.

4. References

- Downs J, 2010, Time-geographic density estimation for moving point objects. *Lecture Notes in Computer Science* 6292:16-26.
- Goodchild M, 2010, Twenty Years of Progress: GIScience in 2010. *Journal of Spatial Information Science* (1):3-20.
- Haagerstrand T, 1970, What about people in Regional Science. *Papers of the Regional Science Association* 24:7-21.
- Hensher D, and A Reyes, 2000, Trip chaining as a barrier to the propensity to use public transport. *Transportation* 27 (4):341.
- Horner M, and M O'Kelly, 2007, Is Non-work Travel Excessive? *Journal of Transport Geography* 15 (6):411-416.
- Kuijpers B, H Miller, T Neutens, and W Othman, 2010, Anchor uncertainty and space-time prisms on road networks. *International Journal of Geographical Information Science* 24 (8):1223.
- Kwan M, 1998, Space-time and integral measures of individual accessibility: A comparative analysis using a point-based framework. *Geographical Analysis* 30 (3):191-216.
- Miller H, 1991, Modelling accessibility using space-time prism concepts within geographical information systems. *International Journal of Geographical Information Science* 5 (3):287-301.
- Neutens T, T Schwanen, and F Witlox, 2011, The Prism of Everyday Life: Towards a New Research Agenda for Time Geography. *Transport Reviews: A Transnational Transdisciplinary Journal* 31 (1):25 - 47.
- Neutens T, N Van de Weghe, F Witlox, and P De Maeyer, 2008, A three-dimensional network-based space-time prism. *Journal of Geographical Systems* 10 (1):89.
- O'Sullivan D, A Morrison, and J Shearer, 2000, Using desktop GIS for the investigation of accessibility by public transport: an isochrone approach. *International Journal of Geographical Information Science* 14 (1):85-104.
- Shaw S, and D Wang, 2000, Handling Disaggregate Spatiotemporal Travel Data in GIS. *GeoInformatica* 4 (2):161.
- Stopher P, C FitzGerald, and M Xu, 2007, Assessing the accuracy of the Sydney Household Travel Survey with GPS. *Transportation* 34 (6):723.
- Winter S, and Z Yin, 2010, The elements of probabilistic time geography. *Forthcoming in GeoInformatica*, DOI 10.1007/s10707-010-0108-1.
- Winter S, and Z Yin, 2010, Directed movements in probabilistic time geography. *International Journal of Geographical Information Science* 24 (9):1349 - 1365.
- Wolf J, R Guensler, and W Bachman, 2001, Elimination of the Travel Diary: Experiment to Derive Trip Purpose from Global Positioning System Travel Data. *Transportation Research Record* (1768):125-134.

SimTraj: An Approach to Similar Queries over Trajectories in Metric Spaces

Fábio Afonso¹, Fernanda Barbosa², Armando Rodrigues¹

¹CITI / Departamento de Informática
 Faculdade de Ciências e Tecnologia da UNL
 2829-516 Caparica, Portugal
 fabio.afonso@gmail.com
 arodrigues@di.fct.unl.pt

²Departamento de Informática
 Faculdade de Ciências e Tecnologia da UNL
 2829-516 Caparica, Portugal
 fb@di.fct.unl.pt

1. Introduction

With the rapid increase in the use of location-acquisition technologies (GPS, GSM networks, etc.), large amounts of spatio-temporal datasets will be accumulated. In different application domains, we need to represent moving entities, i.e. collect the successive location positions of a given entity, which form the trajectory for that entity. The key idea is that moving entities are described by a sequence of positions in a k-dimensional space. Each position in the sequence represents the entity's location at a given time. Thus, a trajectory for a moving entity in a k-dimensional space is viewed as a line in a k+1-dimensional space, where time is an additional dimension.

In many applications, there is a need to analyze the dynamics of moving objects in order to support spatiotemporal decisions. A significant type of query in these applications is the k-nearest neighbours (kNN), which finds the k trajectories more similar (closest) to a given trajectory. For example, in a football match, to identify a player with the most similar trajectory to Ronaldo's, in order to substitute a player, without changing team's strategy; or in a tourist guide application, to find the k most similar bus trajectories to a given touristic route; or to search for the k most similar hurricanes trajectories to Katrina's.

The choice of a distance function, which best represents the degree of similarity, for trajectories in a metric space, depends on the application domain in question. Some of the most used metric functions for moving objects are: the Euclidean Distance (ED), the Manhattan Distance (MD) and the Edit distance with Real Penalty (ERP) (Chen 2005).

In order to have efficient similar searching in metric spaces, several metric data structures have been proposed, which can be classified as cluster-based or pivot-based (Samet 2006; Chávez et al. 2001). Some of these metric data structures are: Recursive Lists of Clusters (RLC) (Mamede 2005; Sarmento 2010) and Metric-Tree (M-Tree) (GBDI 2009; Ciaccia et al. 1997). Both of these metric data structures are dynamic, implemented in secondary memory and seek to minimize the number of distance computations in a similarity search. The RLC is cluster-based, while the M-tree has features common to both, pivot-based and cluster-based.

The main goal of this research is to have a trajectory storage method, based on metric data structures, that speeds up the search by similarity. In the remainder of this paper we will, firstly, describe SimTraj, which is a trajectories storage method in metric spaces that provides efficient kNN searches. Then, we present the evaluation of the performance of SimTraj method in kNN searches. This evaluation involves the two metric data structures, RLC and the M-Tree, in two

metric spaces of hurricanes trajectories. The evaluation involves the two metric data structures, RLC and the M-Tree, in two metric spaces of hurricanes trajectories, using the distance functions, ERP and ED, as similarity functions.

2. SimTraj

SimTraj is a trajectories storage method in metric spaces, which has a distance-based indexing. This means that the trajectories are grouped into partitions, based on distances, between a set of selected trajectories and the remaining trajectories. At search time, the space partitions enable the discarding/retaining of some subsets, without additional calculations of distance, based on the metric properties of the similarity function. This method is a combination of two data structures, RLC (*distance-based partition*), which organizes the trajectories in clusters based on distance, and an in-memory structure (*frontline*), which stores the pointers to the clusters where the trajectories are stored, as showed in Figure 1.

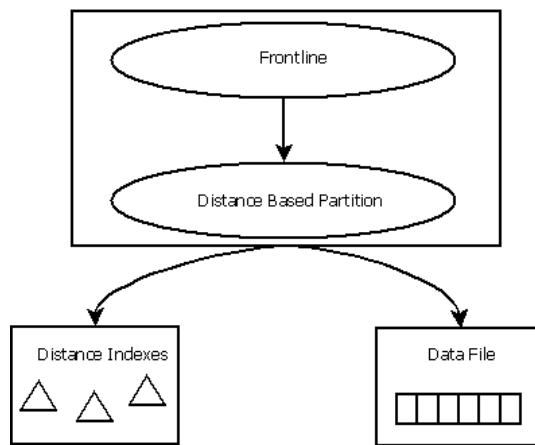


Figure 1. SimTraj Diagram

The *frontline* is used to provide efficient updates in the SimTraj. The *distance-based partition* is used to provide efficient similarity queries, and is implemented with recursive lists of clusters (RLC). A RLC cluster is a triple $\langle c, r, I \rangle$, where c is the centre, r the radius and I the interior of cluster. The interior is composed by elements whose distance to the centre is a value equal or less than the radius, and may be implemented as a list of clusters or as a leaf, depending on the number of elements and on the RLC capacity. Figure 2 shows how the trajectories are stored in the SimTraj method. In this figure, it is possible to see in (a) the partitions of trajectories in the space, and in (b) the organization of these trajectories in SimTraj, which has a RLC with capacity five and with variable radius at clusters.

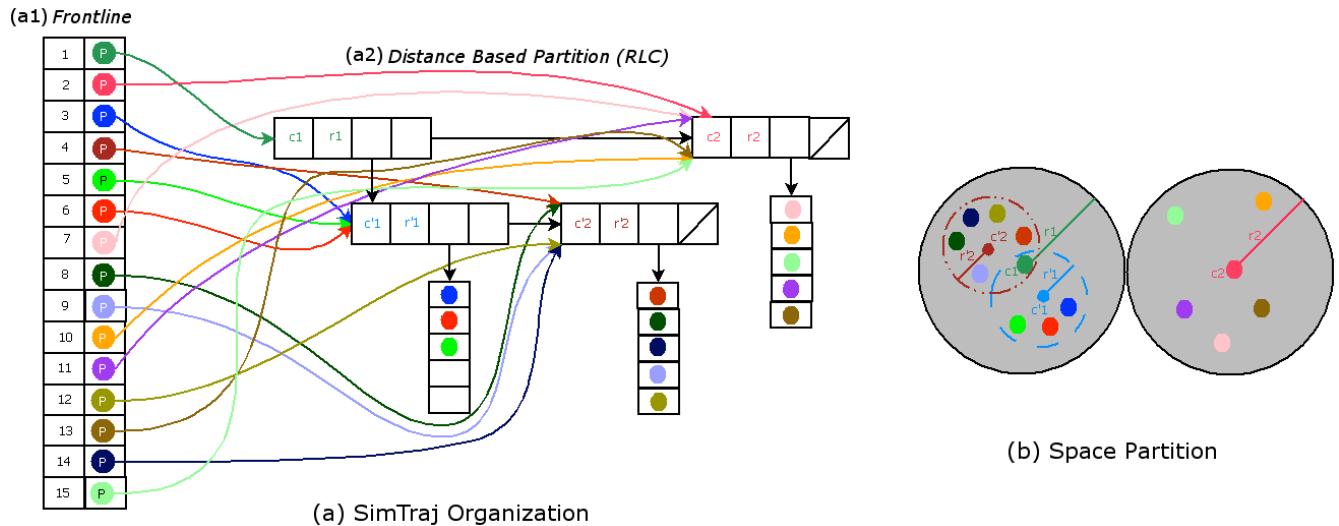


Figure 2. Organization of trajectories in SimTraj

In SimTraj, the insertion of a new trajectory consists in searching for a RLC cluster, in which the distance between the new trajectory and its center is less than or equal to its radius. If it is not found, a new cluster is created. Otherwise, one of two situations can happen: (1) the cluster is a leaf; (2) the cluster is a list of clusters. In (2) the process is applied recursively until it reaches (1). In (1), the first step is to verify if the leaf is not full (RLC capacity). If this is the case, the new trajectory is stored in the leaf. Otherwise, the leaf is transformed into a clusters list and the trajectory is allocated in one of the clusters. Finally, in both cases, the pair composed by the trajectory and the pointer to the RLC cluster is added to *frontline*.

The kNN query is based on a range query (RQ). To perform a kNN search, k elements are obtained from the RLC iterator and sorted downwardly by the distance to the query trajectory. Then a RQ is realized based on the query trajectory and on the largest distance found, which will be used as the query radius. The range query consists in iterating all RLC clusters, and, for each cluster, in finding the trajectories that lie at a distance from the query trajectory that is lower than or equal to a given value (query radius). Using the properties (triangle inequality and symmetry) of the metric function, many of these groupings of trajectories can be immediately discarded or retained, without additional calculations.

To remove a trajectory, *frontline* is consulted in order to obtain the pointer to the RLC cluster that contains the trajectory. When removing from the RLC, one of two situations can happen: (1) the trajectory is the center of a cluster; (2) the trajectory is stored in a leaf. In (1), the cluster is removed from the list, and all trajectories in this cluster will be inserted at the list of remaining clusters. In (2), the trajectory is removed from the leaf. In both cases, the pair associated to the removed trajectory is deleted from *frontline*.

In SimTraj, an update of a given trajectory (a new position in xy-plane at a given time) consists in removing the trajectory followed by the insertion of the updated trajectory.

It should be noted that insertions and deletions of trajectories that lead to a change in the organization of the remaining trajectories (e.g., the removal of a cluster and the reinsertion of its

elements, or the conversion of a leaf in a list of clusters), require an update in *frontline*, for each trajectory, which changed its location in RLC.

A more elaborated description of the RLC can be consulted at (Mamede 2005).

3. Experiments

In this section, we present an experimental evaluation realized in SimTraj, in order to assess the efficiency of k-NN searches (with $k=1$ and $k=5$). This evaluation involves two mechanisms in the *distance-based partition* of SimTraj. These techniques are the use of two metric data structures: RLC and M-tree.

The metric spaces, used in the evaluation, were defined on a hurricanes dataset, which contains all the Atlantic tropical hurricanes between 1851 and 2009 (Unisys 2010) (Figure 3 illustrates 2005).

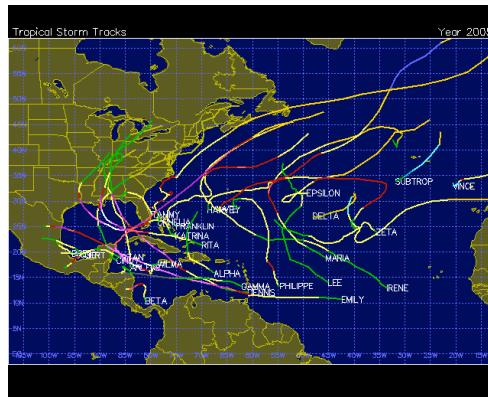


Figure 3. Atlantic tropical hurricanes (year 2005)

In our experiment, the measure of the similarity between two trajectories is based on two distance functions (ERP - Edit distance with Real Penalty, and ED - Euclidean distance). These functions give real values, which represent the degree of similarity between the trajectories. The smaller the function result is, the greater the similarity between the trajectories. ED is L_p -norm with $p = 2$, and ERP can be viewed as a variant of L_1 -norm, which can support local time shifting. To cope with local time shifting, ERP uses an idea from the string edit distance, which represents the number of insert, delete, or replace operations needed to change a string into another string. Note that, in the string edit distance, an added symbol is referred to as a gap element. ERP uses real penalty between two non-gap elements, but a constant value for computing the distance for gaps (origin in the x-y plane). These two functions are metric, as demonstrated in (Chen 2005).

Let $S = \langle(t_{1s}, x_{1s}, y_{1s}), \dots, (t_{ns}, x_{ns}, y_{ns})\rangle$ and $T = \langle(t_{1t}, x_{1t}, y_{1t}), \dots, (t_{mt}, x_{mt}, y_{mt})\rangle$ be two trajectories, let $p = (x_p, y_p)$ and $q = (x_q, y_q)$ points in x-y plane and let d be the distance between two points p and q in the x-y plane, denoted by $d(p, q)$.

The Euclidean distance between S and T , denoted by $ED(S, T)$, is defined in Equation 1. This function can only be applied to trajectories that have the same length ($n=m$). As the length of both trajectories has to be the same, the smaller trajectory needs to be extended. This extension is performed by inserting the start/end point of the small trajectory (point in x-y plane) at the start/end of the sequence, using the respective times of the trajectory with the largest length.

$$ED(S, T) = \sqrt{\sum_{1..n}((x_{it}-x_{is})^2 + (y_{it}-y_{is})^2)} \quad (1)$$

The Edit distance with Real Penalty between S and T, denoted by $\text{ERP}(S,T)$, is defined in Equation 2.

$$\text{ERP}(S,T) = \sum_{1..m} d((x_{it}, y_{it}), (0,0)), \text{ if } n = 0;$$

$$\text{ERP}(S,T) = \sum_{1..n} d((x_{is}, y_{is}), (0,0)), \text{ if } m = 0;$$

$$\begin{aligned} \text{ERP}(S,T) = \min \{ & \text{ERP}(<(t_{2s}, x_{2s}, y_{2s}), \dots, (t_{ns}, x_{ns}, y_{ns})>, <(t_{2t}, x_{2t}, y_{2t}), \dots, (t_{mt}, x_{mt}, y_{mt})>) \\ & + d((x_{1s}, y_{1s}), (x_{1t}, y_{1t})), \text{ERP}(<(t_{1s}, x_{1s}, y_{1s}), \dots, (t_{ns}, x_{ns}, y_{ns})>, <(t_{2t}, x_{2t}, y_{2t}), \dots, (t_{mt}, x_{mt}, y_{mt})>) \\ & + d((0,0), (x_{1t}, y_{1t})), \text{ERP}(<(t_{2s}, x_{2s}, y_{2s}), \dots, (t_{ns}, x_{ns}, y_{ns})>, <(t_{1t}, x_{1t}, y_{1t}), \dots, (t_{mt}, x_{mt}, y_{mt})>) \\ & + d((x_{1s}, y_{st}), (0,0)) \}, \text{ otherwise} \end{aligned} \quad (2)$$

RLC and M-Tree were parameterized in order to have the most efficient performance in the k-NN searches. While the values of RLC parameters come from experimental tests, the values of the M-Tree come from the results obtained in (Ciaccia et al. 1997).

In our experiment, we perform 37 (25% of the size database) kNN, for each k ($k=1$ and $k=5$). The trajectories chosen to perform the searches were chosen randomly from the dataset.

For each search, we calculated the number of disk accesses (SR), the execution time (ET) and the number of distance calculations performed (SD). Figures 4 and 5 show the average results obtained for each search in 1NN and 5NN, respectively.

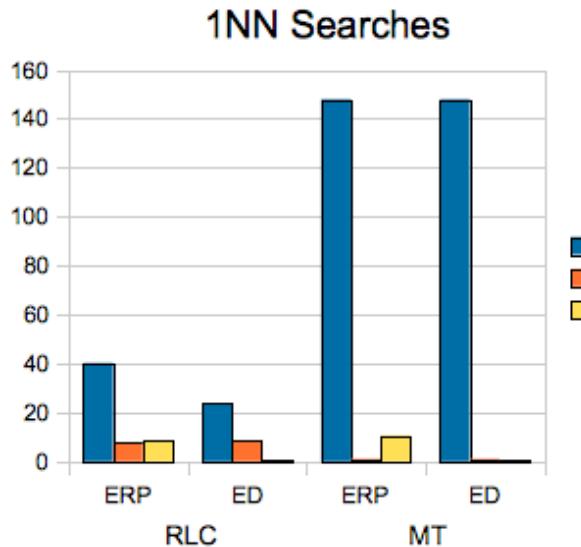


Figure 4. 1NN Searches using ERP and ED at both Data Structures

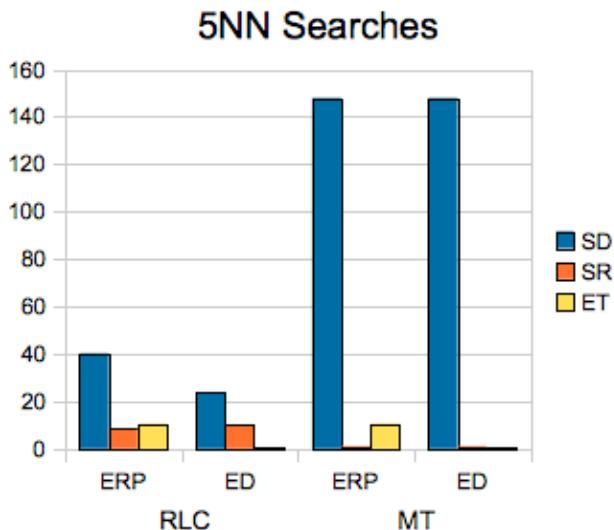


Figure 5. 5NN Searches using ERP and ED at both Data Structures

We can conclude that the RLC and the M-Tree are very competitive. However, the RLC is better at the number of distance calculations and at the execution time. The M-Tree has better values in disk accesses.

Based on disk accesses, one could imagine that the M-Tree would be unbeatable. However such was not true. This happens due the fact that the M-Tree only accesses the disk once, but

performs a higher number of distances calculations than the RLC. Similar results with a different trajectories dataset were obtained in (Afonso et al. 2011).

5. Conclusions and Future Work

In this work we present a trajectories storage method (SimTraj) for efficient similar search in metric spaces. The choice of the RLC data structure to distance-based partition from SimTraj was based on experimental tests performed on two data sets, hurricanes and buses.

An ongoing work is the evaluation the dynamism of the SimTraj. As future work, we aim to explore some interesting scenarios related to the comparison of metric data structures with non-metric ones, as well as the evaluation of SimTraj in a concrete application.

6. References

- Afonso, F., Barbosa, F. & Rodrigues, A., 2011. Trajectory Data Similarity with Metric Data Structures. In *Proceedings of GISRUK 2011*. Portsmouth, United Kingdom.
- Chávez, E. et al., 2001. Searching in Metric Spaces. *ACM Computing Surveys (CSUR)*, 33(3).
- Chen, L., 2005. *Similarity-based Search Over Time Series and Trajectory Data*. Ph.D. Thesis. Waterloo, Ontario, Canada: University of Waterloo.
- Ciaccia, P., Patella, M. & Zezula, P., 1997. M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. In *Proceedings of the 23rd International Conference on Very Large Data Bases (VLDB 1997)*. Athens, Greece: Morgan Kaufmann Publishers, pp. 426-435.
- GBDI, 2009. M-Tree. *Databases and Images Group*. Available at: <http://www.gbdi.icmc.usp.br/en/home>.
- Mamede, M., 2005. Recursive Lists of Clusters: a Dynamic Data Structure for Range Queries in Metric Spaces. In *Proceedings of the 22th International Symposium on Computer and Information Sciences*. Istambul, Turkey: Springer-Verlag, pp. 843-853.
- Samet, H., 2006. Foundations of Multidimensional and Metric Data Structures. In *Foundations of Multidimensional and Metric Data Structures*. USA: Morgan Kaufmann Publishers Inc., pp. 50-89; 270-311; 356-357; 566; 608.
- Sarmento, Â.M.L., 2010. *Estruturas de Dados Métricas Genéricas Memória Secundária*. Master's Thesis in Computer Engineering. Caparica, Portugal: UNL/FCT (in Portuguese).
- Unisys, 2010. Hurricanes Dataset. *Atlantic Tropical Storm Tracking by Year*. Available at: <http://weather.unisys.com/hurricane/atlantic/>.

Measuring Population Shift Bias in Tests of Space-Time Interaction

Nicholas Malizia, Elizabeth A. Mack, and Sergio J. Rey

GeoDa Center for Geospatial Analysis and Computation
School of Geographical Sciences and Urban Planning
Arizona State University
975 S. Myrtle Ave. Tempe, AZ 85287-5302, USA
Email: nmalizia@asu.edu

1 Introduction

Tests of space-time interaction detect clustering of events in space and time in excess of “any purely spatial or purely temporal clustering” (Kulldorff, 1998, pg. 58). These tests are widely employed in studies of crime (e.g. Knox, 2002; Grubesic and Mack, 2008) and disease (e.g. Petridou et al., 1996; Rogerson, 2001). By simultaneously considering both the spatial and temporal dimensions of the event patterns, these methods are capable of identifying certain data generating processes and, as a result, are often used to inform etiological work (Ward and Carpenter, 2000). Most of these tests, however, dubiously assume the underlying susceptible population within a study area to be invariant through time and across space. In settings where this assumption does not hold, these tests will detect space-time interaction due to population changes in addition to interaction resulting from the data generating process of interest. The excess interaction observed due to violating this assumption and by failing to account for the changes in the underlying population is referred to as population shift bias (Kulldorff and Hjalmars, 1999). Although recognized, this bias is often not accounted for in practice and its potential impact on results is not fully explored. This paper carries out a simulation to develop a detailed understanding of the impact of population shift bias on three of the most common tests of space-time interaction: the Knox (1964), Mantel (1967), and Jacquez (1996) tests. Additionally, the simulation demonstrates that contrary to prior claims (i.e. Kulldorff and Hjalmars, 1999; Aldstadt, 2007), population shift bias is problematic even in studies with a short temporal extent. To these ends, we simulate events within the dynamic population of a hypothetical metropolitan landscape over the course of one day. We then quantify the amount of population shift bias affecting each of the space-time interaction tests for a number of different population movement scenarios.

2 Interaction Tests

The space-time interaction tests considered in this study are described in further detail below. The methods have been implemented by the authors in Python and are available in the open-source space-time analysis software, PySAL (Rey and Anselin,

2010). Note that in all cases we employ Euclidean distance metrics. Also, events are never considered adjacent to or neighbours of themselves.

To calculate the Knox (1964) test for space-time interaction, critical space and time distance thresholds (δ and τ , respectively) defining adjacency between events are specified by the user. The test statistic is then calculated as the count of event pairs that are adjacent in both time and space. Formally, the test statistic is specified in Equation 1, where n = number of events, a^s = adjacency in space, a^t = adjacency in time, d^s = distance in space, and d^t = distance in time.

$$X = \sum_i^n \sum_j^n a_{ij}^s a_{ij}^t \quad (1)$$

$$a_{ij}^s = \begin{cases} 1, & \text{if } d_{ij}^s < \delta \\ 0, & \text{otherwise} \end{cases}$$

$$a_{ij}^t = \begin{cases} 1, & \text{if } d_{ij}^t < \tau \\ 0, & \text{otherwise} \end{cases}$$

The Mantel test is a modification of the Knox test that considers the space and time distances between all pairs of events, and not just those within critical thresholds (Mantel, 1967). The test statistic is the sum of the products of the spatial and temporal distances between all event pairs in the dataset. The statistic is specified in Equation 2, where, again, d^s and d^t denote distance in space and time, respectively.

$$M = \sum_i^n \sum_j^n d_{ij}^s d_{ij}^t \quad (2)$$

In an effort to address shortcomings of the previous two methods, Jacquez (1996) developed a test using a similar form, based on nearest neighbour distances. The test locates the k nearest neighbours in both space and time for all events and then counts those common to both dimensions for individual events. Formally, the statistic, J_k is defined in Equation 3, where n = number of cases; a^s = adjacency in space; a^t = adjacency in time.

$$J_k = \sum_i^n \sum_j^n a_{ijk}^s a_{ijk}^t \quad (3)$$

$$a_{ijk}^s = \begin{cases} 1, & \text{if event } j \text{ is a } k \text{ nearest neighbour of event } i \text{ in space} \\ 0, & \text{otherwise} \end{cases}$$

$$a_{ijk}^t = \begin{cases} 1, & \text{if event } j \text{ is a } k \text{ nearest neighbour of event } i \text{ in time} \\ 0, & \text{otherwise} \end{cases}$$

To assess the significance of the results for each of these tests, a Monte Carlo approach is traditionally used where in each permutation the temporal coordinates are shuffled and the statistic is recalculated. This generates a distribution of potential

values for the statistic (specific to the observed event pattern), which is then used to assess the pseudo-significance of the observed test statistic value. While this approach is appropriate in situations where the susceptible population is static across time, it is inappropriate when the distribution changes heterogeneously through time and space. Using this method in such a context introduces the population shift bias mentioned above.

3 Methods

This study measures the bias introduced by failing to account for shifts in the susceptible population for a hypothetical metropolitan area over the course of one day. To measure the bias, events (i.e. crimes, illnesses) are randomly generated within the population in each of four daily movement scenarios: high movement (where 98% of the individuals change spatial unit for some period during the day); moderate movement (59% change unit); low movement (35% change unit); no movement (all individuals remain within unit). The metropolitan area has a population of 640,000 divided equally among its 40 spatial units (see Figure 1). In each of the dynamic scenarios, the population in the spatial units varies heterogeneously over the course of the day. Some spatial units gain population (employment or shopping locations) at certain points of the day while others lose population (bedroom communities). Additionally, we consider the same scenarios with an additional influx of 400,000 individuals to the metro (visitors or commuters) from the periphery during the day.



Figure 1: Simulation study area.

To estimate the population shift bias, we follow the methodology of Kulldorff and Hjalmars (1999). For our experiment, this means events are randomly assigned to spatial units in the metro at different hours in the day based on a probability proportional to the population of the spatial unit at each hour of the day. In this example, all individuals were assumed to be susceptible to the events. For each movement scenario, 1000 replications are run where 100 events are randomly simulated. The significance of the test statistics in each replication is assessed using the Monte Carlo approach described above. For each scenario and test combination, the proportion of significant replications (where $\alpha = 0.05$ and 0.01) is recorded. Because there is no population movement in the static scenario, there is no population shift bias; as a result, the proportion of significant replications for this scenario serves as our baseline. The difference between the proportion of significant replications observed for the dynamic population scenarios and that observed for the static population scenario measures the amount of population shift bias present in each of the tests, for each scenario. The parameters used in this study for the Knox and Jacquez tests are outlined in Table 1, no additional parameters were specified for the Mantel test.

4 Results

The results, shown in Table 1, illustrate the sizable impact population shift bias may have on these tests of space-time interaction, even for the short temporal extent considered. Generally speaking, the Knox test was most affected by the population shifts. As the critical distances used by the test increased, observed bias increased as well, in one case up to 95 times the α value. Although this extreme example is partly an artifact of our experimental design, which intended to promote any potential bias by concentrating mobile individuals in the gaining spatial units, the scenarios designed are not implausible and neither, therefore, are the estimates of the bias. Researchers employing this test, especially in an urban context, need to be aware of this susceptibility. Although still affected, the results for the Jacquez test displayed the least amount of bias, likely due to the relative nature of the nearest neighbour distance metric employed by the test. For all tests, any bias observed was increased by the addition of the influx population to the metro area.

The take-home message from this work is that population shift bias must be accounted for when employing tests of space-time interaction regardless of the test employed or the duration of the study. This can be accomplished by using an unbiased form of the test which takes population shift into account. A general template for such unbiased tests is described in Kulldorff and Hjalmars (1999). Future research should concentrate on specific implementations of this form.

5 Acknowledgements

This work was supported by the U.S. National Science Foundation through a Graduate Research Fellowship to Nicholas Malizia.

Test	Parameters	Without Influx Population						With Influx Population					
		Low Movement		Moderate Movement		High Movement		Low Movement		Moderate Movement		High Movement	
		$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
Jacquez	$k = 1$	0.009	0.006	0.032	0.016	0.108	0.058	0.030	0.005	0.033	0.020	0.103	0.053
	$k = 2$	0.008	-0.004	0.035	0.011	0.139	0.073	0.020	0.009	0.044	0.028	0.167	0.116
	$k = 3$	0.007	0.001	0.050	0.022	0.257	0.140	0.045	0.017	0.074	0.037	0.241	0.149
	$k = 4$	0.003	-0.005	0.039	0.022	0.302	0.176	0.038	0.012	0.096	0.040	0.299	0.188
	$k = 5$	0.007	-0.003	0.049	0.019	0.390	0.228	0.065	0.019	0.121	0.047	0.376	0.244
Knox	$\delta = 0.5, \tau = 0.25$	0.030	0.021	0.024	0.002	0.062	0.045	0.022	0.017	0.028	0.017	0.082	0.100
	$\delta = 1.0, \tau = 0.25$	0.016	0.009	0.038	0.016	0.125	0.062	0.054	0.030	0.068	0.031	0.158	0.118
	$\delta = 2.0, \tau = 0.25$	0.022	0.023	0.035	0.033	0.170	0.084	0.062	0.031	0.092	0.049	0.216	0.153
	$\delta = 5.0, \tau = 0.25$	0.024	0.009	0.071	0.028	0.327	0.171	0.094	0.031	0.127	0.054	0.383	0.223
	$\delta = 0.5, \tau = 0.50$	0.011	0.012	0.042	0.023	0.160	0.079	0.045	0.032	0.068	0.042	0.200	0.147
	$\delta = 1.0, \tau = 0.50$	0.017	0.011	0.056	0.026	0.266	0.146	0.064	0.034	0.141	0.086	0.319	0.220
	$\delta = 2.0, \tau = 0.50$	0.031	0.016	0.088	0.048	0.411	0.261	0.107	0.052	0.237	0.116	0.473	0.348
	$\delta = 5.0, \tau = 0.50$	0.038	0.016	0.154	0.067	0.661	0.499	0.181	0.075	0.349	0.179	0.766	0.599
	$\delta = 0.5, \tau = 1.00$	0.025	0.004	0.079	0.040	0.332	0.207	0.097	0.043	0.180	0.086	0.416	0.303
	$\delta = 1.0, \tau = 1.00$	0.043	0.016	0.116	0.050	0.528	0.376	0.144	0.067	0.276	0.149	0.587	0.461
Mantel	$\delta = 2.0, \tau = 1.00$	0.052	0.019	0.134	0.076	0.709	0.575	0.224	0.104	0.410	0.230	0.777	0.686
	$\delta = 5.0, \tau = 1.00$	0.079	0.039	0.249	0.127	0.896	0.847	0.337	0.176	0.603	0.421	0.926	0.905
	$\delta = 0.5, \tau = 2.00$	0.017	0.009	0.036	0.022	0.433	0.250	0.115	0.050	0.185	0.089	0.501	0.364
	$\delta = 1.0, \tau = 2.00$	0.028	0.011	0.094	0.042	0.649	0.502	0.173	0.072	0.309	0.170	0.737	0.604
	$\delta = 2.0, \tau = 2.00$	0.044	0.016	0.150	0.081	0.790	0.704	0.261	0.123	0.451	0.277	0.852	0.821
Mantel	$\delta = 5.0, \tau = 2.00$	0.071	0.042	0.249	0.150	0.908	0.898	0.362	0.203	0.638	0.487	0.929	0.950
		0.075	0.026	0.139	0.056	0.592	0.407	0.181	0.086	0.216	0.093	0.569	0.384

Table 1: Population shift bias for all combinations of tests and population movement scenarios.

6 References

- Aldstadt, J. (2007). An incremental Knox test for the determination of the serial interval between successive cases of an infectious disease. *Stochastic Environmental Research and Risk Assessment*, 21(5):487–500.
- Grubesic, T. and Mack, E. (2008). Spatio-temporal interaction of urban crime. *Journal of Quantitative Criminology*, 24(3):285–306.
- Jacquez, G. (1996). A k nearest neighbour test for space-time interaction. *Statistics in Medicine*, 15(18):1935–1949.
- Knox, E. (1964). The detection of space-time interactions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 13(1):25–30.
- Knox, E. (2002). An epidemic pattern of murder. *Journal of Public Health*, 24(1):34–37.
- Kulldorff, M. (1998). Statistical methods for spatial epidemiology: tests for randomness. In Gatrell, A. and Löytönen, M., editors, *GIS and Health*, pages 49–62. Taylor & Francis, London.
- Kulldorff, M. and Hjalmars, U. (1999). The Knox method and other tests for space-time interaction. *Biometrics*, 55(2):544–552.
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2):209–220.
- Petridou, E., Revithi, K., Alexander, F., Haidas, S., Koliouskas, D., Kosmidis, H., Piperopoulou, F., Tzortzatou, F., and Trichopoulos, D. (1996). Space-time clustering of childhood leukaemia in Greece: evidence supporting a viral aetiology. *British Journal of Cancer*, 73(10):1278–1283.
- Rey, S. and Anselin, L. (2010). PySAL: A Python library of spatial analytical methods. In Fischer, M. M. and Getis, A., editors, *Handbook of Applied Spatial Analysis*, pages 175–193. Springer.
- Rogerson, P. (2001). Monitoring point patterns for the development of space-time clusters. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1):87–96.
- Ward, M. and Carpenter, T. (2000). Analysis of time-space clustering in veterinary epidemiology. *Preventive Veterinary Medicine*, 43(4):225–237.

Spatio-Temporal Analysis of Air Pollution Data in Malta

Ms Luana Chetcuti Zammit¹, Dr Kenneth Scerri¹, Dr Maria Attard², Ms Thérèse Bajada³, Mr Mark Scerri⁴

¹Systems and Control Engineering Department, Room 419, Engineering Building, University of Malta, MSD2080, Malta.
Tel: +356 2340 2080
E-mail: lche0003@um.edu.mt; kenneth.scerri@um.edu.mt

²Geography Department, University of Malta, OH132, University of Malta, Msida MSD2080, Malta.
Tel: +356 2340 2147
E-mail: maria.attard@um.edu.mt

³Institute for Sustainable Development, First Floor, Regional Building, Triq l-Imhallef Paolo Debono, University of Malta, MSIDA MSD 2033, Malta.
Tel: +356 2340 3404
E-mail: therese.bajada@um.edu.mt

⁴Malta Environment and Planning Authority, Unit D2, Environment Protection Directorate, PO Box 200, Marsa MRS1000, Malta.
Tel: +356 2290 7203
E-mail mark.scerri@mepa.org.mt

1. Introduction

Air pollution measurements display patterns over space and time allowing for spatio-temporal modelling, through which pollution concentrations and trends can be analysed. In Malta, the MEPA (Malta Environment and Planning Authority) collects monthly averaged data for various pollutants from a network of 123 diffusion tubes located around the Islands (Figure 1). This preliminary study uses data associated with traffic, that is nitrogen dioxide (NO_2) and benzene, collected monthly between the period 2004 and 2010 with the objectives to i) develop a computationally efficient method that best describes the data; ii) determine the level of dependency of each site on neighbouring ones and iii) identify any factors that affect the behaviour and patterns of pollution. Results will show that generally there is a low spatial dependency between close sites, thus implying that local sources, rather than diffusion, have a predominant effect on the measurements. This analysis will prove valuable in MEPA's redistribution exercise of the diffusion tube network to determine which sites are necessary to retain and which sites can be removed without significantly affecting the information gathered.

2. The case study - Malta

Malta is located in the centre of the Mediterranean Sea at approximately 100 kilometres south of Sicily. The archipelago consists of three main islands with a population of over 400,000, with the major environmental concern being the air quality (Government of Malta 2002, Office of the Prime Minister 2010). The main contributors to air pollution are the high demands for energy generation and the growth in private car use. Electricity is generated from the combustion of fuel oil at two power stations. In addition, the islands were home to 229,016 private vehicles in 2009 (NSO 2009), one of the highest car ownership rates in the world. These have increased the risks of atmospheric pollution particularly by carbon monoxide, oxides of nitrogen, volatile organic compounds and particulate matter (MEPA 2010).

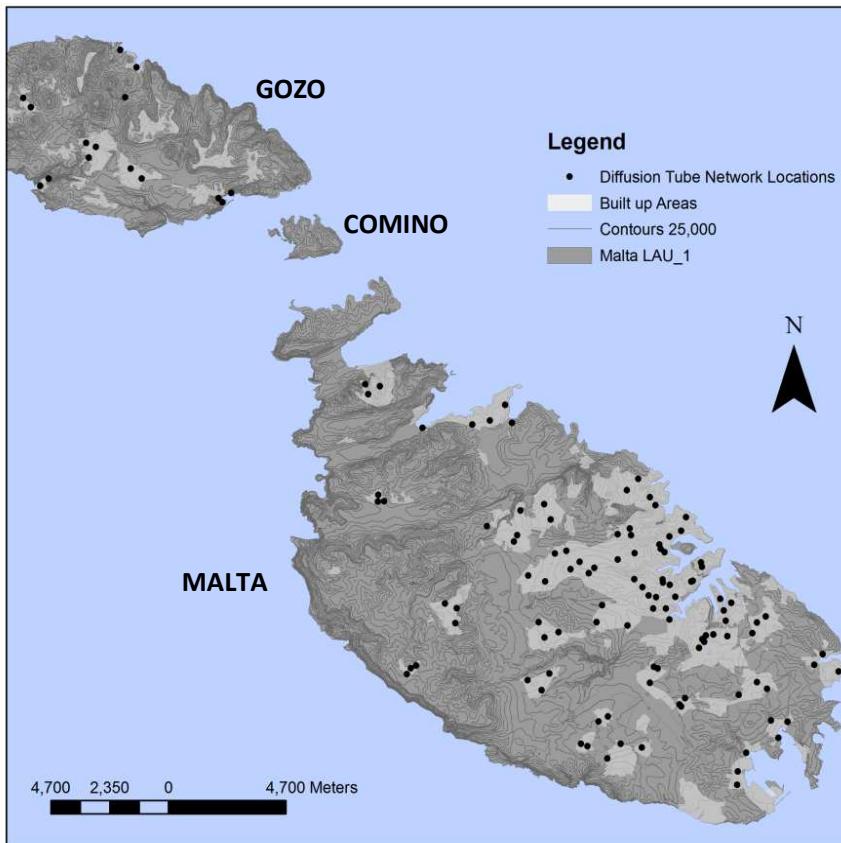


Figure 1. Location of Passive Diffusion Tubes (drawn by author)

3. Data Analysis and Validation

Data-driven modelling strategies (Ljung 1999) are applied in this research to study the dependencies between readings taken at different sites. In particular, statistical multivariate or Vector AutoRegressive (VAR) models are used to represent the spatio-temporal relationships in the data. These methods were first proposed for the study of such phenomena in Pfeifer and Deutsch (1980a, b).

Seasonal and temporal trends in the data were eliminated so as to identify dependencies among sites rather than correlations due to these trends. Detrending was performed by differencing (Chatfield 2004) and the stationary time series obtained were modelled by the $\text{VAR}(p,q)$ models, given by equation 1:

$$\mathbf{z}_t = \mathbf{A}_1 \mathbf{z}_{t-1} + \mathbf{A}_2 \mathbf{z}_{t-2} + \mathbf{A}_3 \mathbf{z}_{t-3} + \mathbf{A}_4 \mathbf{z}_{t-4} + \dots + \mathbf{A}_p \mathbf{z}_{t-p} \quad (1)$$

where p denotes the temporal order, q denotes the spatial order of the system, $\mathbf{A}_i \in \mathbb{R}^{123 \times 123}$ are the

autoregressive terms and $\mathbf{z}_t \in \mathbb{R}^{123 \times 1}$ denotes the air pollution observations of all 123 sites at time t .

Due to the large number of parameters to estimate from the limited data, the computationally advantageous method in de Luna and Genton (2005) was adopted, summarized in Algorithm 1. Some assumptions have been made, namely that the climate remains homogenous throughout all sites under study, that the spatially closer sites have a larger probability of being correlated thus providing a natural ordering for the sites and that only temporal dependences over a monthly period can be captured due to the data's temporal resolution and thus shorter term dependences cannot be ruled out.

```

Iterate for site,  $s_i=1,2,\dots,123$ 
  Order all sites in ascending order of distance relative to  $s_i$ 
  For  $p = 1,2,\dots,n$ 
    For  $q=1,2,\dots,k$  (where  $n$  and  $k$  represent the maximum temporal and spatial order respectively)
    Estimate  $A_1, A_2, \dots, A_p$ 
  Identify the best model orders for  $s_i$  based on some comparative measures

```

Algorithm 1: Iterative model building strategy (de Luna and Genton 2005)

The applied comparative measures are based on model selection criteria, which include the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC) and the Mean Squared Error (MSE). The principle of parsimony (Chatfield 2004) is applied thus aiming to identifying the simplest model that produces good prediction results. Figures 2 and 3 show MSE, AIC and BIC values for different temporal model orders for benzene and NO₂ respectively. Since AIC and BIC values tend to penalize models with larger spatial and temporal orders, these results are weighted more heavily when deciding on the model order.

Note that in Algorithm 1, the computational demand is significantly reduced when compared to estimating full VAR models since only the statistically significant coefficients are estimated. Note also that, to the authors' knowledge these methods have only been applied to datasets with a small number of observation sites (usually less than 10), while in this work the flexibility of this method to solve higher dimension problems (123 sites) has been tested. The predictive accuracy of the models obtained has been tested by a validation data set not used in the estimation procedure. Using this dataset, the one step ahead residues after modelling were found to be temporally white (thus void of any further linear temporal relationship) up to a confidence interval of 88% for benzene and 91% for NO₂. Based on these models the spatial order for both benzene and NO₂ was mostly limited to the three closest neighbours.

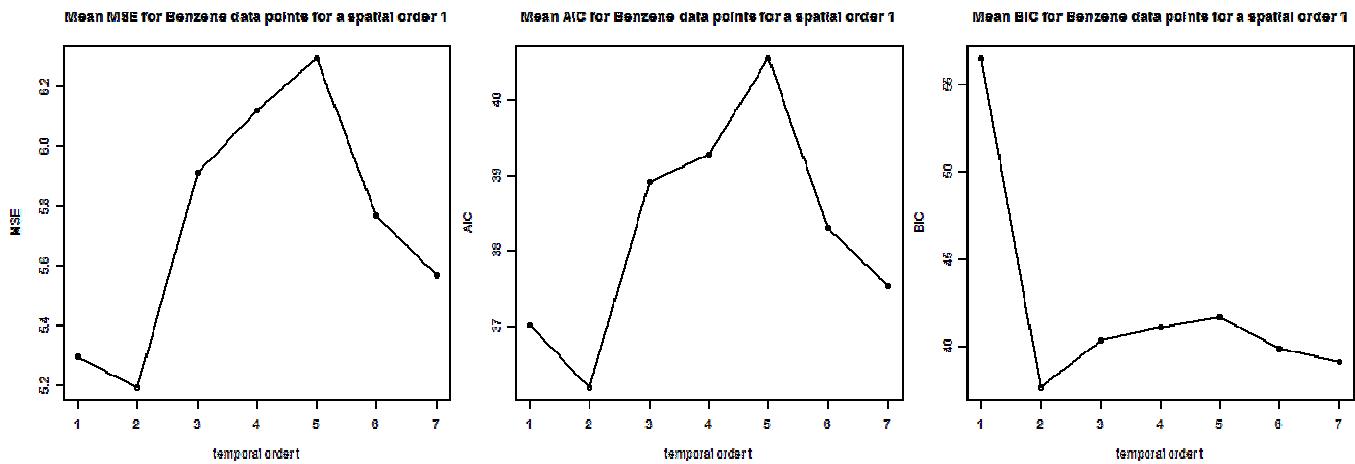
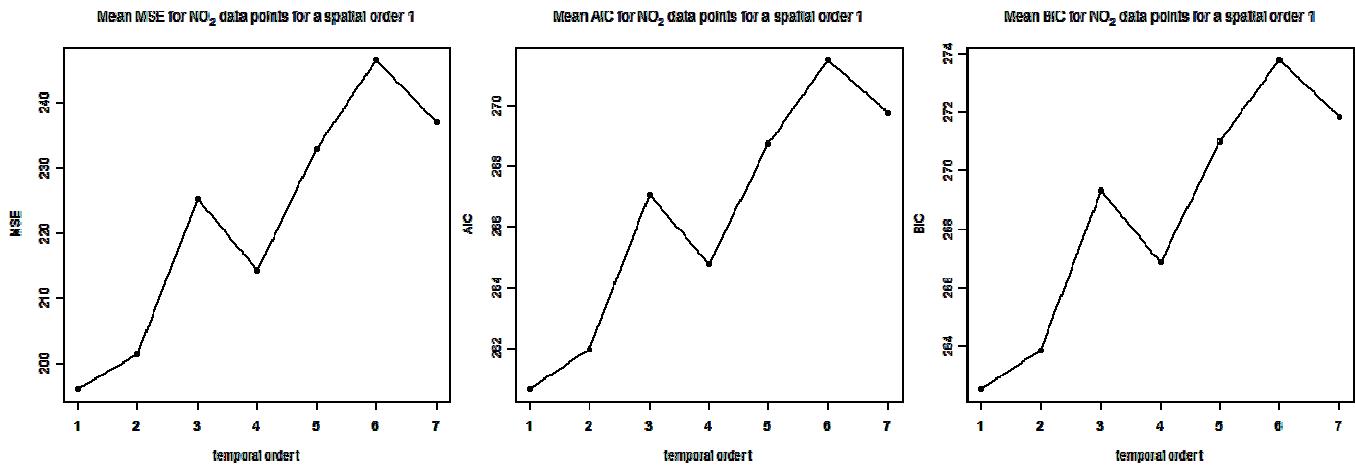


Figure 2. The mean MSE, AIC and BIC for benzene over all sites.

Figure 3. The mean MSE, AIC and BIC for NO₂ over all sites.

4. Conclusions

Figures 4a and 4b give histograms for the number of dependent sites while Figures 5a and 5b show the number of dependent sites for each location for benzene and NO₂ respectively.

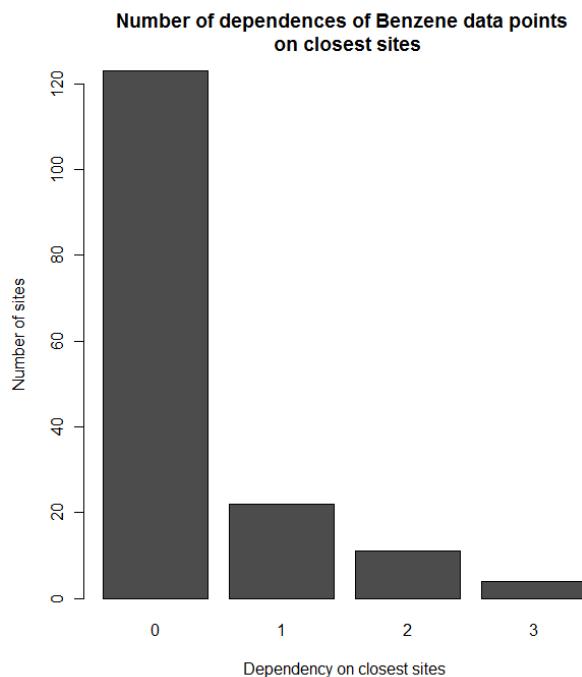


Figure 4a. Histogram showing dependency of Benzene data points on closest sites

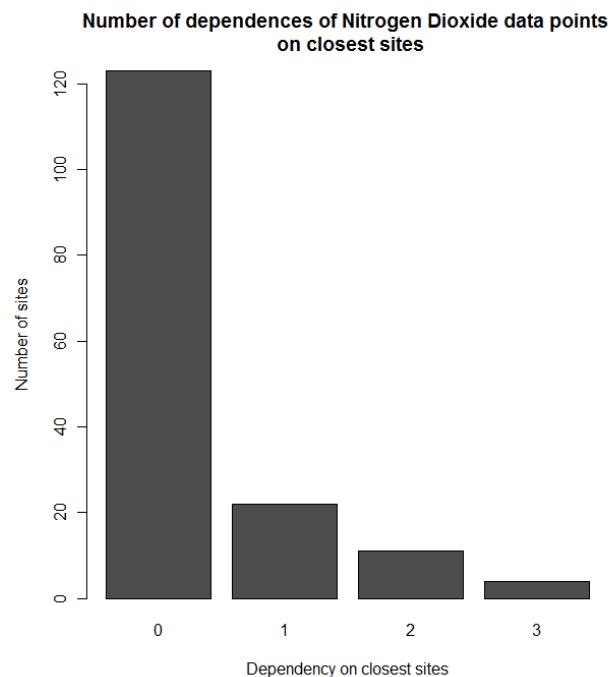


Figure 4b. Histogram showing dependency of NO₂ data points on closest sites

In Figures 4 and 5, a value of 0 indicates that the reading at that particular site is only dependent on previous readings at the same site, while a value of 1 indicates that the measurements are dependent on the site itself and its first closest neighbour, and so on for the other values.

The assumption that dispersal of pollutants is equidistant and therefore one source of pollution in one area has an effect on the neighbouring areas is not supported here. This is further demonstrated by the overlaying of potential sources of pollution in the main island such as traffic density, industrial estates, power stations and the airport. Notwithstanding that most of the points are located relatively close to each other and to these sources, most readings seem to be independent.

The overall spatially independent behaviour of these pollutants would suggest that there are other, more local factors that are affecting air pollution. Some possible interpretations follow.

- Since there is input from a stable source (e.g. traffic) similar temporal patterns can be observed. However, at another location, the source input levels may change (for example, less traffic) and therefore the behaviour of that point, even though it is relatively close, is independent. This is most evident in the area northwest of the Grand Harbour (marked A in Figures 5a and 5b). This is reasonable since in the Maltese urban environment the urban density, urban fabric and traffic change considerably over a relatively short distance.
- A few points experience higher spatial dependencies. These are marked with the letters B and C in Figure 5a. In these cases we note that (i) the pollution values at some of these locations are relatively low, thus affecting the accuracy of the modelling procedure (area marked B) (ii) there are very similar environmental conditions (traffic and urban density) affecting the sites (area marked C).
- These results are validated by MEPA's approach adopted for the location of diffusion tubes. MEPA selects two to three sites per locality, one of which is a traffic site and the other/others are background sites (without traffic).

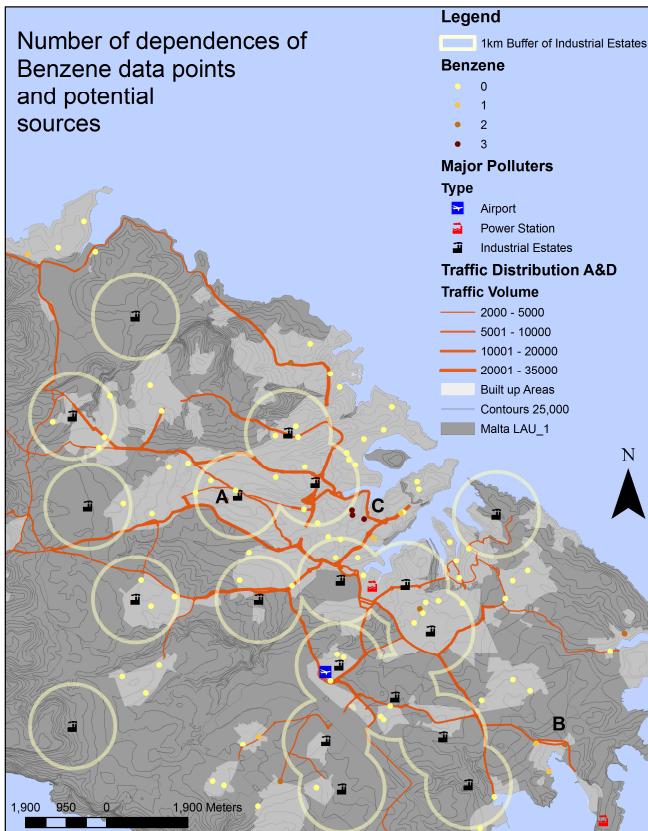


Figure 5a. The number of dependent sites for each location for Benzene.

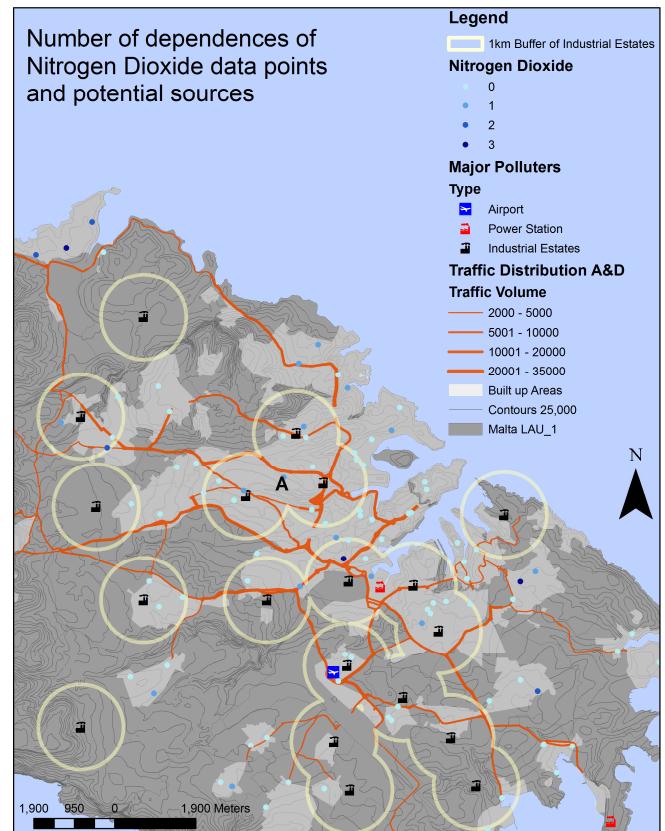


Figure 5b. The number of dependent sites for each location for NO₂

Future work will focus on introducing measured pollution sources to the mathematical model to verify the dependency of the pollution readings on these sources.

5. Acknowledgements

The research work disclosed in this abstract is funded by the Malta Government Scholarship Scheme (MGSS).

6. References

- Chatfield C, 2004, *The Analysis of Time Series*. Chapman & Hall / CRC, USA.
- De Luna X and Genton MG, 2005, Predictive Spatio-Temporal Models for spatially sparse environmental data. *Statistica Sinica*, 15: 547-568. Available online at <http://www3.umu.se/stat/personal/xavier.deluna/papers/stpredict.pdf>. Last accessed 16 March 2011.
- Government of Malta, 2002, Johannesburg Summit 2002, Malta Country Profile.
- Ljung L, 1999, *System Identification – Theory For the User*. Prentice Hall, USA.
- MEPA, 2010, *Air: Sources and Effects*. Malta Environment and Planning Authority, Malta. Available online at <http://www.mepa.org.mt/air-sources>. Last accessed 15 October 2010.
- NSO, 2009, *Motor Vehicles: Q4/2009* News Release. 20 January 2010. 008/2010. Available online at <http://www.nso.gov.mt/statdoc/document file.aspx?id=2669>. Last accessed 15 January 2010.

Office of the Prime Minister, 2010, *Air Quality Plan for the Maltese Islands*. Prepared by the Malta Environment and Planning Authority, Floriana, Malta. Available online at <http://www.mepa.org.mt/air-publications>. Last accessed 5 January 2010.

Pfeifer PE and Deutsch SJ, 1980a, A three-stage iterative procedure for space-time modeling. *Technometrics*, 22(1): 35-47.

Pfeifer PE and Deutsch SJ, 1980b, Identification and Interpretation of First-Order Space-Time ARMA Models. *Technometrics*, 22(3): 397-403.

Toponym disambiguation of landscape features using geomorphometric characteristics

C. Derungs¹, R. S. Purves¹, B. Waldvogel²

¹University of Zürich - Irchel, Winterthurerstr. 190, 8057 Zürich, Switzerland
Email: curdin.derungs, ross.purves@geo.uzh.ch

²Swiss Federal Research Institute WSL, Zürcherstr. 111, 8903 Birmensdorf, Switzerland
Email: bettina.waldvogel@wsl.ch

1. Introduction

Landscape descriptions in natural language are a primary source of what Egenhofer and Mark (1995) call naïve geographical knowledge. Naïve geographical knowledge, however, differs for different people from different cultures and backgrounds (Mark and Turk 2003). For example a description of Uluru in Australia might be very different if given by Dutch tourist in comparison to one given an indigenous inhabitant.

Geoparsing, in particular considering toponym ambiguity is a key task in linking language to space through the assignment of geographic scopes to documents (Clough 2005). Leidner (2007) states that almost all research in geoparsing has focused on *populated places*. ‘Population’ furnishes toponyms with a priori knowledge that is used by state of the art disambiguation approaches (e.g. Purves et al. 2007) using the most populated place as the default toponym in disambiguation.

Landscape descriptions, however, typically contain references to unpopulated places, implying other approaches must be adopted to disambiguate.

Here we generate missing knowledge about toponyms using geomorphometric characteristics, in our case for a landscape feature known as a Hochmoor¹. The toponym knowledge thus created is used for referent disambiguation (i.e. is London, England or London, Ontario relevant) - to our knowledge the first example of *geomorphometric disambiguation*. Our method shows considerable improvement in performance over a baseline disambiguation method. Disambiguation is the first important step towards opening up extensive sources of naïve geographical knowledge in the form of landscape descriptions in natural language which are likely to contain many ambiguous toponyms, which in turn will make such documents more accessible for a wide range of geographically rooted research.

2. Data Center Nature and Landscape

In our investigation we use documents describing Hochmoor in natural language. The documents are part of the Data Center Nature and Landscape (DNL). The DNL was established according to the specifications of the Swiss Nature and Cultural Heritage Protection to manage all Swiss data regarding protected areas of national importance. Information on the condition, composition and location of more than 500 *Hochmoor* in

¹ We use the German term *Hochmoor* which is a geographic object closely related to a high moor or a bog, to avoid semantic confusion through translation.

Switzerland has been collected in a corpus and recorded in separate datasheets (Bauer-Messmer et al. 2009). The datasheets are written in three national languages, French, Italian or German and we investigate German datasheets here ($n=370$). A simple gazetteer lookup performed on the documents using SwissNames² recognizes 600, mostly ambiguous, toponyms that can be referenced to more than 2500 locations in Switzerland.

3. Geomorphometric knowledge for toponym disambiguation

We assume that locations of toponyms used to georeference Hochmoor have a Hochmoor-like topography. Therefore a geomorphometric measurement for Hochmoor is deduced from topography. This measurement is further used as the missing knowledge in disambiguation (Figure 1).

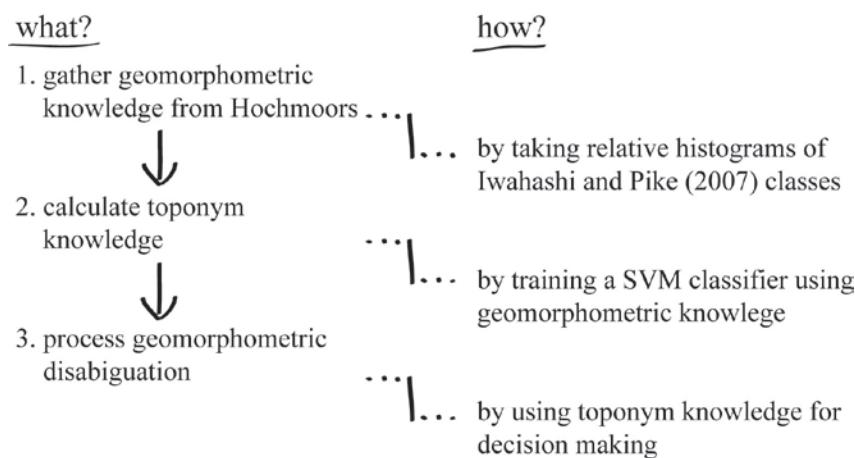


Figure 1. Workflow to process disambiguation with geomorphometric knowledge.

In a first step real Hochmoor locations ($n=100$) are used to infer geomorphometric knowledge. Thus, relative histograms for the 16 geomorphological classes introduced by Iwahashi and Pike (2007) are calculated for two windows of 0.25km and 5km centered on Hochmoor locations. The same is done for 1000 random locations within Switzerland (Figure 2).

What we term geomorphometric knowledge has become a vector with 32 dimensions, one vector for each Hochmoor and random location (16 classes for the 0.25km and 5km window respectively). The geomorphometric knowledge can be summarised as follows: In close proximity to Hochmoor centers (0.25km) topography is characterised by fine textures and gentle slopes (classes 9, 11, 13, 15). Steep slopes and coarse textures become more frequent if we widen the scale to the neighborhood of a Hochmoor (5km; classes 6, 8). This conforms to our notion of Hochmoor being plains in a mountainous environment, a secondary effect of the process of Hochmoor evolution.

The generated geomorphometric knowledge, in terms of location-vectors with 32 dimensions, is used to train a probabilistic SVM classifier (Burges 1998) to distinguish Hochmoor from random locations (probability is equal to the distance between vector and

² <http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/landscape/toponymy.html>

hyperplane). The classifier can be used to quantify geomorphometric Hochmoor probability for each designated set of coordinates. In our case we are interested in geomorphometric Hochmoor probabilities for all 2500 referent locations from the datasheets. At this stage geomorphometric Hochmoor probability has become what we term toponym knowledge.

In a last step we disambiguate toponyms using the generated toponym knowledge. In a most basic disambiguation scenario each of the 600 toponyms are disambiguated with the referent location of maximum geomorphometric Hochmoor probability.

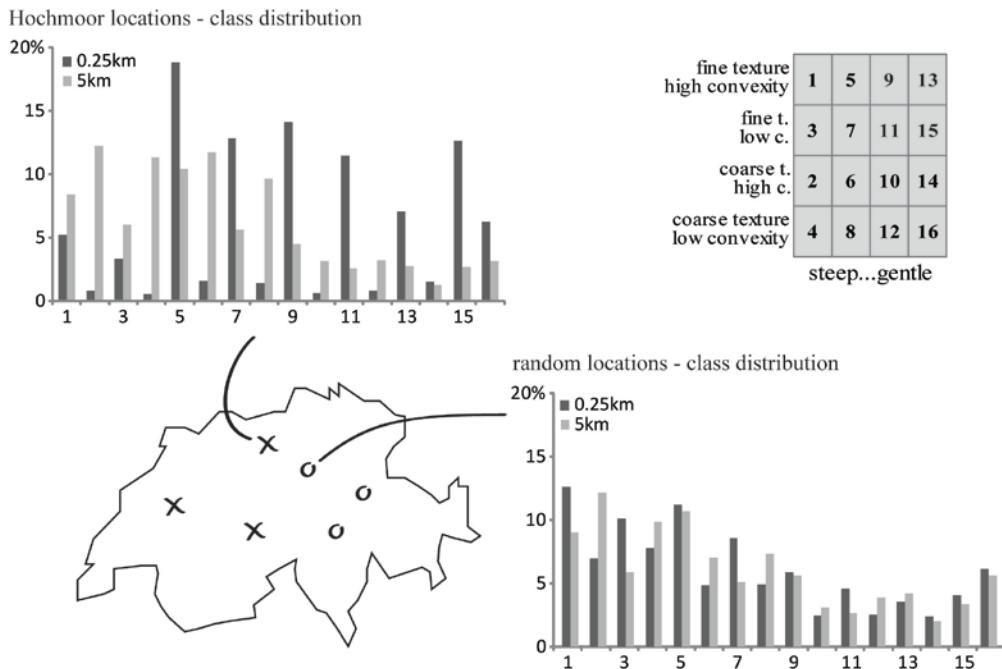


Figure 2. The 16 Iwashashi & Pike classes (upper right) and two typical relative histograms for a Hochmoor and a random location.

4. Geomorphometric disambiguation results

Here we focus on referent disambiguation of datasheets containing a single ambiguous toponym. All toponyms were manually semantic disambiguated in a previous step (e.g. removing instances of Bath where it is a place to wash and not a town).

There are 50 such single toponyms with 330 referent locations covering 20% of all datasheets. Single toponyms are the most complex case of toponym ambiguity, since knowledge gained from other, unambiguous toponyms, in a datasheet cannot be used to aid the process.

As is shown in the previous section only the referent location with the highest geomorphometric Hochmoor probability is resolved. In Figure 3 the Hochmoor probabilities for all 330 referent locations are plotted against the distance to the corresponding Hochmoor.

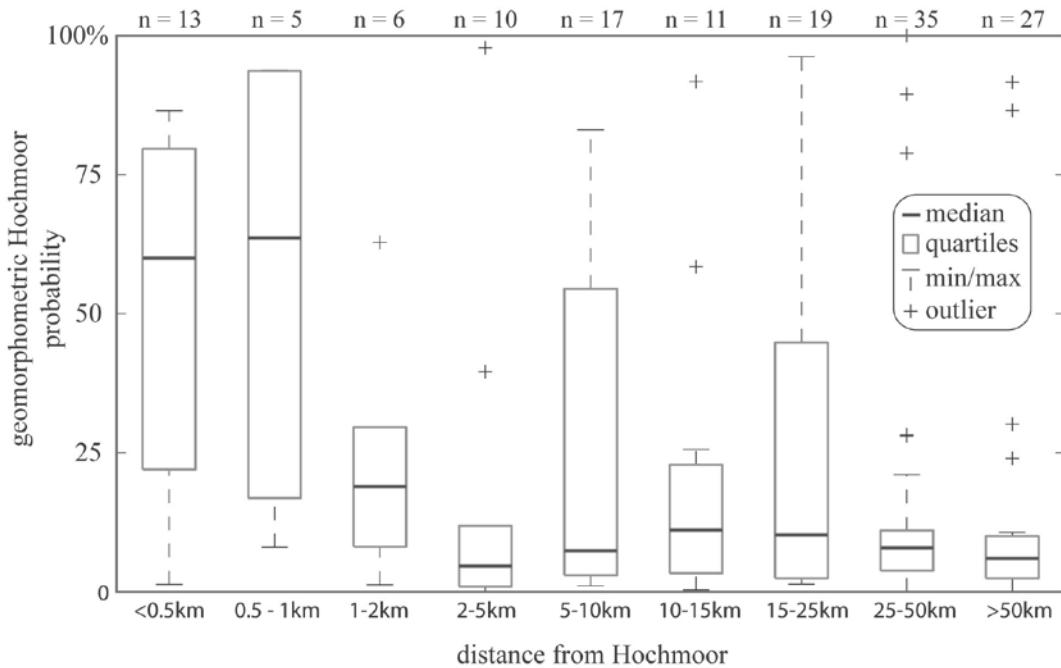


Figure 3. Boxplot of geomorphometric Hochmoor probability and distance to Hochmoor for 330 referent locations.

Figure 3 shows that geomorphometric Hochmoor probability is high for close referent locations and vice versa. In a nutshell, geomorphometric disambiguation allows us to resolve some 58% of the 330 referent locations. The baseline for disambiguation, i.e. the mean probability of successfully disambiguating toponyms by making a random decision, given no other information, is only 23%.

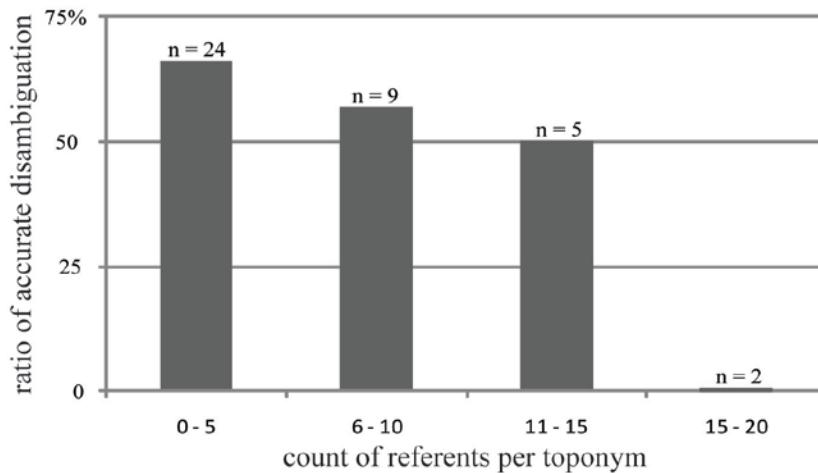


Figure 4. Disambiguation accuracy compared with count of reference locations of toponyms.

In Figure 4 the relationship between disambiguation accuracy and count of potential referents per toponym is visualized. The accuracy drops as the count of referent locations of toponyms increases.

5. Conclusions

Using the knowledge generated from geomorphometric characteristics of Hochmoor makes disambiguation more than twice as precise as the baseline (58% vs. 23%). Topography supplies substitute knowledge for cases where no a priori knowledge is available.

We used a rather basic approach to gather Hochmoor probability from topography. However, the same approach could be applied to all kinds of geographic objects (e.g hills, mountains or lakes).

Disambiguation with many referent locations is still inaccurate (Figure 4). Sometimes topographic Hochmoor probability is considerably higher for locations being far from the actual Hochmoor (Figure 3, outliers >25km). This may be due to false positive classifications, however, our inventory describes Hochmoor as classified at the present time, whilst geomorphometric characteristics describe locations with the affordance of being a Hochmoor, which may have been drained or otherwise altered in the last 200 years, which applies for some 85% of all original Hochmoor (Klaus 2007).

Many referent locations that are close to Hochmoor have rather small geomorphometric Hochmoor probabilities (Figure 3, minimas >1km). The assumption of spatial referents to Hochmoor always having a Hochmoor like topography is therefore clearly not always true.

In further work we will concentrate on resolving semantic ambiguity in landscape descriptions. We will face a very similar problem. Again there is no a priori knowledge that could serve for disambiguation. The general aim is to explicitly link landscape descriptions with space. This is the first important step to make naïve geographical knowledge in landscape descriptions useable.

6. Acknowledgements

The research reported in this paper is funded by the SNF Project 200021-100054.

7. References

- Bauer-Messmer B, Grütter R, Häggi M and five others, 2009, Service Oriented Architecture, Metadata Standards and Semantic Technologies in an Environmental Information System. In: Wohlgemuth V et al. (eds), *Proceedings of the 23rd EnviroInfo*, Berlin, Germany, 101-112.
- Burges CJC, 1998, A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, 2(2):121-167.
- Clough P, 2005, Extracting metadata for spatially-aware information retrieval on the internet. In: Jones C and Purves RS (eds), *Proceedings of the 2005 workshop on GIR*, Bremen, Germany, 25–30.
- Egenhofer M and Mark D, 1995, Naive Geography. In: Frank AU and Kuhn W (eds), *Spatial Information Theory A Theoretical Basis for GIS*, Berlin, Germany, 988:1-15.
- Iwahashi J and Pike RJ, 2007, Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. *Geomorphology*, 86(3): 409–440.
- Klaus G, 2007, Zustand und Entwicklung der Moore in der Schweiz, BAFU, Bern, Schweiz.
- Leidner JL, 2008, *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. PhD thesis, School of Informatics, University of Edinburgh

- Mark D and Turk A, 2003, Landscape categories in Yindjibarndi: Ontology, environment, and language. *Spatial Information Theory*, 28(2):28–45.
- Purves RS, Clough P, Jones CB and eight others, “The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet,” *International Journal of Geographical Information Science* 21, no. 7 (1, 2007): 717-745.

On Estimating Ecotone Occurrence from Land Cover Data Using Type 2 Fuzzy Sets

TUČEK Pavel¹, CAHA Jan², PECHANEC Vilém³

¹Department of Geoinformatics, Faculty of Science, Palacký University, tř. Svobody 26,
Olomouc, 77146, Czech Republic
Telephone: +420585634521
Email:pavel.tucek@upol.cz

²Department of Geoinformatics, Faculty of Science, Palacký University, tř. Svobody 26,
Olomouc, 77146, Czech Republic
Telephone: +420585634578
Email:jan.caha@klikni.cz

³Department of Geoinformatics, Faculty of Science, Palacký University, tř. Svobody 26,
Olomouc, 77146, Czech Republic
Telephone: +420585634579
Email:vilem.pechanec@upol.cz

1. Introduction

The term ecotone was first used in 1905 by F. E. Clements (1905) to describe visually different area between two ecological systems. Lately there has been a lot of attention to model and describe those transitional zones between ecological classes (Kilianová et al., 2009, Arnot and Fisher, 2007, Hufkens, 2008). Most of this research aims either to identify the best border between ecological systems or to identify ecotones as fuzzy objects mainly on data from remote sensing or some other very specific type of data (Fisher, 2006, Hufkens, 2008).

The aim is to identify areas where ecotones are most likely present using land cover and/or land use data, because those type of data are very common and can be easily obtained almost for any area of interest.

The idea of estimating ecotone occurrence from such data is based on several assumptions including facts that specific landscape indexes correlates with ecotone occurrence and that geometric characteristics of adjacent ecological areas can affect quality of the ecotone between those classes. However a great amount of uncertainty is present in this knowledge because so far no study proved exact link between those factors and ecotone presence. For those purposes fuzzy type 2 sets were used to incorporate the correct amount of uncertainty in the output.

2. Theory and Model

According to Holland et al. (1991) ecotones are defined as “zones of transition between adjacent ecological systems, having a set of characteristics uniquely defined by space and time scales and by the strength of interactions between adjacent ecological systems”. Such definition is applicable to ecological systems in any scale and the output ecotones thus may vary in their size from few centimeters to several kilometers (Holland et al.,

1991, Kiliánová et al., 2009). Another definition describes ecotone as area with high rate of change when compare to surrounding areas (Kiliánová et al., 2009). Same sources also claim that ecotone might contain more species and provide very specific conditions that couldn't be found in any of neighbouring area. Ecotones based on its characteristics can be linked with many ecological factors such as barrier, corridor or edge effect which makes the important part of landscape matrix. Because of the given characteristics is identification and monitoring of those spatial structures crucial to understanding biodiversity (Holland et al., 1991)

Several approaches on mapping ecotones exist. It is possible to represent them as crisp areas or as lines that have no area (Arnot and Fisher, 2007), but none of those is precise enough because the first treats ecotone as homogenous area which according to its definition isn't correct and the second omits the fact that ecotone may occupy quite significant area and thus representing it as line is too much generalization. The most correct representation of ecotone based of several sources (Arnot and Fisher, 2007, Kiliánová et al., 2009) that follows its definition is such where ecological systems are represented as spatial fuzzy sets and ecotone is an area that has specific degree of membership to more than one fuzzy set (Fig. 1). Different variations of this approach are presented in several sources (Arnot and Fisher, 2007, Hufkens, 2008). Given those reasons the fuzzy representation of ecotone seems the best for modeling both spatial extent as well as quality.

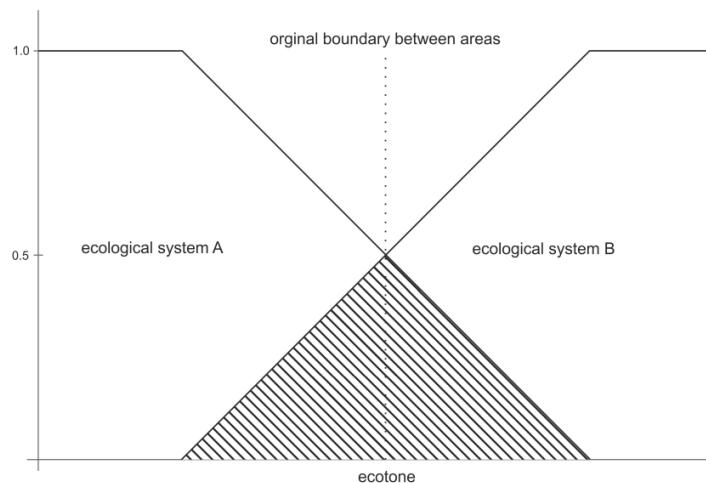


Figure 1. Representation of ecotone as intersection of two fuzzy sets

The main step in estimating the occurrence of ecotone with use of fuzzy sets is to fuzzify the input land cover data. As suggested above the landscape indexes, geometric properties and indexes of area and the relation between neighboring areas affect spatial extent and quality of ecotone. Fig. 1 shows how result of such fuzzifying may look like. Areas with membership value 1 are so called core areas of the ecological unit. Original boundary shows where originally was the border when area was classified into crisp sets of land cover categories. Wide of support of fuzzy set is defined by function that derive its result from values of several landscape indexes, geometric properties of area and relation to neighbor. In practical example the wide of support of fuzzy set for forest with very complex shape in highly heterogeneous landscape that neighbors meadow will be

much higher than for field with almost geometric shape in homogeneous landscape that neighbors road. This is based on premises that ecotones tend to be of higher quality and have bigger spatial extent in more heterogeneous landscape, on border of areas that have more complex shapes and between ecologically more stable and quality areas. All of mentioned parameters have impact on creating each area's fuzzy sets that determines areas zone of influence. Result ecotone is then created as intersection of two or more fuzzy sets. The quality of ecotone is determined based on ecotone's geometric properties and spatial statistics of overlapping fuzzy sets. The area occupied by the ecotone and the grade of union of membership values are the factors that are used in this part of model. This assessment of quality helps in estimating the uncertainty with which was the given ecotone's spatial extent calculated. Low quality ecotones tend to be of lower spatial extent, resulting in extreme cases to state called ecoline, almost crisp border between two ecological systems.

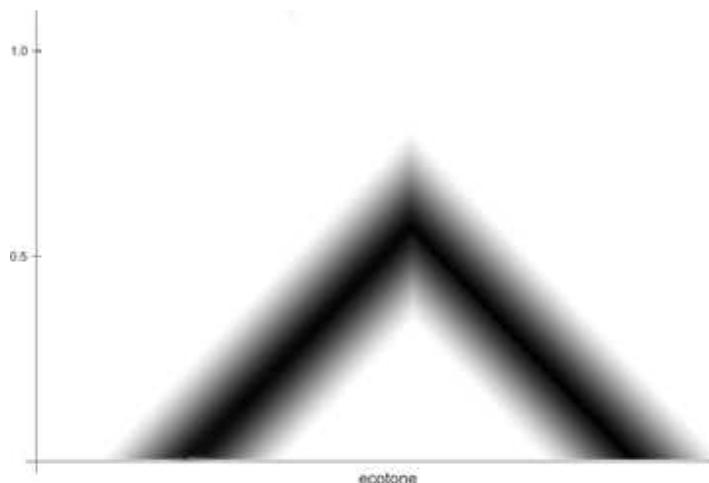


Figure 2. Representation of spatial extent of ecotone as type 2 fuzzy set

Quality of ecotone is in the model perceived as type 2 fuzzy set that modifies membership value into interval of values (fig. 2). The extent of this type 2 fuzzy set indicates how precise the estimation of spatial extent of ecotone is. This brings to the model fact, that for ecotones with low quality it could be much more complicated to estimate its occurrence and such ecotones are also much vaguer than the ones with higher quality.

Proposed model estimates occurrence of ecotones from common land cover and/or land use datasets and is suitable for modeling of ecotones in big scales. In the case study the aim was to catch even small ecotones that occur between roads and meadows as well as rivers and forests.

3. Case study

Area of interest is protected landscape area Litovelské Pomoraví located at north part of central Moravia between cities Mohelnice and Olomouc (fig. 3) with city Litovel being located almost exactly in the middle of protected area. The main reasons for protection are natural meanders of river Morava and floodplain forests that surround the river. The area is characterized by having many small ecological systems resulting in quite

often transitions between those various ecological units. Such locality provides optimal space for testing proposed model because it provides great diversity in land cover/land use types.

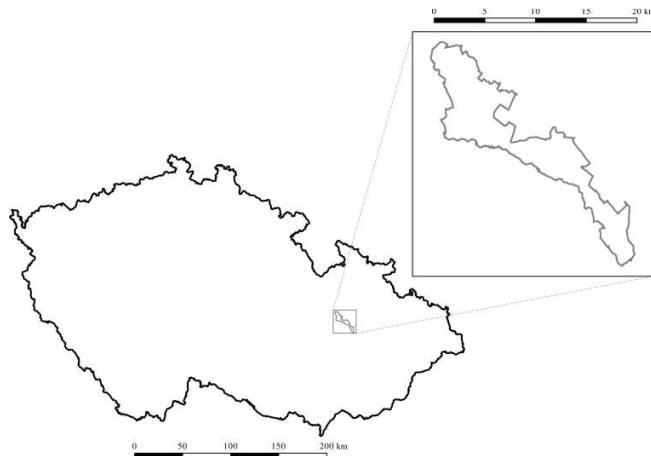


Figure 3. Localization of protected landscape area Litovelské pomoraví

6. Acknowledgements

This work has been supported by the Operational Program Education for competitiveness–European Social Fund (CZ.1.07/2.2.00/15.0276) and the by the Ministry of Education, Youth and Sports of the Czech Republic

7. References

- Arnot C and Fisher P, 2007, Mapping the Ecotone with Fuzzy Sets. In: Morris A and Kokhan S (eds), *Geographic Uncertainty in Environmental Security*, Dordrecht, The Netherlands, 19-33.
- Clements F E, 1905, *Research methods in ecology*. University Publishing Company, Lincoln, Nebraska USA.
- Fisher P et al., 2006, Detecting change in vague interpretations of landscapes. *Ecological Informatics*, 1:163-167.
- Holland M M et al, 1991. *Ecotones: the Role of Landscape Boundaries in the Management and Restoration of Changing Environments*. Chapman and Hall, New York, USA, 142 pp.
- Hufkens K et al., 2008, Estimating the ecotone width in patchy ecotones using a sigmoid wave approach. *Ecological Informatics*, 3:97-104.
- Kiliánová H et al., 2009, *Ekotony v současně krajině*. Vydavatelství UP, Olomouc, Czech republic.
- Rocchini D, 2010, While Boolean sets non-gently rip: A theoretical framework on fuzzy sets for mapping landscape patterns. *Ecological Complexity*. 7: 125-129.

Modeling Spatial Relevancy in Context-Aware Systems

Using Fuzzy Intervals

Najmeh Samany

PhD student, GIS Division, Dept. of Surveying and Geomatics Eng., College of Eng., University of Tehran,
Tehran, Iran
nneysani@ut.ac.ir

Mahmoud Reza Delavar

Center of Excellence in Geomatics Eng. and Disaster Management, Dept. of Surveying and Geomatic Eng.,
College of Eng., University of Tehran, Tehran, Iran
mdelavar@ut.ac.ir

Nicholas Chrisman

Department of Geomatic Science, University of Laval, Pavillon Casault, Québec, Canada
Nicholas.Chrisman@geoide.ulaval.ca

Mohammad Reza Malek

Dept. of GIS, Faculty of Geodesy and Geomatic Eng., K.N. Toosi Univ. of Technology, Tehran , Iran
mrmalek@kntu.ac.ir

Abstract

With an ongoing variety of pervasive computing devices integrated in our environment and an increasing mobility of users, it is necessary for mobile systems and services to be context-aware. Introducing relevant contexts to the user is the main properties of context-aware systems especially spatial relevant contexts. Most often as situations change gradually, there is no sharp boundary about how far one can see some relevant objects. It seems that contexts do not have crisp borders where they are true on one side but false on the other side. On the other hand, every context and mobile user has an influence interval in an urban network. So applying fuzzy spatial intervals for contexts and mobile users and defining their spatial relationships could effectively model spatial relevancy parameter. The main contribution of this paper is introducing fuzzy interval algebra for modeling spatial relevancy in context-aware systems. The proposed algorithm is implemented in a context-aware tourist guide system. The experimental results showed that the algorithm could accurately detect the spatial contexts.

Keywords: Fuzzy Spatial Interval; Context-aware; Interval Algebra; Tourist.

1 Introduction

Context-aware systems are computer systems that use context to provide more relevant services or information to support users performing their tasks, where context is any information that can be used to characterize the situation in which something exists or occurs (Vieira et al., 2010). The major challenge of the context-aware systems is to find an acceptable degree of information reduction to the relevant ones (Reichenbacher, 2005). Relevancy is a parameter which depends on the contexts supported by the system. Brown (1996) described that “context awareness”, is a

term that describes the ability of the computer to sense and act upon information about its environment, such as location, time or user identity". Because of the importance of location in fieldwork applications, the hand-held computers used are normally connected to a GPS receiver.

Reichenbacher (2005) modeled relevancy parameters and proposed some general rules of thumb for the assessment of relevancy that build a kind of hierarchy of relevant geospatial objects. He claimed that the bases of finding relevant contexts are physical and spatial relationship. Kwon and shin (2007) implemented a context-aware system "Location-aware COoperative Query system (Laco)". They modeled the spatial relations with metric distance and applied shortest path.

Review of the related researches proved that spatial relationship between the user and the contexts is a dominant factor for finding relevant objects in context-aware systems. However, it seems that more research to explore qualitative and quantitative spatial relevancy modeling is still needed.

The objective of this paper is to provide relevant information to the right situation for mobile users. We aim to model spatial relevancy parameters via spatial relationships between the user and his/her contexts. The main contribution of this paper is using fuzzy spatial interval algebra to model spatial relevancy in context-aware systems. It is assumed that the locations of users and the related contexts have a fuzzy spatial interval. A fuzzy spatial interval is a spatial interval which is not crisp and follows a fuzzy membership function. The spatial relationships between them model the spatial relevancy in the algorithm. The model is implemented in a tourist guide system scenario. The study area is a part of Tehran, capital of Iran.

2 Background

This section, briefly explains the concept of context-awareness and spatial relevancy. Then it is concentrated on the fuzzy interval algebra and its components.

2.1 Context-awareness and Spatial Relevancy

According to Dey's definition, context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves. Also a system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task (Dey, 2001 and Saracevic, 1996). Context appears as a fundamental key to enable systems to filter relevant information from what is available, to choose relevant actions from a list of possibilities (Hong et al., 2009; Chedrawy and Abidi, 2006), or to determine the optimal method of information delivery (Decouchant et al., 2009; Pan et al., 2007).

Saraceviev offers a general definition of relevance derived from its general qualities: "Relevance involves an interactive, dynamic establishment of a relation by inference, with intentions towards a context. Relevance may be defined as a criterion reflecting the effectiveness of exchange of information between people (or between people and objects potentially conveying information) in communication relation, all within a context" (Saraceviev, 1996, p.205).

Collecting data and acquiring context out of this data is inherently bound to a location. The information is fully relevant at this position. Generally, the relevance of the data declines with the distance from its point of origin (Schmidt, 2002). As seen from these observations locality of context is quite important and should therefore be included in the model as one of the basic relevant parameters which is called "spatial relevancy". Modeling this type of relevancy is necessary for context-aware services to provide appropriate information (Reichenbacher, 2005).

2.2 Fuzzy Spatial Relations in Context-aware Systems

It seems that contexts do not have crisp borders where they are true on one side but false on the other side. This fading, or fuzziness, is related to the relevance of the context. In fuzzy sets the main idea is that the membership of a component to a set is not just binary. It is rather fuzzy – meaning that an element has a degree of membership to a set (Schmidt, 2002).

The vagueness of the boarder of context is stemmed from the movement of the user, so by modeling the position of user with fuzzy spatial interval we could model spatial relevancy in an effective way. A fuzzy spatial interval is a spatial interval which is not crisp and follows a fuzzy membership function. The fuzziness of the spatial interval of the user has some characteristics including the following ones:

- 1) The most spatial relevancy is at the center of spatial interval which is the position of user called origin. The membership degree of origin is "1".
- 2) With increasing of distance from the origin, vagueness is increased and membership degree is decreased tending to "0".

Regarding these matters trapezoidal membership function is selected in this paper.

- 3) However, regarding the movement of the user with the car, we could specify the certain interval (the part of interval with membership degree equal to "1") rather than origin. This interval is determined by the velocity of movement. As the velocity increases, the distance of the interval increases.

Figure 1 illustrates the vagueness of the spatial interval of the user in a three-dimensional model in all directions. Trapezoidal function is used whose argument is the distance between the point of origin of the context value and any other point.

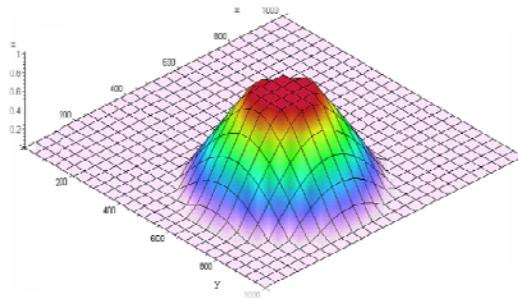


Figure 1: A three-dimensional model of spatial interval of the user in all directions

3 The Proposed Method

In the model, we define two main elements including the context and user. The user and context have an influence interval along a path. So applying spatial intervals for the users and contexts and defining the possible spatial relationships between them could model the spatial relevancy. As the user is moving along a road, a crisp spatial interval could not be defined for him/her. So a fuzzy spatial interval is assumed for the user. Fuzzy membership function of this interval is shown in Figure 2. The relationships between the spatial interval of the context and fuzzy spatial interval of the user model the spatial relevancy between them. If the membership degree of the fuzzy interval is equal to “1”, no vagueness exists in spatial relationships. Regarding to our implementation study area which is an urban space with the semi-congested traffic, the common velocity of the moving user in the main streets is supposed to be 60km/h, in this research, During 3 seconds, 50m distance is considered for the domain of user with membership degree of “1”.

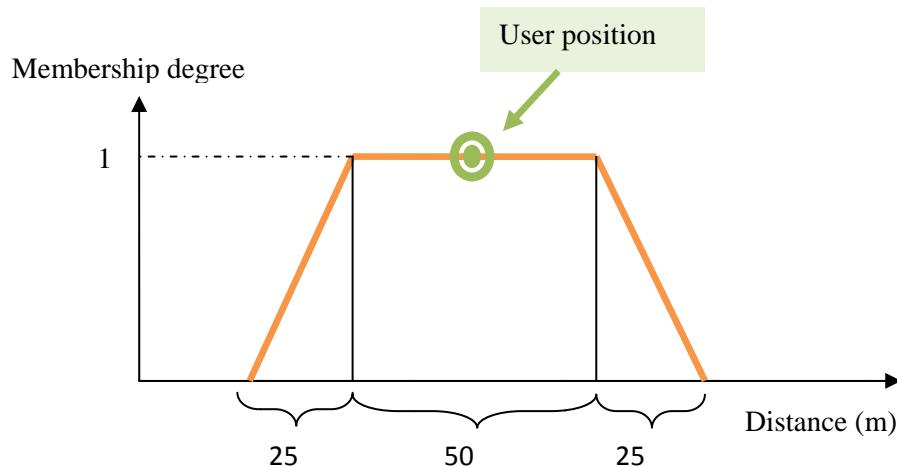


Figure 2: Fuzzy spatial interval

To model spatial relationship between the fuzzy spatial interval of the user and the spatial interval of the context, we applied Renz's (2001) spatial odyssey of interval algebra to define the spatial relationships. These relations are overlap (2), meet (2), Contain and inside (2), Covered

by (2), Covers (2), disjoin (2) and equal (1). Renz (2001) explained 26 spatial relationships between the directed intervals, however as we consider the intervals of the contexts non-directed, we are left with 13 relations (Table 1).

Table 1. The 13 basic relations for spatial relevancy model

Fuzzy Spatial Interval's Base Relations	Symbol
x behind = y	b=
x behind #y	b#
x meet from behind =y	mb=
x meet from behind# y	mb#
x overlaps from behind =y	ob=
x overlaps from behind #y	ob #
x contained-in = y	c=
x contained-in # y	c#
x contained-in the back of = y	cb=
x contained- in the back of # y	cb#
x contained-in-the-front-of # y	cf=
x contained-in-the-front-of # y	cf#
x equals = y	eq=

4 Case Study

We implemented the algorithm in Vb.net and developed a prototype in a tourist guiding system which consists of a mobile phone and GPS. The study area is in a part of Tehran.

The model is evaluated in a directed urban network for a user with different origins and destinations in the study area. Then the achieved results and predicted outputs are compared. 20 different routes for the tourists are considered. In each route a number of contexts have been considered as control points and the system is run while the user moves. Then the numbers of detected contexts compared with the control contexts are counted. Figure 3 depicts the detected contexts and control contexts. This comparison proved that the proposed approach could effectively model spatial relevancy parameters in the location-aware system.

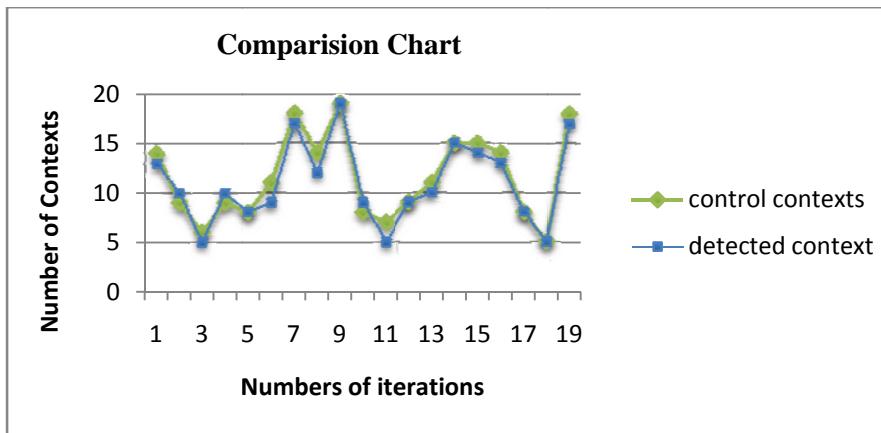


Figure 3. Diagram of the comparison between the control objects and detected contexts

Also we use the statistic indices for specifying the accuracy of the system. We compared the value of the achieved RMSE with the equivalent valid confidence level in 95% and 99% confidence levels. In other words, the value of RMSE which is equal to 1.38 is smaller than the value of errors at 99% confidence level (1.4) and 95% confidence level (1.6). So our statistics demonstrates that the proposed approach could effectively model spatial relevancy parameter in the context-aware system.

5 Conclusions

A fuzzy context model for spatial relevancy parameters in context-aware system has been proposed in this paper. Based on the proposed model, a fuzzy spatial interval is defined for the user and a spatial interval for the context are determined. The prototype system and evaluation results verified the performance. The model is able to meet the requirements of context-aware systems concerning limited memory and CPU resources in pervasive computing environments. The implementation of the context-aware system in an urban area is carried out based on the fuzzy model for a moving tourist. The experimental results show that the proposed approach would effectively detect spatial relevant contexts.

As a continuation to this work, we plan to work on modeling time as a context and presenting a spatio-temporal model for detecting spatio-temporal relevant contexts.

6 References

Brown P.J., 1996. The stick-e document: A framework for creating context-aware applications. In Proceedings of EP'96, Palo Alto, pp: 259-272.

- Chedrawy Z. and Abidi S.R., 2006. Case-based reasoning for information personalization: Using a context-sensitive compositional case adaptation approach. Proc. IEEE International Conference on Engineering of Intelligent Systems, Islamabad, Pakistan, Sep. 18, 2006, pp: 1-6.
- Decouchant D., Imaz G., Enriquez A. M., Mendoza S. and Muhammad A., 2009. Contextual awareness based communication and coauthoring proximity in the internet. Expert Systems with Applications, 36, pp: 8391–8406.
- Dey A. K., 2001. Understanding and using context. Personal and Ubiquitous Computing, 5: 4–7.
- Hong J., Suh E.-H., Kiim J. and Kim S., 2009. Context-aware system for proactive personalized service based on context history. Expert Systems with Applications, 36, pp: 7448–7457.
- Kwon O. and Shin M.K., 2007. LACO: A location-aware cooperative query system for securely personalized services, Expert Systems with Applications, 34(4): 2966-2975.
- Pan W., Wang Z., and Gu X., 2007. Context-based adaptive personalized web search for improving information retrieval effectiveness. Proc IEEE International Conference on Wireless Communications, Networking and Mobile Computing . Shanghai, China, Oct. 8-10 2007, pp: 5427–5430.
- Reichenbacher T., 2005. The concept of relevance in mobile maps, In Location Based Services and TeleCartography. Lecture Notes in Geoinformation and Cartography, Section III, pp: 231-246.
- Renz J., 2001. A spatial odyssey of the interval algebra: Directed intervals. In B. Nebel (ed.), Proc. of the 17th Znt 'I Joint Con on Artificial Intelligence. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, pp: 51-56
- Saracevic T., 1996, Relevance reconsidered, Proceeding, The Second Conference on Conceptions of Library and Information Science (CoLIS2), Copenhagen, Denmark, 14-17 Oct., 1996, pp: 201-218.
- Schmidt, A. 2002. Ubiquitous Computing – Computing in Context, PhD Thesis, Lancaster University.
- Vieira V., Tedesco P. and Salgado A.C., 2010. Designing context-sensitive systems: An integrated approach. Journal of Expert Systems with Applications, 38 (2): 1119-1138.

Map Comparison for the Evaluation of Spatial Bayesian Models

C. Robertson¹, J.A. Long², T.A. Nelson³, F. Nathoo⁴

¹Wilfrid Laurier University, 3rd Floor Arts Bldg,
75 University Ave West, Waterloo, ON, Canada, N2L 3C5
Telephone: (+1.519.884.0710 ext.4757)
Email:crobertson@wlu.ca

²University of Victoria,
Victoria, BC, Canada
Telephone: (+1.250.853.3271)
Email: jlong@uvic.ca

³University of Victoria,
Victoria, BC, Canada
Telephone: (+1.250.472.5620)
Email: trisalyn @uvic.ca

⁴University of Victoria,
Victoria, BC, Canada
Telephone: (+1.250.472.4693)
Email: nathoo@uvic.ca

1. Introduction

Spatial Bayesian models are increasingly developed to model spatially explicit processes in ecology and epidemiology. Spatial models of ecological spread processes, such as an infectious disease propagating through a human population or an invasive species spreading through a landscape, often have spatially distributed model parameters that account for variation spread rate (i.e., deviations from a global model such as a travelling wave)(e.g., Smith et al. 2002, Wheeler & Waller 2008). Local parameters are spatially varying coefficients with an estimated value at each spatial unit. In Bayesian models, each spatial parameter has a full posterior distribution available for inference. One advantage of aforementioned models is that each spatial unit has parameter estimates that can be used to provide spatial context about the spread process.

Validation of spatial models presents unique challenges. Typical validation approaches include some form of spatially global comparison between observed and expected values, such as the chi-squared (χ^2) test (e.g., Dice 1945). One initial problem with a global approach is that obtaining the true value for a theoretical spatial parameter describing some property of a complex ecological process is often difficult. Typically, assumptions are made based on results of field experiments taken over limited spatial scales (e.g., dispersal range in mark-recapture studies). Second, there may be spatial structure in the way that parameters themselves fit the data, and understanding the spatial structure of parameter estimates may reveal systematic errors that can be used to further refine the model. These two issues form the basis for the current research.

We employ a spatially explicit approach to the evaluation of spatial parameters in a Bayesian model-checking framework. Two approaches, posterior predictive checks

(Gelman 2005) and map comparison (Wang et al. 2004) are combined to provide evidence of model fit that includes information on spatial structure. We examine spatial structure when comparing maps of parameter estimates from a spatially local model describing the rate of spread across a study area. Our approach addresses the second problem of evaluating spatially local models. The first problem, knowing the true values of the parameters, is handled via Bayesian model checking. Simulation-estimation is a common approach whereby the fitted model is used to estimate new data which are then used to test model fit via a measure of discrepancy (Gelman et al. 1996). A case study using simulated data describing different spatial-temporal spread patterns is used to highlight our methodology.

2. Methods

2.2 Simulation-Estimation

In Bayesian modelling, uncertainties in parameter estimates are evident in the properties of the posterior distribution. If values are tightly clustered around the mean, there is strong evidence that the mean is a good estimate. Checking the model as a whole is more complicated. Posterior predictive checking is an approach whereby random draws from the posterior distributions of all model parameters are used to simulate new data sets, generally denoted as Y_{rep} , which we define as simulated replicates of the observed data. The Y_{rep} can be used to measure model fit with a general discrepancy measure such as the Deviance Information Criterion (Spiegelhalter et al. 2002), or more specific model test statistics. We re-estimate the model using the Y_{rep} datasets and compare the parameter estimates with known true values (i.e., those used to simulate the data). The comparison of these values at each spatial location forms the central problem of this research.

2.3 Map Comparison

The objective of map comparison here is to uncover similarities (or differences) in the spatial structure of expected and observed parameter maps. Examining spatial structure provides improved confidence in observed parameter estimates over purely aspatial comparisons. We selected the structural similarity (SSIM) index as an exploratory statistic for comparing maps (Wang et al. 2004). SSIM incorporates a Gaussian weighting function, to assess similarity across spatially local *regions*. SSIM does not require direct pixel to pixel comparisons, which ignore spatial structure and often produces overly critical comparison statistics (Pontius 2000). SSIM considers three components for map comparison: luminance, contrast, and structure, relating to local differences in mean, variance, and covariance respectively (Wang et al. 2004). Note that these three components are relatively independent, and changes in one component will not necessarily affect others (Wang et al. 2004). SSIM takes the following spatially local form, computing a similarity statistic for each spatial unit:

$$SSIM(x, y) = [l(x, y)]^\alpha \cdot [c(x, y)]^\beta \cdot [s(x, y)]^\gamma \quad (1)$$

where (x, y) denotes the spatial unit, l the luminance component, c the contrast component, and s the structure component (Wang et al. 2004). The exponents α , β , and γ can be used to weight individual components, with default values taken as $\alpha = \beta = \gamma = 1$. We report a mean global statistic for each of the three components and overall similarity. When two maps are identical, $SSIM = 1$, and values decrease as similarity decreases.

Expected and observed maps with low similarity in the luminance component are interpreted differently from those low in the structure component.

2.4 Case Study

To demonstrate the importance of spatial structure in model validation we implement the SSIM statistic comparison of data simulated from a model with spatially local parameters describing a spreading process. We specify a logistic model for a spatial spread process similar to Smith et al. (2002) where the logistic probability of an uninfected region (i) becoming infected at time t is defined as:

$$\log\{p_{it}/(1-p_{it})\} = \mu_t + \lambda_i NN_{[i,t-1]} \quad (2)$$

Here μ_t is a time varying parameter representing a baseline probability of infection; $NN_{[i,t-1]}$ is the number of infected neighbors of region i at time $t-1$; and λ_i is a spatially varying parameter quantifying the impact of infected regions on their uninfected neighbors. Our research here focuses on investigating the spatial structure of differences between the true values for λ and those estimated by the model. Values for spread were simulated as in Figure 1. These values were used to simulate data describing a spreading process on 40x40 grids over 100 time periods. As such, these represent the true values against which model estimates from the Y_{rep} data are compared via map comparison

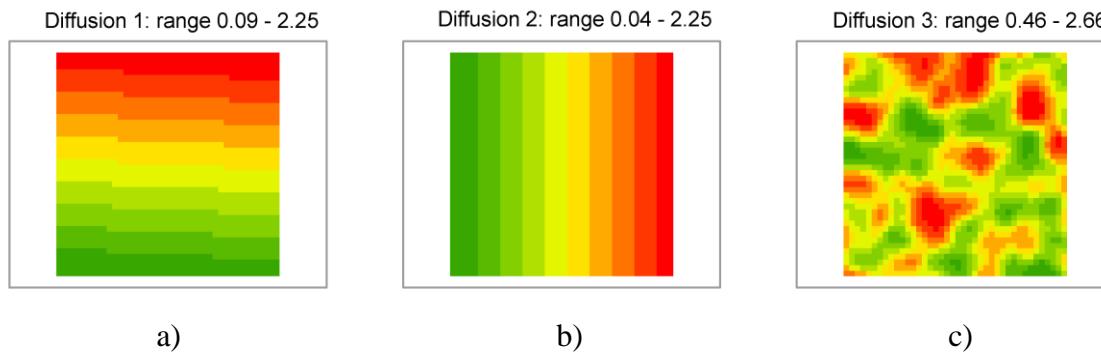


Figure 1. True values of diffusion parameters (λ) in three scenarios of a spatial spread process: a) Λ_1 , b) Λ_2 , c) Λ_3 . A range of μ -scenarios (M_1, M_2, M_3) were also used (not shown), generating nine spread scenario combinations.

3. Preliminary Results

Results of map comparison analysis on three types of diffusion spread represented in Figure 1.

	M	Luminance	Contrast	Structure	SSIM
Λ_1	1	0.924	0.864	0.889	0.710
	2	0.231	0.739	0.890	0.152
	3	0.951	0.881	0.899	0.753
Λ_2	1	0.681	0.870	0.898	0.532
	2	0.214	0.872	0.930	0.174
	3	0.824	0.858	0.897	0.634
Λ_3	1	0.974	0.827	0.697	0.561
	2	0.903	0.726	0.692	0.454
	3	0.972	0.856	0.720	0.599

Table 1. Map comparison analysis results comparing estimated diffusion to the true diffusion used to simulate data.

3. Discussion

Map comparison revealed that in some cases observed spread values were different from expected in terms of magnitude but the general spatial pattern of spread (structure component) was retrieved. The SSIM method enables creation of maps of local differences in mean, variance, and covariance, providing information on the spatial structure and differences in each which can be further explored to reveal systematic deficiencies in model development. Models that fit well based on aspatial validation tests do not always demonstrate good spatial agreement, warranting such a spatial approach. The approach we present for model evaluation is relatively simple and can be easily implemented with existing models (not exclusively Bayesian) providing valuable and unique insight on how the spatial structure of parameters relate to model performance.

4. References

- Dice LR, 1945, Measures of the amount of ecologic association between species. *Ecology* 26, (3): 297-302.
- Gelman A, Meng XL, and Stern H, 1996, Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733-759.
- Gelman A, Carlin JB, Stern HS, and Rubin DB, 2004, Bayesian data analysis. 2nd edition. Chapman & Hall, CRC Press, New York.
- Pontius Jr RG, 2000, Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering and Remote Sensing*, 66:1011-1016.
- Smith DL, Lucey B, Waller LA, Childs JE, and Real LA, 2002, Predicting the spatial dynamics of rabies epidemics on heterogeneous landscapes. *Proceedings of the National Academy of Sciences* 99:3668-3672.
- Spiegelhalter DJ, Best NG, Carlin BP, and Van Der Linde A, 2002, Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4): 583-639.
- Wang Z, Bovik AC, Sheikh HR, and Simoncelli EP, 2004, Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600-611.
- Wheeler DC, and Waller LA, 2008, Mountains, valleys, and rivers: The transmission of raccoon rabies over a heterogeneous landscape. *Journal of Agricultural, Biological, and Environmental Statistics*, 13:388-406.

On the Use of Grey Information Theory as a Conceptual Framework for Treating Uncertainty in Spatial Systems

M.W. Horner¹

¹Florida State University, Tallahassee, FL, USA
Email: mhorner@fsu.edu

1. Introduction

Developing effective means of theorizing and dealing with uncertainty continues to attract a great deal of interest in GIScience (Brown 1998, Rashed and Weeks 2003, MacEachren et al. 2005, Klugl et al. 2006, Lilburne and Tarantola 2009). During the last few decades, grey information theory has emerged in the engineering sciences as a means of better understanding uncertain information and processes in human and physical systems (Liu and Yi 2006). This paper explores the potential implications of grey information theory for conceptualizing uncertainty in GIScience, emphasizing its possible role in spatial modelling and geocomputation.

2. Background

Many spatial problems consist of both known and unknown elements. For example, in spatial models of facility location, candidate sites for prospective new retail stores are often known, but the demand for sited facilities' goods or services may not be certain (Hale and Moberg 2003, Snyder 2006). Similarly, when modeling urban land use change, base year land use (known) is subjected to a host of hypothesized processes (some unknown) to estimate future land characteristics (Li and Yeh 2000, Al-Ahmadi et al. 2009). In other situations, such as emergencies or extreme weather events (Elsner et al. 2006), there simply may not be sufficient historical spatial information to model a particular human behavioral response using traditional statistical approaches (Liu and Yi 2006).

One approach to deal with such situations is to treat uncertain problem constructs as 'grey' information. In this regard, grey information theory may be a useful way to ascertain uncertainty and provide an overarching formal organizational framework in many GIScience arenas. Essentially, grey information theory recognizes that some systems may consist of both completely known information and unknown information. It differentiates such information in terms of white (known) and black (unknown) information and their interrelationships (Huang and Fan 2005). A key idea in grey theory is the identification of the 'whitening function,' with the purpose of improving the level of uncertainty of a particular system parameter, process, or data instance. When this concept is extended to spatial systems, grey theory could be used to guide the design of new geocomputational tools intended to address and resolve uncertainty in modeling situations.

More broadly, the implications of grey information theory for GIScience are quite numerous, including the possibility of developing new ways of formalizing uncertainty in spatial problems and the logical decomposition of spatial problems into known and unknown components. A scan of recent research reveals there is relatively little exploration of grey concepts in the GIScience literature, with the few papers in

existence focussing on empirical and computational applications of selected grey concepts. (Yeh and Li 2001, Yang et al. 2009, Horner 2010). The present paper will comprehensively analyze some of the potential linkages between grey information theory and GIScience.

3. Overview

This paper consists of three major components. First, a more complete background on grey information theory is given, including providing general formulations for grey systems concepts including the whitenization functions and their possible instances in GIScience. Secondly, an effort is made to compare grey information theory with other related but different conceptualizations of uncertainty in GIScience, particularly those involving fuzziness (Fisher 2000; Rashed and Weeks 2003; Silvan-Cardenas et al. 2009), stochasticity (Sahinidis 2004; Lilburne and Tarantola 2009), as well as notions of complexity in general (Manson 2001; Ligmann-Zielinska and Jankowski 2007). Third, examples of grey systems are drawn from the author's work in spatial modelling for hazard management (Horner and Widener 2009), network uncertainty (Horner 2010), and household energy conservation behaviour (Horner et al. 2010). In these cases, grey information theory is critically discussed as a possible organizing principle for conducting uncertainty experiments and simulations of human behavioural response.

References

- Al-Ahmadi K, L. See, A Heppenstall, and J Hogg, 2009, Calibration of a fuzzy cellular automata model of urban dynamics in Saudi Arabia. *Ecological Complexity* 6 (2):80-101.
- Brown D, 1998, Classification and boundary vagueness in mapping pre-settlement forest types. *International Journal of Geographical Information Science* 12 (2):105-129.
- Elsner J, T Jagger, and A Tsomis, 2006, Estimated return periods for Hurricane Katrina. *Geophysical Research Letters* 33 (8):L08704.
- Fisher P, 2000, Sorites paradox and vague geographies. *Fuzzy Sets and Systems* 113 (1):7.
- Hale T, and C Moberg, 2003, Location science research: A review. *Annals of Operations Research* 123 (1-4):21-35.
- Horner M, 2010, Exploring the Sensitivity of Jobs-Housing Statistics to Imperfect Travel Time Information. *Environment And Planning B-Planning & Design* 37 (2):367-375.
- Horner M, and M Widener, 2009, Hurricanes, Networks, and Disaster Relief Planning. Paper read at North American Meetings of the Regional Science Association, November 2009, at San Francisco, CA.
- Horner M, T Zhao, and T Chapin, 2010, Exploring Research Opportunities for GIScience and Energy Sustainability. Paper read at Association of American Geographers Annual Meeting, April 2010, at Washington, DC.
- Huang C, and Y Fan, 2005, Knowledge structuring and evaluation based on grey theory. In *Fuzzy Systems and Knowledge Discovery, Pt 1, Proceedings*, 26-30.
- Klugl F, R Herrler, and G Andriotti, 2006, Coupling GIS and multi-agent simulation - towards infrastructure for realistic simulation. *Computer Systems Science and Engineering* 21 (3):197-206.
- Li X, and A Yeh, 2000, Modelling sustainable urban development by the integration of constrained cellular automata and GIS. *International Journal of Geographical Information Science* 14 (2):131-152.
- Ligmann-Zielinska A, and P Jankowski, 2007, Agent-based models as laboratories for spatially explicit planning policies. *Environment And Planning B-Planning & Design* 34 (2):316-335.
- Lilburne L, and S Tarantola, 2009, Sensitivity analysis of spatial models. *International Journal of Geographical Information Science* 23 (2):151-168.
- Liu S, and L Yi, 2006, *Grey Information: Theory and Practical Applications*. Edited by L. Jain and X. Wu, *Advanced Information and Knowledge Processing*. London: Springer.

- MacEachren A, A Robinson, S Hopper, S Garder, R Murray, M Gahegan, and E Hetzler, 2005, Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know. *Cartography and Geographic Information Science* 32 (3):139-160.
- Manson S, 2001, Simplifying complexity: a review of complexity theory. *Geoforum* 32 (3):405-414.
- Rashed T, and J Weeks, 2003, Assessing vulnerability to earthquake hazards through spatial multicriteria analysis of urban areas. *International Journal of Geographical Information Science* 17 (6):547-576.
- Sahinidis N, 2000, Optimization under uncertainty: state-of-the-art and opportunities. *Computers & Chemical Engineering* 28 (6-7):971-983.
- Silvan-Cardenas, J, L Wang, and F Zhan, 2009, Representing geographical objects with scale-induced indeterminate boundaries: A neural network-based data model. *International Journal of Geographical Information Science* 23 (3):295-318.
- Snyder L, 2006, Facility location under uncertainty: a review. *IIE Transactions* 38 (7):537-554.
- Yang F, G Zeng, C Du, L Tang, J Zhou, and Z Li, 2009, Integrated Geographic Information Systems-Based Suitability Evaluation of Urban Land Expansion: A Combination of Analytic Hierarchy Process and Grey Relational Analysis. *Environmental Engineering Science* 26 (6):1025-1032.
- Yeh A, and X Li, 2001, A constrained CA model for the simulation and planning of sustainable urban forms by using GIS. *Environment and Planning B-Planning & Design* 28 (5):733-753.

GeoComputation 2011

Eveline Wandl-Vogt, Mag^a

Project Manager Digital Dialect Lexicography
 Institute of Lexicography of Austrian Dialects and Names (I DINAMLEX)
 Austrian Academy of Sciences (ÖAW)
 1040 Vienna | Wohllebengasse 12-14/2

eveline.wandl-vogt@oeaw.ac.at
www.oeaw.ac.at/dinamlex
<http://wboe.oeaw.ac.at/projekt/personen/wandlvogt>

Abstract

GeoComputation and Dialect Lexicography: Ways to increase insight during an interdisciplinary partnership

Much information is inherently spatial in dialectology and dialect lexicography: the distribution of a word variant, the areal extent of a specific phonetic type, the movement patterns of morphologic types during time, the spread of a specific semantic realization, the source of a bibliographic reference, the birth location of a collector, etc. Both, geographers and dialect lexicographers have assembled massive amounts of analogue and digital information with spatial attributes.

In this talk the author reflects the role of GeoComputation and Geoinformation concerning projects of the human sciences dealing with examples of the field of dialect lexicography.

The author will give an overview of this interdisciplinary partnership and examples for increasing insight within the last 100 years (cf Lameli 2010, Ramisch et al 1997, Scholler 1973, Schrambke 2010, Schreibmann et al 2004). Furthermore she will discuss future prospects of the development (cf Göbl 2008, Perea 2008, Rumpf et al 2010, Wandl-Vogt 2010) of this fruitful partnership in the cyberscience surroundings (cf Nentwich 2003).

First, the author of this paper will discuss the role space plays in dialectology and – especially – in dialect lexicography and give some overview of the origins of dialect-lexicographic endeavour (cf Moulin 2010).

Second, she will inform about the lexicographers work on the example of the major dialect-lexicographic enterprises, also known as territorial dictionaries '*Territorial-wörterbücher*' or diatopic dictionaries '*diatopische Wörterbücher*', of the upper German dialect family (cf Badisches Wörterbuch, BWB, Ostfränkisches Wörterbuch, Schweizerisches Idiotikon, Schwäbisches Wörterbuch, SdWb, WBÖ). She will give a short overview of the mapping of them (cf Moulin 2010) and will focus mainly the handling of space and location information units in the dictionary context.

In the main part she will discuss how Geography (e.g. Cartography, Geoinformation, Geocomputation) and dialect lexicography were matched together formerly (e.g. on

dialect atlas projects, cf Lameli 2010, Schrambke 2010) and nowadays (cf Praxmarer 2010, Perea 2008, Rumpf et al 2010, Wandl-Vogt 2006 and 2010).

She will pick up especially the example of the project framework *Database of Bavarian dialects of Austria electronically mapped (dbo@ema)* [2007-2010] of the Institute for Lexicography of Austrian Dialects and Names of the Austrian Academy of Sciences to exemplify data analysis tools and presentation of results (cf Scholz et al 2008, Wandl-Vogt 2006 and 2010).

Within the framework of this project a web based system (cf wboe.oeaw.ac.at) consisting of a database for heterogeneous dialect data, a desktop-application to edit the data, a website to present the data and a web-application to visualize the data was established. The main focus of materials open to public since 1st of July 2010 is a collection of fungi (cf Piringer et al 2010) and source material belonging to the *Dictionary of Bavarian dialects of Austria (WBÖ)*, the mayor Austrian dialect-lexicographic enterprise of the Austrian Academy of Sciences (establishment of the nowadays institute in 1911, collection of materials since then, published since 1963, data base development since 1993).

She will give an outlook of future prospects; efforts that are at the time being part of linguistic yet not lexicographic research: She will discuss on some examples, how new methods of interdisciplinary dialectology and geo-information or geo-statistics could increase insight of dialect lexicography: dialectometry and dialect-lexicography (cf Göbl 1998, ders. 2008, Rumpf et al 2010), GIS-systems and dialect-lexicography (cf Praxmarer 2010, Scholz et al 2008, Wandl-Vogt 2010), analyzing tools for dialect-lexicography (cf Nerbonne 2010, Praxmarer 2010), visualising tools (maps for non-linguists cf Upton 2010) and perceptual dialectology (cf Preston 1989 and 2010). She will open the discussion with some proposals she suggests.

References

Peter Auer / Jürgen Erich Schmidt (2010): Language and Space. An International Handbook of Linguistic Variation. Volume 1: Theories and Methods. Berlin / New York. (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science / Manuels de linguistique et des sciences de communication 30.1).

Badisches Wörterbuch (1925-lfd.). Lahr bzw. München.

Bayerisches Wörterbuch (BWB) (1995-lfd.). München. (Bayerisch-österreichisches Wörterbuch. II. Bayern).

Hans Göbl (1984): Dialektometrische Studien. Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF. 3. vols. Tübingen.

Hans Göbl (2008): Die korrelative Dialektometrie. Eine Kurzvorstellung anhand von Beispielen aus AIS und ALF. In: Gerald Bernhard / Heidi Siller-Runngaldier (Eds.): Sprache im Raum - Raum in der Sprache. Frankfurt/Main u.a.: 67-90 (Spazi comunicativi / Kommunikative Räume, vol. 4).

Alfred Lameli (2010): Linguistic atlases – traditional and modern. In: Peter Auer / Jürgen Erich Schmidt (2010): 567-591.

Alfred Lameli / Roland Kehrein / Stefan Rabanus (2010): Language and Space. An International Handbook of Linguistic Variation. Volume 2: Language Mapping. Berlin / New York. (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science / Manuels de linguistique et des sciences de communication 30.2. + Maps).

Claudine Moulin (2010): Dialect dictionaries – traditional and modern. In: Peter Auer / Jürgen Erich Schmidt (2010): 592-612.

Michael Nentwich (2003): *Cyberscience. Research in the Age of the Internet*. Wien.

John Nerbonne / Wilbert Heergina (2010): Measuring dialect differences. In: Peter Auer / Jürgen Erich Schmidt (2010): 550-567.

Jacob Hald Pedersen / Thomas Eske Rasmussen (2007): Strategi I navigation. In: LexicoNordica 14 (2007) 71-80.

Maria-Pilar Perea (2008): Mapping morphological and phonetic features of Catalan. A general template for contemporary atlases and corpus. In: Dialectologia 1 (2008): 121-135.

Eva Pfanzelter / Christoph Praxmarer (2010): Geographische Informationssysteme (GIS): Ein Ort in den Geisteswissenschaften. In: Marin Gasteiner / Peter Haber (Eds.): *Digitale Arbeitstechniken für die Geistes- und Kulturwissenschaften*. Wien / Köln / Weimar: 251-259.

Barbara Piringer / Eveline Wandl-Vogt (2010): Österreichische Pflanzennamen. Eine Webapplikation für ein thematisches Korpus. In: Anne Dykstra / Tanneke Schoonheim (Eds.): *Proceedings of the XIV Euralex International Congress (14th EURALEX International Congress)*. Afük: 183; 774-779 (CDR).

Christoph Praxmarer (2010): Dialect relations in EDD. In: Barry Haselwood / Clive Upton (Eds.): *Proceedings of Methods XIII: Papers from the Thirteenth International Conference on Methods in Dialectology, 2008*. Frankfurt / Main: 153-159.

Dennis R. Preston: Mapping geolinguistic spaces of the brain. In: Alfred Lameli / Roland Kehrein / Stefan Rabanus (2010): 121-141.

Dennis R. Preston (1989): *Perceptual dialectology: nonlinguists' views of areal linguistics*. Dordrecht.

Heinrich Ramisch / Kenneth Wynne (1997): *Language in Time and Space. Studies in Honour of Wolfgang Viereck and the Occasion of his 60th Birthday*. In: *Zeitschrift für Dialektologie und Linguistik. Beihefte 97*. Stuttgart.

Jonas Rumpf / Simon Pickl / Stephan Elspaß / Werner König / Volker Schmidt (2010): Quantification and statistical analysis of structural similarities in dialectological area-class maps. In: *Dialectologia et Geolinguistica 18*: 73-98.

Hermann Scheuringer (2010): Mapping the German language. In: Alfred Lameli / Roland Kehrein / Stefan Rabanus (2010): 158-179.

Harald Scholler / John Reidy (Eds.) (1973): *Lexicography and Dialect Geography. Festgabe for Hans Kurath*. In: *Zeitschrift für Dialektologie und Linguistik, Beiheft Neue Folge 9*. Wiesbaden.

Johannes Scholz / Norbert Bartelme / Günther Fliedl / Marcus Hassler / Christian Kop / Heinrich Mayr / Jost Nickel / Jürgen Vöhringer (2008): Mapping languages – Erfahrungen aus dem Projekt dbo@ema. In: Josef Strobl et al (Eds.): *Angewandte Geoinformatik 2008. Beiträge zum 20. AGIT-Symposium*. Heidelberg: 822-827.

Renate Schrambke (2010): Language and space: Traditional dialect geography. In: *HSK 30.1*: 87-106.

Susan Schreibmann / Ray Siemens / John Unsworth (2004): *A Companion to Digital Humanities*. Oxford. <<http://www.digitalhumanities.org/companion/>> (2011.01.30)

Schwäbisches Wörterbuch (1904-1936). Tübingen.

Schweizerisches Idiotikon. Wörterbuch zur schweizerdeutschen Sprache (1881-1fd.). Frauenfeld.

Sudetendeutsches Wörterbuch. Wörterbuch der deutschen Mundarten in Böhmen und Mähren-Schlesien (1988-1fd.). München.

Clive Upton (2010): Designing maps for non-linguists. In: Alfred Lameli / Roland Kehrein / Stefan Rabanus (2010): 142-157.

Eveline Wandl-Vogt (Ed.): Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema). Vienna 2010-. <wboe.oeaw.ac.at> [2011.01.28]

Eveline Wandl-Vogt (2006): Mapping dialcets. Die Karte als primäre Zugriffsstruktur für Dialektwörterbücher. In: Karel Kriz et al (Eds.): Kartographie als Kommunikationsmedium / Cartography as a Communication Medium. Wiener Schriften zur Kartographie 17: 89-97.

Eveline Wandl-Vogt (2010): Point and find: the intuitive user experience in accessing spatially structured dialect dictionaries. In: Slavia Centralis 2 (2010) III, 35-53.

Wörterbuch der bairischen Mundarten in Österreich (WBÖ). Wien 1963-lfd. (Bayerisch-österreichisches Wörterbuch. I. Österreich).

Evaluation of ASTER and LISS III Data in Identification of Saline Soils, Case Study: Regions of Iran

S. K. Alavipanah(1), H. R. Matinfar(2), N. Sarmasti(3), M. Jafarbeglou(4), S. Goodarzi
mehr(5)

(1) Professor, Faculty of Geography, Department of Cartography, University of Tehran, Iran,
Email: salavipa@ut.ac.ir

(2) Assistant professor, Faculty of Agriculture, Department of soil science, Lorestan University, Iran,
Email: matinfar44@gmail.com

(3) High Expert, Department of Cartography, University of Tehran, Tehran, Iran,
Email: nsarmasty@yahoo.com

(4) Assistant professor, Faculty of Geography, Department of Cartography, University of Tehran, Iran.
Email: mjbeglou@ut.ac.ir

(5) M Sc student, Department of Cartography, University of Tehran, Tehran, Iran,
Email: goodarzi_1900@yahoo.com

Abstract

Salts tend to concentrate on the soil surface in dry and irrigated areas. As salinity increases, more salts will appear at the soil surface, favouring the use of conventional remote sensing tools. Rapid identification and large-scale mapping of salt-affected soils will help improve salinity management in watersheds and ecosystems. Potentiality of various sensors is important in detecting saline soils and salt crusts. Therefore, in this study we evaluated the data of ASTER and LISS III sensors in Playas of DAMGHAN, KASHAN and MAHARLOO regions, IRAN. The first, the imageries corrected and then we used PCA, NDVI and band ratioing in detecting of saline soils. In band ratioing method, two index were applied, NDSCI and RSCI. Investigation of feature space graphs in saline and non-saline soils indicated that NDSCI and RSCI had the best separability. Then, the maps of this soils prepared. In this maps, saline soils with salt crusts were perfectly clear. Relative calibration of visible and near infrared in applied approach showed, band ratioing and using of these two indexes(NDSIC and RSCI) were very efficient. In addition, in general the ASTER data were better than of LISS III data in separation of saline soils and non-saline.

Keywords: ASTER Data; Calibration; DAMGHAN; IRAN; KASHAN; LISS III; MAHARLOO; NDSCI; RSCI; Salinity; Salt crusts.

2.Methods

2-1-Study area

In this research three desert area include DAMGHAN, KASHAN and MAHARLOO were chosen. Figure 1 shows the situation of these areas in IRAN.

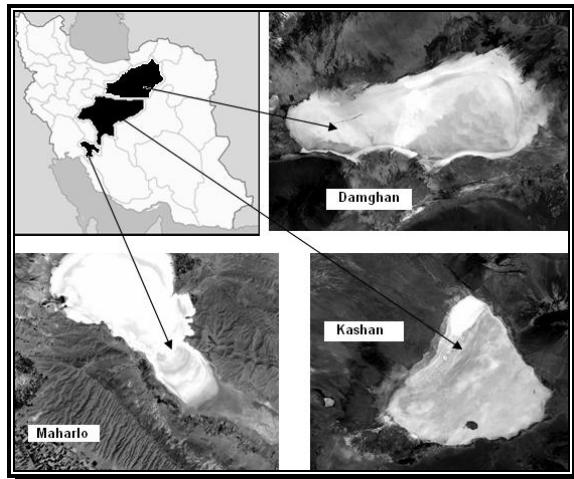


Figure 1. The location of study areas

2-2-Research methodology

2-2-1-Satellite images.

Table 1 shows the summary characteristics of sensors used.

Table.1 summary characteristics of sensors

Acquisition date	area	sensor
2007 NOV	DAMGHAN	
2007 JUL	KASHAN	
2006 JUL	MAHARLOO	LISSIII
2002 SEP	DAMGHAN	
2000 JUN	KASHAN	
2002 JUL	MAHARLOO	ASTER_L1B

2-2-2 Pre-processing of satellite data

In order to control the quality of used data and identify what extent of the systematic and non-systematic errors are fixed or remain after the systematic correction, the data were verified and determined all data used in this study, have standard corrections.

2-2-3 Alternative Calibration Method

For sensor calibration, due to the unavailability of sufficient information of study areas atmospheric conditions, the alternative method of relative calibration was devised. In this research the salt crusts of study areas were used for the relative calibration of LISSIII and ASTER sensors based on spectral reflectance method.

2-2-3-1 LISSIII sensor calibration

Using expression (1), the pixel DN of LISSIII sensor calibrated data convert to radiance:

(Slater, 1999). In this study, the following equation for the calibration of bands 2, 3 and 4 of LISSIII sensor were used:

$$L_{\lambda} = \left(\frac{L_{\max\lambda} - L_{\min\lambda}}{Q_{cal\max} - Q_{cal\min}} \right) Q_{cal} + L_{\min\lambda} \quad (1)$$

$$L_2 = \frac{148.005}{128} \times DN$$

$$L_3 = \frac{156.644}{128} \times DN$$

$$L_4 = \frac{164.543}{128} \times DN$$

2-2-3-2 ASTER sensor calibration

To convert from DN to spectral radiance of ASTER L1B calibrated data used equation(2):

$$L_{\lambda} = (DN - 1) \cdot UCC \quad (2)$$

Where L_{λ} is the sensor spectral radiance in $w/(m^2 \cdot sr \cdot \mu m)$, UCC is unit conversion coefficient in $(w/(m^2 \cdot sr \cdot \mu m))$ (Markham, 2005). In this study, the following equation for the calibration of bands 1, 2 and 3 of ASTER sensors were used :

$$L_1 = (DN - 1) \times 0.676$$

$$L_2 = (DN - 1) \times 0.708$$

$$L_3 = (DN - 1) \times 0.862$$

After convert and calculate the radiations of visible and NIR bands of LISSIII and ASTER sensors, the equation (3) were used to convert the radiance to reflectance:

$$\rho_p = \frac{\pi \cdot L_{\lambda} \cdot d^2}{ESUN_{\lambda} \cdot \cos \theta_s} \quad (3)$$

Where ρ_p is the amount of reflectance that is a quantity without unit, L_{λ} is the input spectral radiance of sensors ($w / m^2 \cdot sr \cdot \mu m$), d is the Earth-Sun distance in astronomic unit depends on Day, Year, Solar Zenith Angle, the time of image taking, latitude and longitude. $ESUN_{\lambda}$ is the solar exoatmospheric irradiance in the top of the atmosphere in band λ ($w / (m^2 \cdot sr \cdot \mu m)$), and θ_s is the Solar Zenith Angle at the image acquisition time in degree (Markham, 2004).

3- Results

3-1- Correlation of sensor bands LISSIII

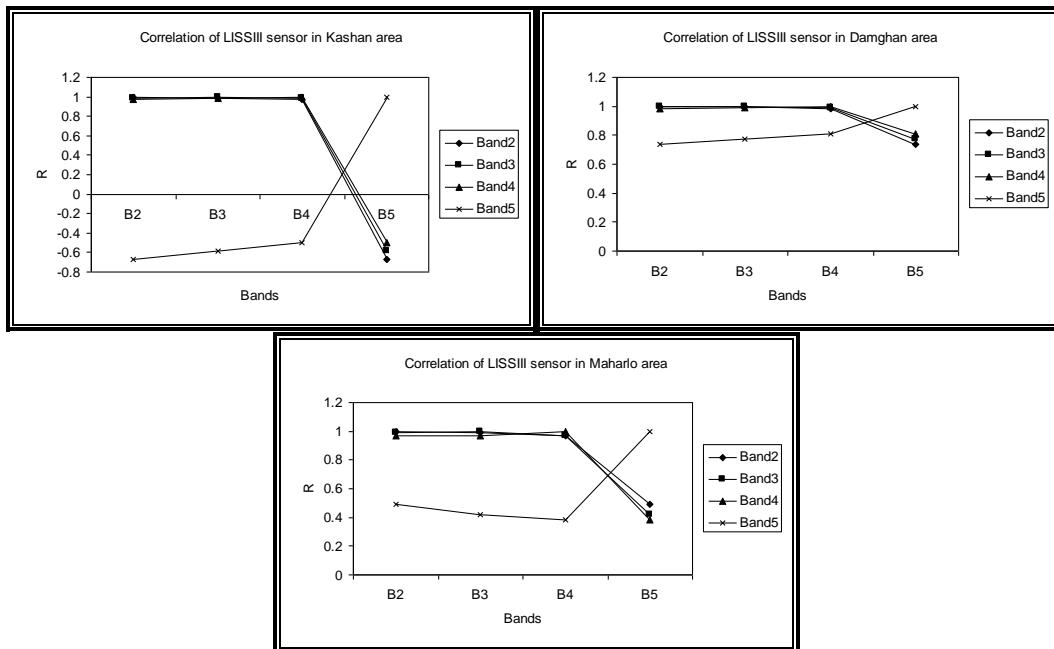


Figure 2. Correlation trend charts of LISSIII sensor in DAMGHAN, KASHAN and MAHARLOO area

As seen in figure (2), bands 2, 3 and 4 have a high correlation with each other. Band 5 shows less correlation with other bands, so this band has diverse information than others, thus band5 low correlation with other bands indicating the existence of fairly useful information on this band and requirement of its applications to identify the salt crusts.

3-2- Surveillance of bands correlation of ASTER sensor

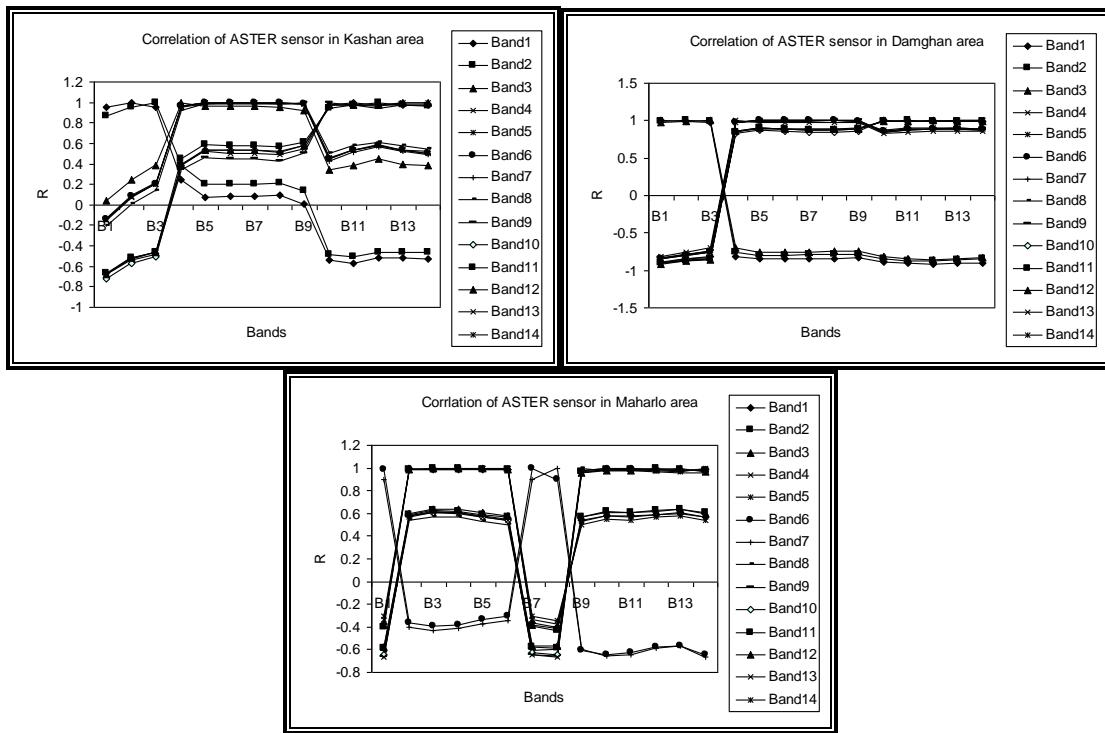


Figure 3. Correlation trend charts of ASTER sensor in DAMGHAN, KASHAN, and MAHAROO area

As shown in figure 3, band1 and band 3 have a great correlation. Near infrared, middle infrared and thermal infrared bands have a great correlation too. Therefore low correlation between near infrared, middle infrared and thermal infrared bands with visible bands indicates existence of useful information in these bands and necessity of use this information to recognize salt crust.

3-3- Principal Component Analysis

Variance percentage of principal component analysis (Figure 4) shows that more than 90% of information is concentrated on first and second component, after compressing. Thus, first and second component have useful information about salt crusts which extractable in both form of visual and digital.

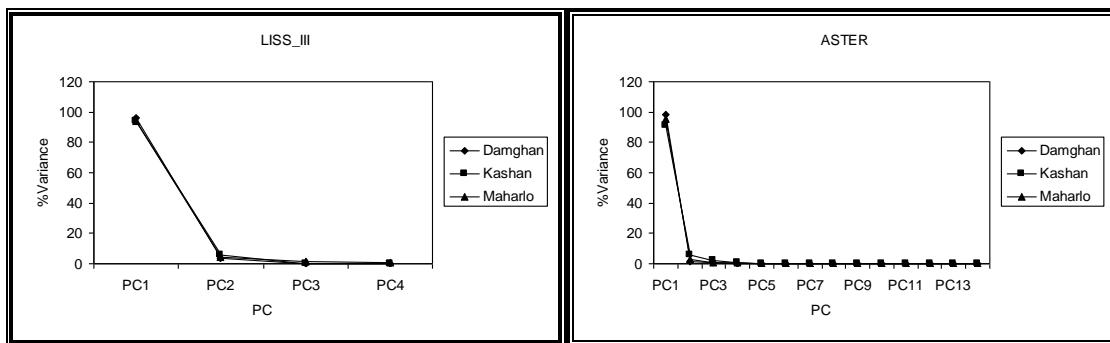


Figure 4. PCA Chart percent variance of LISSIII and ASTER sensors in DAMGHAN, KASHAN, MAHAROO

3-4- Spectral Rationing

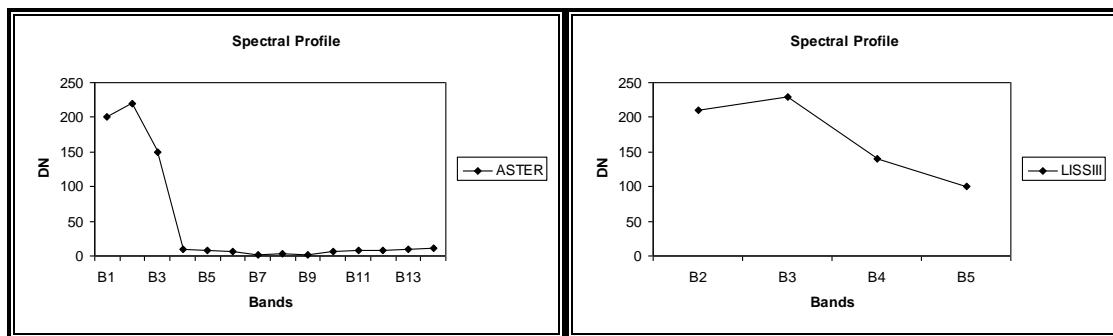


Figure 5. Average spectral reflection curves of DAMGHAN, KASHAN, and MAHARLOO salt areas in spectral band of LISSIII and ASTER sensor

On the basis of analysis of salt crusts spectral reflectance average curve, we can realize that difference value of salt crusts spectral reflectance in visible and middle infrared, is too high. In this research two new indices of salt crust have been introduced with regard to the characteristics of ASTER and LISS III sensors in arid and semi-arid condition of desert region.

3-4-1- Ratio Salt Crust Index

Ratio salt crust index is the simplest salt crust index and is defined as;

$$RSCI = \frac{LISSIII5}{LISSIII3} \cong \frac{ASTER4}{ASTER2} \quad (4)$$

This index, have a simple formula from the point of view of calculation and its values domain is between 0 and 1, so that the values toward 0, indicates salt crust.

3-4-2- Normalized Difference Salt Crust Index

Normalized difference salt crust index is defined as;

$$NDSCI = \frac{LISSIII3 - LISSIII5}{LISSIII3 + LISSIII5} \cong \frac{ASTER2 - ASTER4}{ASTER2 + ASTER4} \quad (5)$$

This index, have a similar treatment with RSCI salt crust index from the viewpoint of operation.

3-5- Thresholding

For evaluating separation capability of first and second component, NDVI vegetation index also RSCI and NDSCI indices, first, images of ASTER and LISS III sensor bands with

images obtained from first and second component, also NDVI vegetation index and RSCI and NDSCI salt crust indices, were put into a map list and then pseudoscopic color image were produced. Through visual interpretation on pseudoscopic color image, training classes of salt crust and non-salt crust were selected and displayed on monitor by feature space diagram simultaneously. After samples were selected, feature space diagram of pixel dispersion were evaluated and at the end, samples were modified by evaluation of feature space diagram of training classes with regard to this issue that pure pixel of salt, must separate from other pixels, to identify salt crust. Therefore capability of images which obtained from first and second component, also NDVI vegetation index and RSCI and NDSCI salt crust indices, was evaluated by feature space diagram. After repetitious examination of feature space diagrams, specified that RSCI and NDSCI salt crust indices shows the best discrimination capability for salt crust and non-salt crust classes. (Figure 6 and 7). Thus, in final step, threshold values between 0 and 255 were used to have more certainty about RSCI and NDSCI salt crust indices discrimination values. Different thresholds for RSCI and NDSCI salt crust indices were determined by trial and errors and finally thematic maps that include salt crust and non-salt crust classes were obtained, in which salt crust discriminated perfectly. (figure 8 and 9)

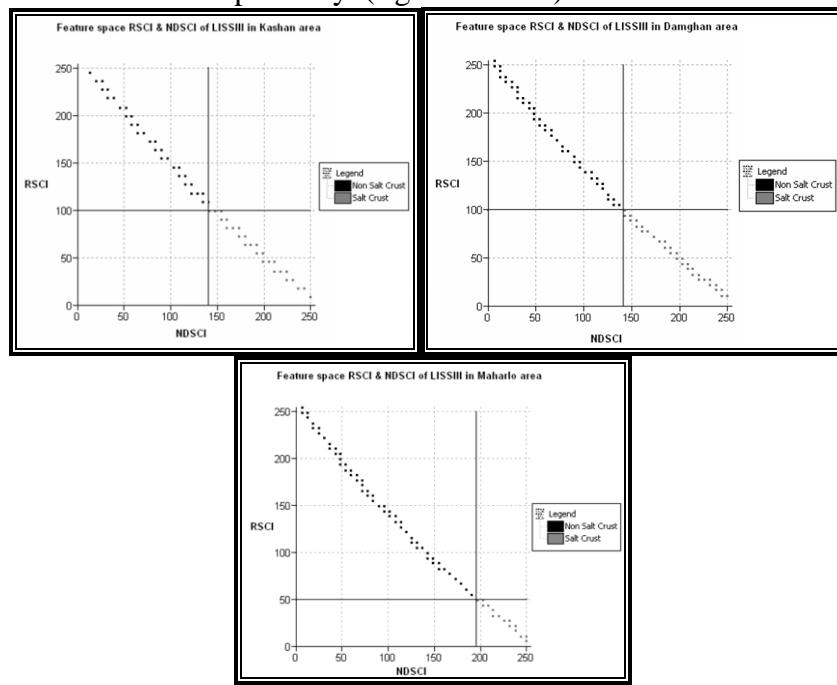


Figure 6. The scatter of pixels in two dimension space for RSCI and NDSCI indices of LISSIII in DAMGHAN, KASHAN, MAHARLOO areas

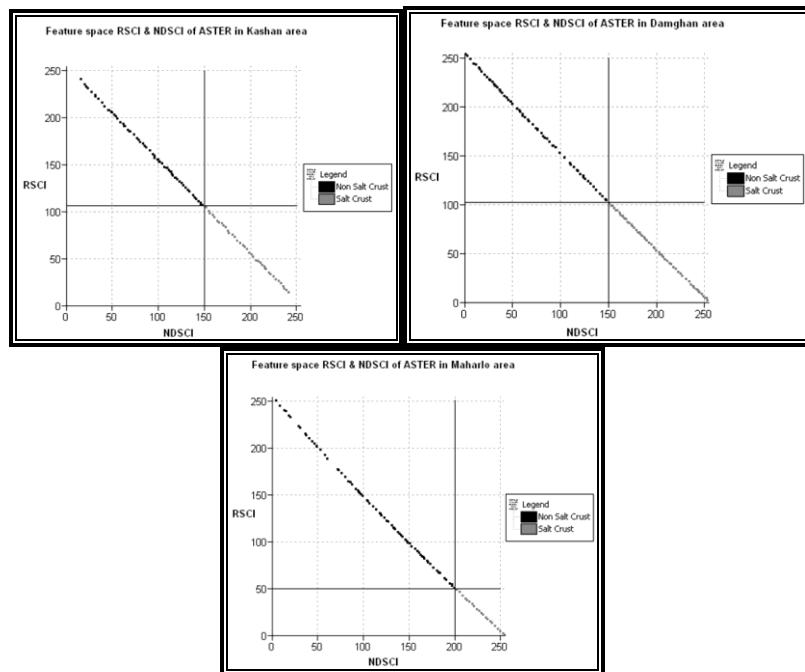


Figure 7. The scatter of pixels in two dimension space for RSCI and NDSCI indices of ASTER in DAMGHAN, KASHAN, MAHARLOO areas

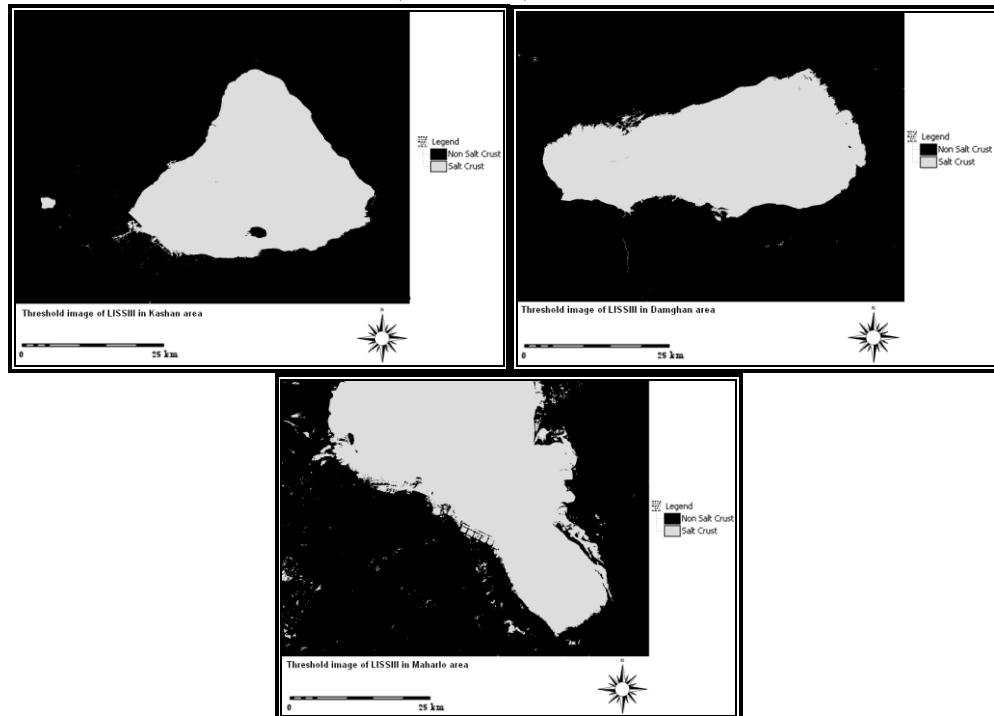


Figure 8. The image results of RSCI and NDSCI indices thersholding for LISSIII sensor in Damghan, Kashan and Maharloo area

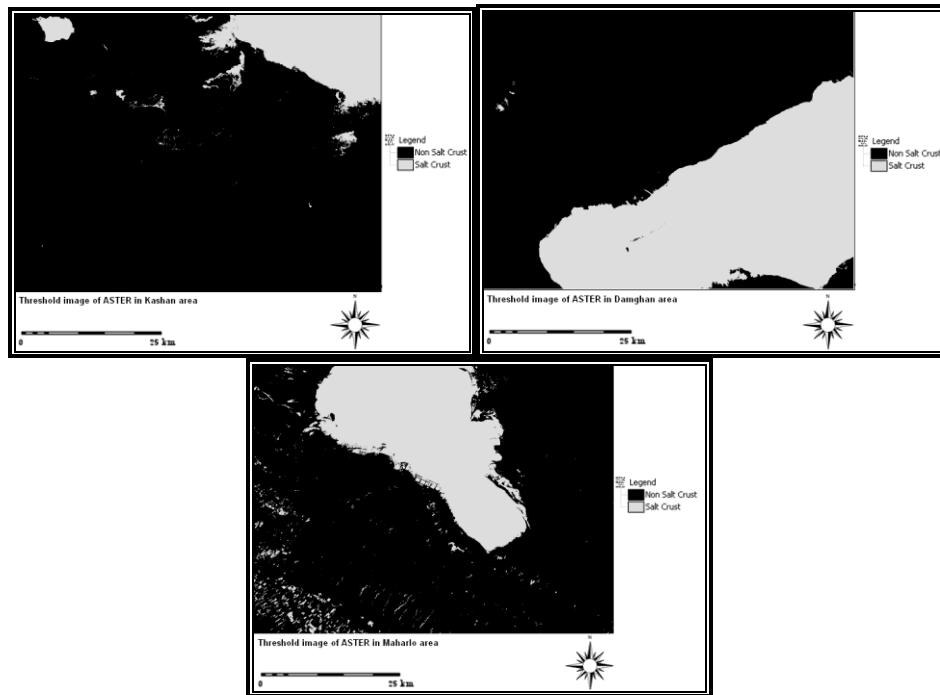


Figure 9. The image results of RSCI and NDSCI indices thersholding for ASTER sensor in Damghan, Kashan and Maharloo area

3-6- Superseding Calibration of ASTER and LISS III Sensors

After evaluation of salt crusts in studied areas, brightness value according to 1 and 2 equations, converted to receiving spectral radiation of satellite sensor. Correlation between brightness value and spectral reflectance was evaluated by calibration curve through overlaying of information layers of brightness value and calculated spectral reflectance for each visible and near infrared bands of ASTER and LISS III sensors, also performing cross instruction in ILWIS software. Figure 10 shows calibration curve of each visible and near infrared bands of LISS III sensor and Figure 11 shows calibration curve of each visible and near infrared bands of ASTER sensor in Damghan region. Calibration curve of Kashan and Maharlou, are the same with Damghan calibration curve.

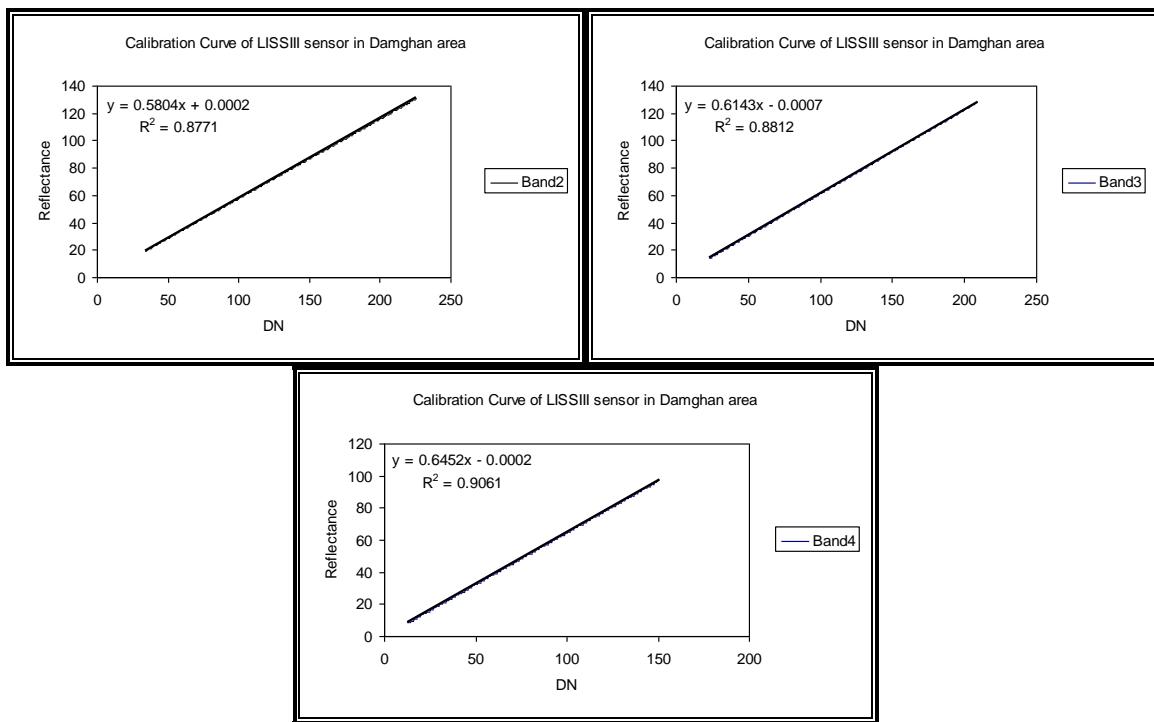


Figure 10. Calibration curve of visible and NIR bands for LISSIII sensor in Damghan area

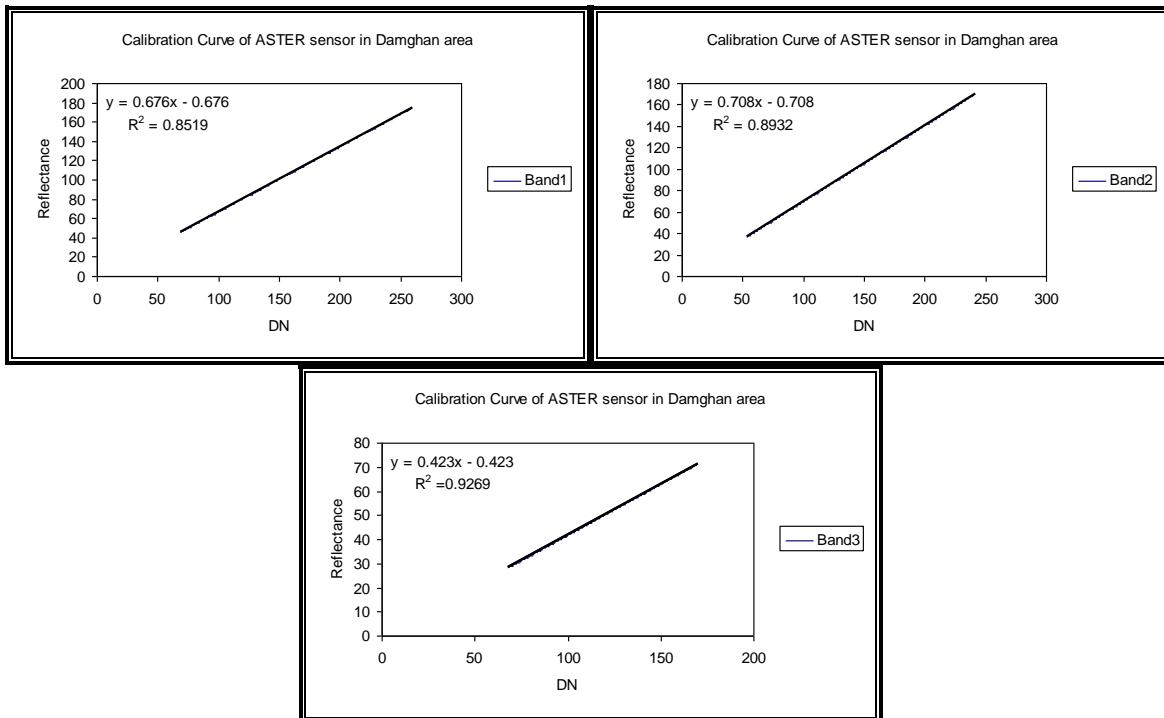


Figure 11. Calibration curve of visible and NIR bands for ASTER sensor in Damghan area

4- Conclusion

1- Analyzing the result of information obtained from ASTER and LISS III sensors shows that every three groups of ASTER bands have different band width and brightness value with respect to LISS III sensor .Thus the information of ASTER does not compare with the information of LISS III. In addition to, correlation process between the bands of LISS III and ASTER sensors is different with regard to geographic situation and salt marsh conditions of Damghan, Kashan and Maharloo, such as salt type, humidity, density of vegetation coverage and etc. Different status of salt marshes condition in different months or different seasons of the year is one of the effective parameters on band correlation.

2- Evaluating spectral reflection of salt marshes in different sensor bands can help us to study and identification of those. Optimum selection of bands is an important factor to salt marshes identification in spectral rationing method. In this method, reliable land data and information or personal experiment is important and required. Beside of complementary information requirement, much dependence on used sensor and its characteristics such as number of bands, band width in electromagnetic spectrum and its resolution, is one of the weak points of spectral rationing method. Also, accuracy of the results, depend on bands combination which is used for spectral rationing. Visible and middle infrared parts of spectral bands are suitable for identification of salt marshes. The reason of this circumstance, is, high reflection of salt marshes n visible and high absorption in middle infrared parts of electromagnetic spectrum. We can enhance and recognize the regions that have high percentage of salt by generate ratio between visible and middle infrared parts of electromagnetic spectrum. Therefore with accurate selection of visible and middle infrared spectral bands and generating ratio between them, salt marshes are recognized better. In this research by definition of salt marshes indices such as RSCI and NDSCI with regard to the characteristics of ASTER and LISS III sensors, salt marsh recognized better. The importance of salt marsh indices, RSCI and NDSCI, depend on the data type which used and salt covering surface. These indices can be use to study of the salt marshes of desert regions.

3- Thresholding is an information exploitation method to identify salt marshes which assign the amplitude pixel value to the desired class. Although accurate determine the desired amplitudes are not possible easily and always performed by trial and error. For perform thresholding, quantities of amplitudes are determine by user with regard to situation of the area, complementary information, science and adequate experiment. The best and optimum quantities are usually obtain in a trial and error process and examination of pixel dispersion in two-dimension diagram.

4- Comparison the correlation between brightness degree and spectral reflection of calibration diagram of each bands of visible and near infrared which obtained from ASTER and LISS III sensors, shows that use of salt marsh in maximum spectral reflection quantities, will reduce errors arising from atmospheric effects and sensor calibration. Existence of linear relation and high specification index in each visible and near infrared bands of ASTER and LISS III sensors, indicate that high percentage of changes (more than 80%) depend on spectral reflectance of salt marsh and low percentage of changes (less than

20%) impressed by unwanted parameters such as atmosphere and etc. therefore salt marshes can be used to calibration of satellite sensor in visible and near infrared bands.

References

- [1] Ehsani, A., 2001, "Study and separation Playa marginal soils by use of ETM+ Landsat sensor digital data in Damghan" Master's Thesis, Department of Natural Resources, Tehran university, 199 p.
- [2] Janfaza, E., 2007, "Study of salinity and type of solute by use of satellite data in Damghan area", Master's Thesis, Department of Natural Resources, Tehran university, 100 p.
- [3] komaki, CH. B., and S. Alavipanah, 2004, " Study of information classes spectral separation for Loot desert by use of satellite data", Geography researches, vole 54, p.13-28.
- [4] Alavipanah, S. K., Ehsani ,A and Omidi, P, 2003, " Desertification and land use change of Damghan Palaya study with use of multi temporal and multi spectral satellite data" , Desert jornal, J. 9, vol 1, P. 143-154.
- [5] Alavipanah, S. K, 2002, "Application of remote sensing in Earth science", Tehran university press, First edition, 478 p.
- [6] Matinfar, H., 2005, "Study of MSS, TM, ETM+, LISSIII and ASTER data to identify soil based on field studies and by use of Geographic Information System in Damghan area", PHD thesis, department of Agriculture, Tehran university, 400 p.
- [7] Matinfar, H., S. K. Alavipanah and F. Sarmadian, 2004, "Study of drylands soil spectral characteristics (Kashan), Desert jornal, vol 2, p 10.
- [8] Nematollahi, M. J., 2006, "Study of separation Damghan Playa marginal soils by use of ASTER sensor data", Masters Thesis, Department of Natural Resources, Tehran university, 199 p.
- [9] Arai, K., 2007, "Vicarious calibration of the solar reflection channels of radiometers onboard satellites through the field campaigns with measurements of refractive index and size distribution of aerosols", Remote Sens., Vol. 39, PP. 13_19.
- [10] Chen J., M. Y. Zhang, Le. Wang, H. Shimazaki, & M. Tamura, 2005, "A new index for mapping lichen-dominated biological soil crusts in deserts areas," Remote Sens. Env., Vol. 96, PP. 165_175.
- [11] Farifteh, J. A. & R. J. Farshad, 2006, "Assessing salt-affected soils using remote sensing", Solute modeling and geo., Vol. 130, PP. 191_206.
- [12] Howari, F. M., P. C. Goodel, & S. Miyamoto, 2002, "Spectral properties of salt crusts formed on saline soils", Environ. Qual., Vol. 31, PP. 1453_1461.
- [13] IRS_P6 users handbook , 2003, Hyderabad, India: NRSA. [Online]. Available: <http://www.nrса.gov.in/index.html>
- [14] "Landsat_7 science data users handbook", 2006, MD: NASA goddard space flight center. [Online]. Available:http://Ltpwww.gsfc.nasa.gov/IAS/Handbook/Handbook_toc.html
- [15] Markham, B. L., W. C. Boncyk, J. L. Barker, E. Kaita & D. L. Helder, 1996, "Landsat_7 ETM⁺ in_flight radiometric calibration", In Workshop on calibration of optical and thermal sensors, CNES, Toulouse.
- [16] Markham. B., & G. Chander, 1998, "Revised landsat_5 TM radiometric calibration procedures and post_calibration dynamic ranges", IEEE Trans. Geosic. Remote Sens., Vol. 41, PP. 2674_2677.
- [17] Markham. B. L., & G. Chander, 2003, "Landsat TM and ETM⁺ thermal band calibration", Can. J. Remote Sens., Vol. 29, PP. 141_153.
- [18] Markham. B. L, 2005, "Vicarious calibration of ASTER thermal infrared bands", IEEE Trans. Geosic. Remote Sens., Vol. 43. PP. 2733_2746.
- [19] Slater, P. N., S. F. Bigger, R. G. Holm, R. D. Jackson, Y. Mao, M. S. Moran, J. M. Palmer & B. Yuan, 1987, "Reflectance_based and radiance_based methods for the in_flight absolute calibration of multispectral sensors", Remot Sens. Env., vol. 22, PP. 11_37.
- [20] Slater, N., F. Bigger and J. Thome, 1996, "Vicarious radiometric calibration of EOS sensors", Remote Sens., Vol.13, PP. 349_358.
- [21] Scott, K. P., K. J. Thome, & M. R. Brownlee, 1996, "Evaluation of the Railroad Valley Playa for use in vicarious calibration," Proc. SPIE Conf., Vol. 2818.
- [22] Stuart F. & J. Thome, 2003, "Vicarious radiometric calibration of EO_1 sensors by reference to high_reflectance ground targets", IEEE Trans. Geosci. Remote Sens., Vol. 41, PP. 11174_1179.

- [23] Thome, K. J., "Absolute radiometric calibration of landsat_7 ETM⁺ using the reflectance_based method", 2001, Remote Sens. Env., Vol. 78, PP. 27_38.
- [24] Thome, K., B. Markham, J. Barker, P. Slater, & S. Bigger, 1997, "Radiometric calibration of Landsat", Photogram. Eng. & Remote Sens., Vol. 63, PP. 853_858.
- [25] Thome, K., S. Schiller, J. Conel, K. Arai, & S. Tsuchida, 1998, "Results of the 1996 earth observing system vicarious calibration campaign at Lunar Lake Playa, Nevada (USA)", Merologia., Vol. 35, PP. 631_638.
- [26] Tonnoka, H., F. D. Palluconi, S.J. Hook, & T. Motsunaga, 2005, "Vicarious calibration of ASTER thermal channels", IEEE Trans. Geosci. Remote Sens., Vol. 43, PP. 2733_2746.

Accuracy assessment for Fuzzy classification in Tripoli, Libya

Abdulhakim khmag, Alexis Comber, Peter Fisher

¹Department of Geography, University of Leicester, Leicester, LE1 7RH, UK

Tel. 00441162525148

Email: aek9@le.ac.uk

²Department of Geography, University of Leicester, Leicester, LE1 7RH, UK

Tel. 00441162523812

Email: ajc36@le.ac.uk

³Department of Geography, University of Leicester, Leicester, LE1 7RH, UK

Tel. 0044116253853

Email: pff1@le.ac.uk

Abstract

Satellite imagery is a longstanding and effective resource for environmental analysis and monitoring at local, regional and global scales. Thematic map accuracy continues to be problematic; especially when Boolean representations are used as each image pixel is assumed to be pure and is classified to one and only one class. In reality the pixel may be mixed, containing many classes. This paper will describe the field work that was undertaken to validate the fuzzy change estimates arising from fuzzy set classification. The main objective of this paper to carry out a comparative study of different accuracy assessment measures to check the accuracy of fuzzy classified images. By using different models to determine the validation of soft classification, to check the accuracy of fuzzy classified images, complete information about the class proportions in each pixel are required to be known.

Fuzzy classifications may be useful as multiple class memberships are assigned. A membership function is defined for each class against the feature value (digital numbers) and membership values of a class to belong to a particular pixel are determined based on function definition. Quantifying classification accuracy is an important aspect of map production as it allows confidences to be attached to the classifications for their effective end use. Accuracy measures serve as the analysis of

errors, arising from the classification process due to complex interactions between the spatial structure of landscape, classification algorithms, land cover change and sensor resolutions.. Therefore, other accuracy measures may appropriately including the fuzziness in the classification outputs and/or reference (ground) data. These include . Measure of closeness distance, Euclidean Distance, fuzzy set operators, and fuzzy error matrix based measure. Generally, the confusion matrix compares ground observations for a given set of validation samples with the classification result.

From the results of accuracy indices for user defined and actual classification, it can be said that all of the measures methods can be used successfully to check the fuzzy accuracy of classification

1. Introduction

The study area is located in North West Libya (the capital city Tripoli and surrounding regions) and this area contains different types of land use and land cover. These include urban, forest; agriculture area .The extent of land patches is frequently small leading to a prevalence of mixed pixels. The study area is subject to rapid changes in land cover and land use due to increases in population, and human activity and requirements for, more urban land, and food production.

In generally the accuracy assessment is based on the accuracy or confusion matrix, which compares ground truth data with the equal classification for a given set of validation samples (Congaltion et al., 1999; Foody, 2002). The accuracy matrix enables the source of the most common evaluation criterions firstly overall accuracy, secondly producer accuracy, finally user accuracy. A detailed overview is given by (Foody 2002; Congaltion et al. 1999).

For the assessment of soft classifications in general, various suggestions have been made such as fuzzy error matrix, Entropy, cross Entropy and cross tabulation (Binaghi et al., 1999; Foody, 1995; Woodcock et al. (2000); Green et al., 2004; Lewis et al., 2001; Pontius et al., 2006; Townsend, 2000). The fuzzy error matrix Binaghi et al. (1999) is one of the most attractive approaches, as it represents a generalization (grounded on the fuzzy set theory) of the traditional confusion matrix. Specifically, for a

cross-comparison to be consistent with the traditional confusion matrix, it is popular that the cross-comparison results in a diagonal matrix when a map is compared to itself, and that its marginal totals match the total of membership grades. More significantly, a cross comparison should convey readily interpretable information on the confusion between the classes. To date, the applicability of the fuzzy error matrix has been mostly concentrated on generating accuracy indices such as the overall accuracy, the user and producer accuracy, the kappa, and the conditional kappa coefficients (Binaghi et al., 1999; Okeke et al., 2006; Shabanov et al., 2005).

2. Field survey

The fuzzy land cover information have been generated from remotely sensed data (different fuzzy classification) identifies fuzzy memberships to five land cover classes (urban, vegetation, woody land, grazing land and bear area). There are five predicted fuzzy membership values for each pixel. I undertook some field work, recording the sub-pixel memberships at 210 locations. Each of the 210 pixels was sub-divided into 16 and the land cover recorded at each point. This gives me observed fuzzy memberships for the same five classes. In this paper we will compare the two sets of predicted and observed fuzzy memberships to determine some measure of fuzzy accuracy

3. Result and dissociation

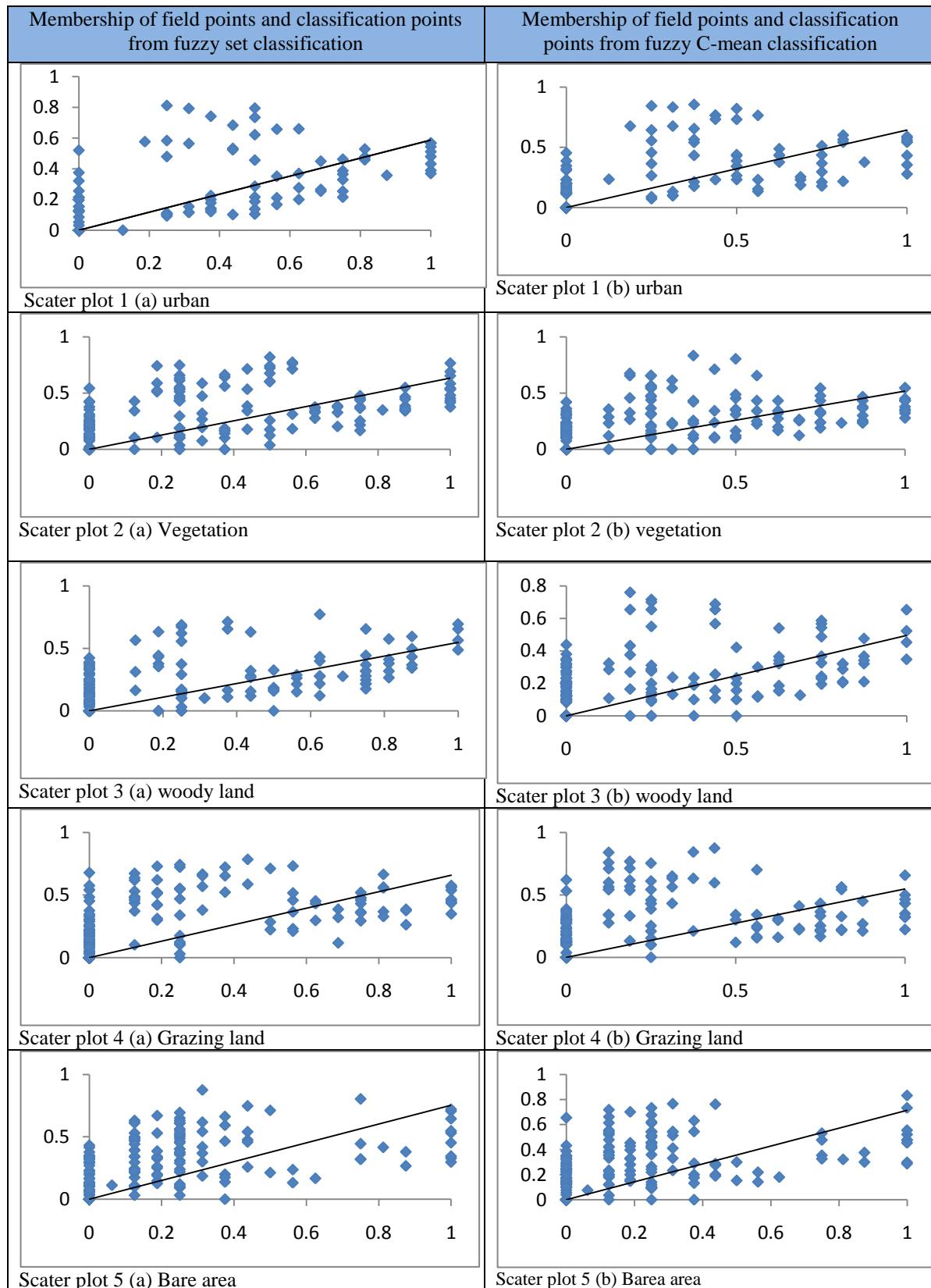


Table 1 Membership of field points and classification points

Generally these plots in table 1 show the degrees of membership of field points and classification points for all the classes, from the scatter plots there are many points scattered and there is a variation between the field points and classification points. The first column illustrates the field points and fuzzy set classification, the second column illustrates the field points and fuzzy C-mean, there is a bit difference between the two classification, these difference from training set which was taken it is not the same for both method. Generally the distribution of the points in both classifications is acceptable.

3. Regression

The regression was used to compare between the referenced data from the field and data from classification image. Table 2 illustrated regression statistics for multiple R and R^2 in the classes urban, vegetation, woody land, Grazing land and bare area, the result from fuzzy set classification and fuzzy C-mean, when the R and R^2 are high that means there are a good correlation and good classification. From the table we can see that the R^2 and multiple R is higher in fuzzy set compared with fuzzy C-mean in all the classes and the value of R and R^2 in the urban class is the highest in fuzzy set ($R=0.71725$, $R^2=0.51445$) and in fuzzy C-mean is ($R=0.69495$, $R^2=0.48295$), the lowest value of R and R^2 in the bare area class in fuzzy set is ($R=0.56127$ and $R^2=0.31663$), and in fuzzy C-mean ($R=0.48901$, $R^2=0.23917$) this gives indication that the urban class more accurate than the others, the reason for that the bare area and vegetation classes were changing from time to time and from season to season.

Class	R fuzzy set	R C-mean	R^2 fuzzy set	R^2 C-mean
Urban	0.71725	0.69495	0.51445	0.48295
Vegetation	0.62448	0.49900	0.38497	0.24987
Woody land	0.61410	0.51514	0.33712	0.26538
Grazing land	0.58384	0.44515	0.34087	0.22763
Bare area	0.56217	0.48901	0.31663	0.23917

Table 2 illustrated regression statistics for R^2 and multiple R for fuzzy set classification and fuzzy C-mean

4. Conclusion

Accuracy assessment of soft classifiers is still a big issue. This study studied methods to evaluate the performance of soft classifiers but they are sensitive to the use of a higher accurate proportion coverage of each informational class per pixel as a soft ground truth data which in practical situations is sometimes a bit difficult to obtained. It is needed to conduct further investigation on how we can assess soft classifiers taking into consideration the multiclass assignment problem and using soft ground truth data. Among these the Euclidean distance may be stated to be the best method since this measure takes into account the ambiguity and vagueness in the data, can be used for any probability distribution and provides a suitable accuracy index of classification also.

References

- Binaghi, E., P.A. Brivio, P. Ghezzi, and A. Rampini. "A Fuzzy Set-Based Accuracy assessment of Soft Classification." *Pattern Recognition Letters* (Elsevier Science), 1999: 935-948.
- Congalton, R.G., Green, K., 1999. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Lewis, Boca Raton.
- Foody, G.M., 2002, Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80, pp. 185–201.
- Foody, G.M. (1995). Cross-entropy for the evaluation of the accuracy of a fuzzy land cover classification with fuzzy ground data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 80, 185–201.
- Foody, G.M. "Approaches for the Production and Evaluation of Fuzzy Land Cover Classifications from Remotely-Sensed Data." *International Journal of Remote Sensing* 17, no. 7 (1996): 1317-1340.
- K., & Congalton, G. (2004). An error matrix approach to fuzzy accuracy assessment: The NIMA geocover project. In R. S. Lunetta & J.G. Lyon (Eds.), *Remote sensing and GIS accuracy assessment* (pp. 163–172). Boca Rato: CRC Press.
- Lewis, H. G., & Brown, M. (2001). A generalized confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*, 22, 3223–3235.

- Okeke, F., & Karnieli, A. (2006). Methods for fuzzy classification and accuracy assessment of historical aerial photographs for vegetation change analyses. Part I: Algorithm development. *International Journal of Remote Sensing*, 27(1-2), 153–176.
- Pontius, R. G., Jr., & Connors, J. (2006). Expanding the conceptual, mathematical and practical methods for map comparison. Proc. of the Spatial Accuracy Meeting 2006. Lisbon, pp., Portugal 16 (available from www.clarku.edu/~rpontius).
- Shabanov, N. V., Lo, K., Gopal, S., & Myneni, R. B. (2005). Sub pixel burn detection in Journal Moderate Resolution Imaging Spectro radio meter 500-m data with ARTMAP neural of networks. *Geophysical Research*, 110, 1–17.
- Townsend, P. A. (2000). A quantitative fuzzy approach to assess mapped vegetation classifications for ecological applications. *Remote Sensing of Environment*

Application of Data Mining In Micro-scale Urban Feature Analysis

A. Sokmenoglu¹, G. Cagdas², S. Sarıyıldız³

¹Istanbul Technical University & TU Delft
ITU Mimarlik Fakultesi Taskisla Taksim Istanbul, Turkey
+90 532 3420885
ahusokmenoglu@yahoo.com

²Istanbul Technical University
ITU Mimarlik Fakultesi Taskisla Taksim Istanbul, Turkey
+90 212 2931300
cagdas@itu.edu.tr

³Delft University of Technology
Design Informatics Faculty of Architecture,
TU Delft, Julianalaan 134, 2628 BL Delft, The Netherlands
+31 (0)15 27 85997
I.S.Sariyildiz@tudelft.nl

1. Introduction

This abstract introduces an ongoing research project addressing multi-dimensional and relational complexity of urban environments by the application of data mining as a methodology of knowledge discovery in micro-scale urban feature analysis. This research is an attempt to establish a link between knowledge discovery methodologies and automated urban feature analysis. After presenting our motivation, research questions and our methodology, an application of data mining of urban features will be briefly introduced in this abstract.

2. Motivation

By the beginning of the 1960's, as planning as a design-led practice seemed to fail to explain how urban processes occur, many urban theorists started to criticize the analysis of urban system from the perspective of few interrelated factors, without considering the multi-dimensionality of the system in a deductive fashion (Jacobs, 1961, Lefebvre, 1970, Harvey, 1973, Alexander, 1979). Hence, in the scope of this research, main motivation is that, in urban analysis, there is a need to advance from traditional one-dimensional (Marshall, 2004) description and classification of urban forms (e.g. Land-use maps, Density maps) to the simultaneous consideration of multi-dimensional aspects of urban systems. For this purpose, data mining is proposed as an analysis methodology for urban feature analysis. When applied to discover relationships between urban attributes, data mining can constitute a methodology for the analysis of multi-dimensional relational complexity of urban environments (Gil, et al., 2009). There are several recent studies of data mining applications in the domain of urban and geographical research such as works

of Demsar, 2006, Reffat, 2008, Behnisch and Ultsch, 2008, Liu and Seto, 2008, Cheng and Wang, Cheng and Anbaroglu, 2009 Christopoulou, 2009, Gil, et al., 2009.

3. Research Questions and Methodology

This research aims to address multi-dimensional and relational complexity of urban environments by applying data mining as a methodology of knowledge discovery in urban feature analysis, with a particular interest in exploring the patterns and relationships of micro-scale data in Beyoglu (a historical neighbourhood of Istanbul) as an application area. Two main research questions are formulated:

- What knowledge can be extracted from existing conventional urban analysis maps of Beyoglu, by the application of data mining methodologies? How this knowledge can be represented?
- Could data mining of urban attributes can produce valuable results and assist architects and urban planners at design, policy and strategy levels?

Within the scope of this research, a methodology is developed specifically for formulation and analysis of an urban database of Beyoglu. This methodology consists of the application of data mining into a GIS based urban database built out of official real data of Beyoglu, operating in three stages; Database formulation, Database analysis and Database evaluation. This methodology, applied in Beyoglu, is illustrated in Figure 1.

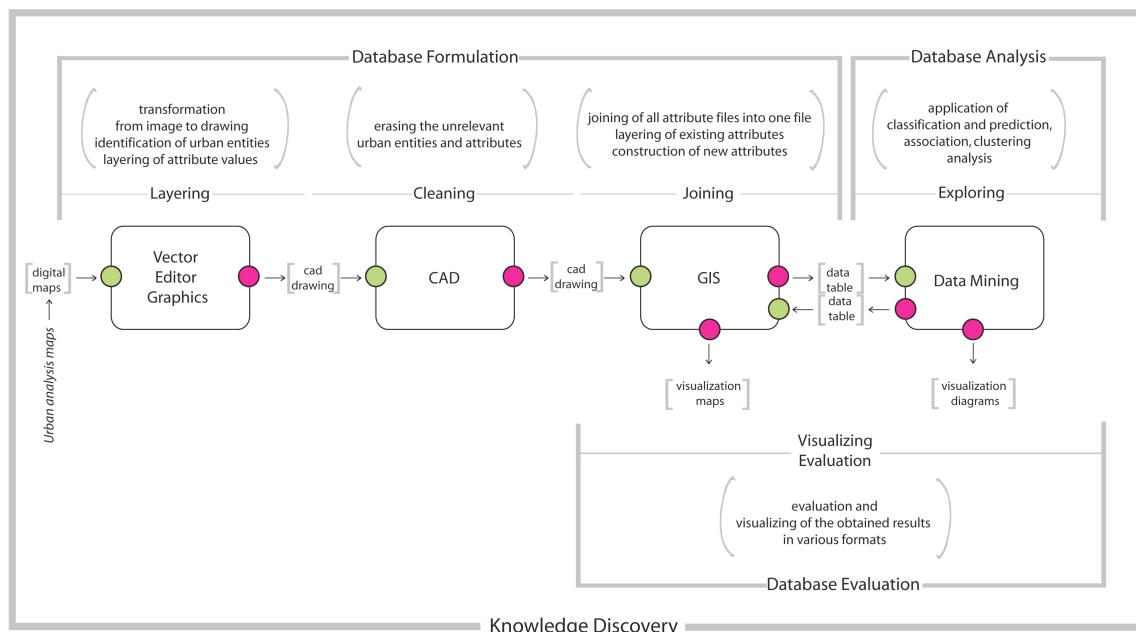


Figure 1. Methodology of knowledge extraction from urban data by data mining

In the following section, an application of the developed methodology will be explained through its stages.

3.1 Database Formulation

In this first stage, micro-scale urban data of Beyoglu is extracted from the various urban analysis maps of 2008 Master Plan of Preservation of Beyoglu provided by the Istanbul

Metropolitan Municipality (IBB). In Table 1, there is a list of urban feature data included in the database as urban attributes.

Attributes			
Att.1-17	Land Use_Ground Floor,	Att.16	Land Use_2 nd Penthouse
Att.2	Land Use_1 st Floor	Att.17	Land Use_3 rd Penthouse
Att.3	Land Use_2 nd Floor	Att.18	Neighborhood Name
Att.4	Land Use_3 rd Floor	Att.19	Density (Person/Ha)
Att.5	Land Use_4 th Floor	Att.20	Presence in the Bosphorus Silhouette
Att.6	Land Use_5 th Floor	Att.21	Building Maintenance Conditions
Att.7	Land Use_6 th Floor	Att.22	Building Construction Style
Att.8	Land Use_7 th Floor	Att.23	Empty floor ratio
Att.9	Land Use_8 th Floor	Att.24	Ownership
Att.10	Land Use_9 th Floor	Att.25	Density of Registered Buildings
Att.11	Land Use_10 th Floor	Att.26	Factor of Constructable Land (k.a.k.s)
Att.12	Land Use_1 st Basement Floor	Att.27	Registered Places for Preservation
Att.13	Land Use_2 nd Basement Floor	Att.28	Ground floor surface area
Att.14	Land Use_3 rd Basement Floor	Att.29	Distance to Galatasaray
Att.15	Land Use_1 st Penthouse	Att.30	Distance to Taksim

Table 1. Classification of processed urban attributes of Beyoglu

Available data of the historical neighbourhood of Istanbul covers several scales (from district to block, street, building and building floor) and different forms of classification themes including density, land-use, land value, ownership, material, physical conditions, road attributes, geological attributes and mobility infrastructure and more. There are 11,985 buildings, 700 building blocks, 30 neighbourhoods included in the urban database of Beyoglu preservation area (approx. 3,500,000 m²). The attributes (namely urban features of Beyoglu) of these buildings, building blocks and neighbourhoods are stored in the attribute table available in GIS. So far, in total, there are 30 attributes processed in the form of data table, ready for data mining, 27 attributes gathered from the Beyoglu Master Plan Analysis maps and 3 attributes calculated in GIS are processed in the form of data table.

3.2. Database Analysis and Evaluation

After the formulation of a micro-scale urban feature database for Beyoglu, this urban database is analyzed by Rapid Miner open-source software and the results are evaluated. The data mining analysis is concerned with the investigation of these generic questions;

- Are there significant recurrence patterns of attributes of the land? (Identification of groups, clusters, strata, or dimensions in data that display no obvious structure)
- How dependent and independent are these attributes? (Identification of associations and links among attributes, factors that are related to each other)

- How influential are these attributes on a particular urban phenomenon?
(Identification of factors that are related to a particular outcome of interest
(root-cause analysis)

Specifically, an analysis of data mining will be briefly introduced here, as an attempt to investigate second question listed above. Naïve Bayesian Method of Classification is applied for predicting the land use value of ground floor (Att.1) of the buildings by means of other attributes; land use value of first floor (Att.2), density of person (Att.19) living in the building and neighborhood (Att.18) where the building is located, distance to Taksim (Att.30), distance to Galatasaray (Att.29), building surface area (Att.28). In Table 2, below, there is a list of these attributes and their value range, subject to this data mining application.

	Attribute	Urban Entity Level	Values	Value Type
Att.1	Land Use_Ground Floor	Building Floor	{Residential, Business-Shopping, Social Infrastructure, Technical Infrastructure, Accomodation, Open Space, Empty, Other}	8 nominal categories
Att.2	Land Use_1st Floor	Building Floor	{Residential, Business-Shopping, Social Infrastructure, Technical Infrastructure, Accomodation, Open Space, Empty, Other}	8 nominal categories
Att.18	Neighborhood Name	Neighborhood	{Arap Camii, Asmalimescit, Bedrettin, Bereketzade, Bostan, Bulbul, Catmalimescit, Cihangir, Cukur, Emekyemez, Evliya Celebi, Firuzaga, Gumussuyu, Hacimimi, Huseyinaga, Kalyoncu Kullugu, Kamer Hatun, Katip Musafa, Kemankes, Kilicali Pasa, Kocatepe, Kuloglu, Mueyyetzade, Omeravni, Purtelas, Sahkulu, Sehitmuhtar, Sururi, Tomtom, YahyaKahya}	30 nominal categories
Att.19	Density (Person/Ha)	Building Block	{0-100, 100-200, 200-300, 300-500, 500-750, 750-1000, 1000-1500, 1500-2000, 2000+, non person living}	10 nominal categories
Att.28	Ground floor surface area	Building	{0-34 m ² , 35-48 m ² , 49-61 m ² , 62-81m ² , 82-114 m ² , 115-187 m ² , 187-17928 m ² } (Quantile Classification Method)	7 numeric categories
Att.29	Distance to Galatasaray	Building	{0-293 m., 294-451 m., 452-588 m., 588-721 m., 722-872 m., 873-1048 m., 1049-1508 m.} (Natural Breaks, Jenks Classification Method)	7 numeric categories
Att.30	Distance to Taksim	Building	{0-450 m., 451-693 m., 694-919m., 920-1178, 1179-1453m., 1454-1728m., 1729-2071m.} (Natural Breaks, Jenks Classification Method)	7 numeric categories

Table 2. Selected urban entities, their attributes and range of attribute values

Below in Figure 2, there is a Rapid Miner screenshot illustrating the process of data mining consists of applying a Naïve Bayesian learning operator and a cross-validation in order to estimate the performance of the learning operator.

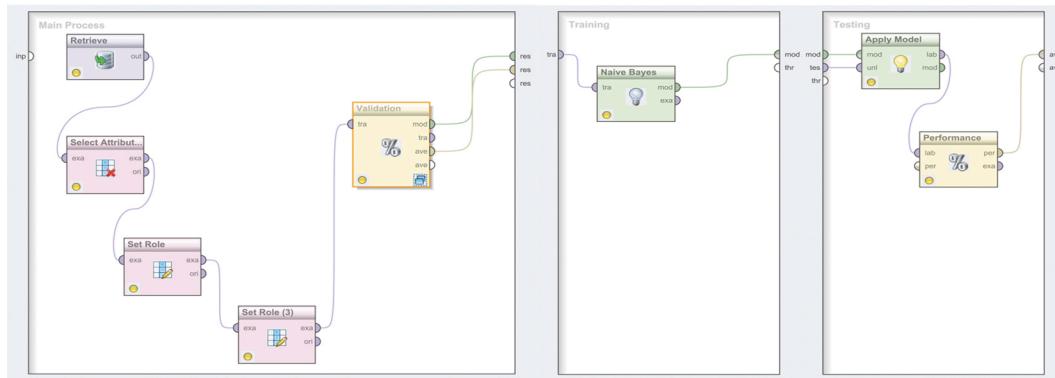


Figure 2. Process of data mining in Rapid Miner

First test is to predict Att.1 by Att.2. Results of this process can be seen in Figure 3, in the form of accuracy table, given by Rapid Miner software.

accuracy: 74.63% +/- 1.02% (mikro: 74.63%)											
	true Other	true Residential	true Business-Shoppin	true Accommodation	true Sociocultural Infra	true Technical Infrastru	true Empty	true Open Space	class precision		
pred. Other	688	138	428	4	409	19	64	1	39.29%		
pred. Residential	11	3689	1115	0	8	1	110	0	74.77%		
pred. Business-Shoppin	9	18	2923	0	12	1	86	0	95.87%		
pred. Accomodation	1	0	17	149	0	0	0	0	89.22%		
pred. Sociocultural Infra	0	2	69	0	296	8	3	0	78.31%		
pred. Technical Infrastru	0	0	1	0	0	10	0	0	90.91%		
pred. Empty	1	21	470	0	0	0	1009	0	67.22%		
pred. Open Space	0	0	0	0	0	0	0	139	100.00%		
class recall	96.90%	95.37%	58.19%	97.39%	40.83%	25.64%	79.32%	99.29%			

Figure 3. Accuracy table

As seen in the table, the overall accuracy of prediction is 74.63 %, which is significant in terms of claiming a dependency relationship between the land-use values of ground floor and first floors of the buildings in Beyoglu, in general. In case of residential use of ground floor for instance, the model predicts 3689 of the residential as residential and 179 of the residential as false, which gives a 95.37% class recall. More, the model predicts 149 of the accommodation as accommodation and 4 of the accommodation as false, which gives a 97.39% class recall. The model is successful in predicting the land-use values in case of other uses (96.90%), residential (95.37%), accommodation (97.39%), empty (79.32%) and open spaces (99.29%) uses. On the other hand, the model do not return significant results in case of business-shopping (58.19%), socio-cultural infrastructure (40.83%) and technical infrastructure (25.64%). This means that, to some extend in general, land use value of first floor of the building is dependent on the land use value of the ground floor. This hypothesis is especially valid in case of other uses, residential, accommodation, empty and open spaces uses.

After completing all the tests of this analysis with the rest of the attributes (Attributes 18, 19, 28, 29, 30) similar to the test introduced above, briefly we found that to a large extend in general, land use value of first floor is the most influential attribute among others, on determining the land use value of the ground floor. Neighborhood of the building and density of person living in the building are influential on determining the

land use value of the ground floor only in case of residential use with a class recall over %80. Distance to Taksim and Galatasaray (major transportation nodes in Beyoglu) are both not influential on determining the land use value of the ground floor, although in residential and business-shopping cases it can be claimed that there is a small degree of dependency which is over 60%. Surface area of the building is not influential on determining land use value of the ground floor in general, except significantly, in case of business-shopping use there is an accuracy level of 74.50%. These hypotheses must be certainly verified by means of other analysis methods in order to test their validity. Still the results are inspiring enough to expect that this kind of relational analysis methods of urban features could result in valuable site-specific knowledge.

4. Conclusion

Methodology of urban feature analysis applied in this research provides a multi-dimensional study of urban entities meaning that how each attribute of an entity is related to the other(s). Not only one kind of attribute is in interest, many of them are considered in a simultaneous manner. Departing from classical one-dimensional description of urban features' attributes, by means of the computational methods, this research looks for capturing the interrelations among those attributes. Hence, the focus of the analysis is on the relationships that exist within the order of an urban area rather than a conventional description of this urban order. More, microscopic or detailed view of urban system proposed in this research by relying on micro-scale data, provides a way of exploring urban system as complex as it is, allowing a deeper understanding of the system. Finally, data mining seem to provide a promising way of addressing multi-dimensional and relational complexity of urban environments by enabling to explore hidden patterns and relationships among urban features.

5. Acknowledgments

We would like to thank to Nuffic, the Netherlands Organization for International Cooperation in Higher Education, for funding of this research.

6. References

- Alexander, C, 1979, *The Timeless Way of Building*, Oxford University Press
- Behnisch, M, Ultsch, A, 2008 'Urban Data Mining Using Emergent SOM' in Preisach, C., Burkhardt, H., Schmidt-Thieme L., and Decker, R. (eds.), *Data Analysis, Machine Learning and Applications*, Springer Berlin Heidelberg.
- Behnisch, M and Ultsch, A, 2009, 'Urban data-mining: spatiotemporal exploration of multidimensional data', *Building Research & Information*, 37(5): 520-532.
- Cheng, T. and Wang, J, 2008, Integrated Spatio-temporal Data Mining for Forest Fire Prediction. *Transactions in GIS*, 2: 591–611.
- Cheng T., Anbaroglu B, 2009 Spatio-Temporal Outlier Detection in Environmental Data. *Spatial and Temporal Reasoning for Ambient Intelligence Systems Workshop*; 20- 25 September 2009, France.
- Christopoulou, K, 2009 *A Geographic Knowledge Discovery Approach to Property Valuation*, PhD Thesis.UCL
- Demsar, U, 2006, *Data Mining of Geospatial Data: Combining Visual and Automatic Methods*, PhD Thesis. KTH
- Gil, J, Montenegro, N, Beirao, JN, Duarte, JP, 2009 'On the Discovery of Urban Typologies: Data Mining the Multi dimensional Character of Neighbourhoods', in Çağdaş, G. and Çolakoglu, B. (eds.),

- Proceedings of 27th Conference on Education of Computer Aided Architectural Design in Europe*, Istanbul, Turkey, pp. 269-278.
- Harvey, D 1973, *Social Justice and The City*, University of Georgia Press.
- Jacobs, J, 1961, *The Death and Life of Great American Cities*, Random House INC., New York.
- Lefebvre, H, 1970, *The Urban Revolution*, University of Minnesota Press.
- Liu, W, Seto, KC, 2008, 'Using the ART-MMAP neural network to model and predict urban growth: a spatiotemporal data mining approach', *Environment and Planning B*, 35(2): 296 – 317.
- Marshall, S, 2004, *Streets and Patterns: The Structure of Urban Geometry*, Routledge.
- Master Plan of Preservation of Beyoglu, 2008, Istanbul Metropolitan Municipality (IBB).
- Rapid Miner Community Edition; <http://rapid-i.com/content/blogsection/7/82/lang,en/>
- Reffat, R, 2008, 'Investigating Patterns of Contemporary Architecture using Data Mining Techniques', in Muylle, M. (ed.), *Proceedings of 26th Conference on Education of Computer Aided Architectural Design in Europe*, Antwerpen, Belgium, pp. 601-608.

MetaHeuristics for a Non-Linear Spatial Sampling Problem

Eric M. Delmelle

Department of Geography and Earth Sciences

University of North Carolina at Charlotte

eric.delmelle@uncc.edu

1 Introduction

In spatial sampling, once samples of the primary variable have been collected, it is possible to augment the initial set by collecting additional measurements at other locations, a method known as second-phase sampling (Cressie 1991, Muller 1998, van Groenigen and Stein 1998 and recently de Gruitjer *et al.* 2006). Following a first sampling phase, the kriging variance is computed at each location using a covariogram function. Generally, additional observations are gathered away from existing points, that is where the kriging variance is large (see for instance Van Groenigen and Stein 1998). However, when the process under study is not stationary, sampling efforts should be directed in those strategic locations exhibiting strong spatial variation locally (Delmelle and Goovaerts 2009). In this paper, we formulate these two objectives into a single weighted-objective function -referred to as the weighted kriging variance-, where the weights reflect the roughness of the spatial process.

This objective function is highly non-linear (inversion of covariance matrices), and calls for robust heuristic methods. Additional samples can be collected *sequentially*, for instance by adding one sample at a time to the initial set. This procedure may be suboptimal but fast since it requires the inversion of a matrix augments by only one entry.

Practically, a covariogram summarizing the spatial variation in the observed variable with distance is determined following the collection of initial samples. Based on the covariance structure, the kriging variance is computed at each grid node, and weighted by the local variation at that node. The objective consists of locating those additional samples strategically to maximize the change in weighted kriging variance. Heuristic methods decide on the location of new samples. For instance a greedy algorithm will allocate additional observations on the peaks of the weighted kriging variance

surface but these local maxima may not be optimal to the objective function. In this paper, we propose a combination of heuristic methods: first, additional samples are determined using a sequential greedy algorithm and the objective function evaluated. Second, the points obtained using a greedy algorithm are used as a starting solution in simulated annealing. Through a swapping procedure, additional points are exchanged for other potential points, while the objective function is recomputed. This *metaheuristic* procedure combines the advantage of the greedy algorithm, that is its rapidity, with simulated annealing, which is recognized for its convergence towards optimal solutions.

2 Additional sampling methodology

A variable of interest Y has been measured at m locations within a study region, \mathfrak{D} . Measurements are denoted $y(\mathbf{s}_i)$, $\forall i = 1 \dots m$ (Goovaerts 1997). Using data values of the primary variable and a covariogram function, the kriging variance at a gridpoint \mathbf{s}_g :

$$\left(\sigma_k(\mathbf{s}_g) \right)^2 = \sigma^2 - \mathbf{c}^T(\mathbf{s}_g) \cdot \mathbf{C}^{-1} \cdot \mathbf{c}(\mathbf{s}_g), \quad (1)$$

where \mathbf{C}^{-1} is the inverse of the covariance matrix \mathbf{C} based on the covariogram function. The term \mathbf{c} is a column vector and \mathbf{c}^T its corresponding row vector. The Average Kriging Variance (*AKV*) is obtained by integrating Equation 1 over the area \mathfrak{D} . Computationally, discretizing \mathfrak{D} over a fine grid of points (set G):

$$AKV = \int_{\mathfrak{D}} \left(\sigma_k(\mathbf{s}_g) \right)^2 \approx \frac{1}{|G|} \sum_{g \in G} \left(\sigma_k(\mathbf{s}_g) \right)^2 \quad (2)$$

Our first objective $Z[S]$ is to select a set of n points to our existing set of m samples, which will maximize the change in kriging variance by as much as possible. This process can be thought as a simulation of what the change in kriging variance is expected to be, without having to collect additional points, assuming the covariogram structure would remain constant (Burgess, Webster and McBratney 1981 as well as Cressie 1993). Specifically:

$$\underbrace{\text{MAXIMIZE}_{\{\mathbf{s}_{m+1}, \dots, \mathbf{s}_{m+n}\}}}_{Z[S]} Z[S] = \frac{1}{|G|} \sum_{g \in G} \left(\sigma_k^{\text{old}}(\mathbf{s}_g) \right)^2 - \left(\sigma_k^{\text{new}}(\mathbf{s}_g) \right)^2, \quad (3)$$

where S denotes the sampling scheme. The set \mathbf{P} of p potential points is obtained by discretizing \mathfrak{D} , generating a total of $\binom{p}{n}$ possible sampling combinations.

The kriging variance is unfortunately misused as a measure of reliability of the kriging estimate, as noted by several authors (Deutsch and Journel 1992; Armstrong 1994).

It is merely a function of the sample pattern, sample density, the numbers of samples and their covariance structure. The kriging variance assumes that the errors are independent of each other, which means that the process is stationary, an assumption violated in practice. Figure 1 illustrates the limitation of the kriging variance (Armstrong 1994), the objective being to interpolate the value of the inner grid point, highlighted with a question mark. The interpolation is a function of the values at the four surrounding observations. In scenario *b*, three very similar values and an extreme one. The scenario in *a* however shows four data values in a very narrow range. Assuming a similar spatial structure in both cases and given that the configuration of the data points is the same, the kriging variances are identical, and so are the kriged estimates. Nevertheless, since there is much less variation among its neighbors, the left-hand side scenario is a much safer option than the right hand-one when it comes to estimating the value of the primary variable.

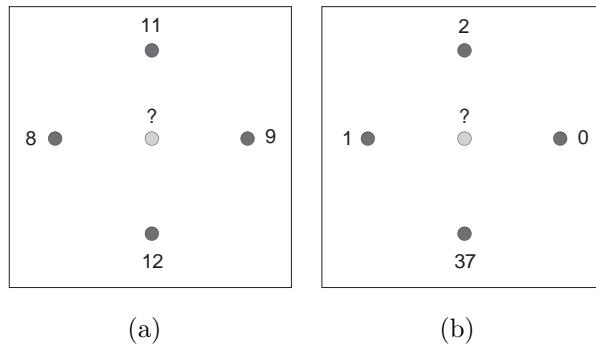


Figure 1: Example of two-dimensional nonstationarity. Dark points are used as data values to interpolate the center point (light gray). After Armstrong (1994).

This example illustrates the importance to account for local variations in the observed variable. Let $\hat{y}(\mathbf{s}_g)$ be the interpolated value of the primary variable Y at a grid node \mathbf{s}_g . Estimating by how much that grid node is different in value from its surrounding points \mathbf{s}_j ($j = 1, 2, \dots, J$) is possible through a filter process, specifically, a circular filter is constructed around each grid node \mathbf{s}_g that encompasses its neighbors. For illustration purposes, Figure 2 illustrates a 3 by 3 window, however the methodology can handle various neighborhood sizes. To determine an appropriate moving window size J , we compute the squared difference in interpolated value between the central grid node $\hat{y}(\mathbf{s}_g)$ and the surrounding ones $\hat{y}(\mathbf{s}_j)$. We also introduce a distance factor $d(\mathbf{s}_j, \mathbf{s}_g)$ and a parameter β , both regulating the importance given to nearby points. This is then summed over the set G . The weight $\lambda(\mathbf{s}_g)$ becomes:

$$\lambda(\mathbf{s}_g) = \sum_{j=1, j \neq g}^J \frac{d(\mathbf{s}_j, \mathbf{s}_g)^{-\beta} \cdot (\hat{y}(\mathbf{s}_j) - \hat{y}(\mathbf{s}_g))^2}{\sum_{j=1, j \neq g}^J d(\mathbf{s}_j, \mathbf{s}_g)^{-\beta}} \quad (4)$$

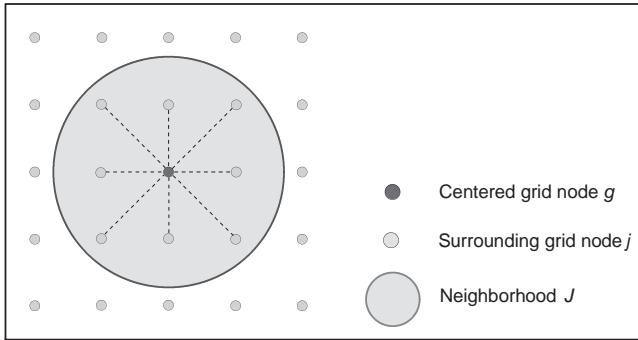


Figure 2: A 3×3 moving window: a circle is passed around a grid node within a specific distance.

If the neighborhood J is kept constant, $\lambda(\mathbf{s}_g)$ will exhibit great values when $\beta < 1$, because more weight is given to far away data points. As β increases, $\lambda(\mathbf{s}_g)$ decreases and flattens out for high values of β . If J is too large, zones of rapid changes may go undetected. Equation 3 should be account for spatial variation of the primary variable. As such, a weighted second-phase sampling problem can be formulated as a single-weighted objective (Cressie 1991) where the kriging variance is weighted by Equation 4:

$$\underbrace{\text{MAXIMIZE}_{\{\mathbf{s}_{m+1}, \dots, \mathbf{s}_{m+n}\}}}_{Z[S]} = \frac{1}{[G]} \sum_{g \in G} \lambda(\mathbf{s}_g) \cdot \left| \left(\sigma_k^{\text{old}}(\mathbf{s}_g) \right)^2 - \left(\sigma_k^{\text{new}}(\mathbf{s}_g) \right)^2 \right| \quad (5)$$

3 Application

In this paper, we use a sequential approach to strategically allocate new observations. To illustrate our methodology, we use primary data on soil concentration of Chromium (Cr) in a study area near La Chaux de Fonds, in the Swiss Jura (see, Goovaerts 1997 for the dataset). The Cr-concentration $\frac{mg}{kg}$ represents the quantity of the heavy metal per kilogram of soil sampled.

Sequential addition assumes that one additional point has to be added to the initial set n -times. Once the first point has been selected and added to the initial set M , $n - 1$ additional locations are to be chosen in a similar, sequential fashion. The sequential addition approach is illustrated using algorithms such as random strategy, total enumeration, greedy, simulated annealing and simulated annealing with greedy start. The greedy approach has the drawback of getting stuck at local optima, while total enumeration is not time-efficient. Since simulated annealing has the inherent property of jumping out of a local optimum, we capitalize on this technique for finding the optimal solution S^* to the sequential addition, using a cooling factor κ at the end of a fixed number of iterations T_{it} . Similarly, the step size for determining new neighbors (for swapping purposes) was reduced by a factor δ . A large initial step size

δ of 3 kilometers -corresponding to approximately half the size of the study area \mathfrak{D} - was chosen to permit wide jump swaps.

Figure 3 to the left illustrates the performance of the greedy algorithm and total enumeration in maximizing the change in weighted kriging variance with the addition of new samples, against the changes obtained using naive addition. The total enumeration evaluates all possible solutions to the sequential addition, but may still be suboptimal. To check on global optimality (14.879%), we ran a simultaneous simulated annealing (see Van Groenigen and Stein 1998) and found that sequential results were very close to the optimal. The sequential total enumeration yielded a 14.8% improvement in the objective function. For the naive (random) addition, a total of 1500 simulations were performed, providing a good lower bound to evaluate other heuristics. In the best-case scenario, a reduction of 7.52% was obtained, in comparison with a change of 4.89% in the worst case. Table 1 reports on the computational time. When simulated annealing is used (Figure 3 to the right), the algorithm returns near optimal solutions, even more so when the algorithm uses a lower cooling schedule (κ closer to 1), and a greater number of iterations per temperature steps T_{it} .

Sequential heuristic	Time (min)	Reduction (%)	Optimality gap (%)
Total enumeration	229.72	14.768	.75
Naïve	8.56	[2.869; 7.521]	[80.72; 49.45]
Average naïve	8.56	4.892	67.12
Greedy	8.04	12.537	15.74
SA-Greedy($\kappa = .875, \beta = .9$)	106.76	14.768	.75
SA-Greedy($\kappa = .35, \beta = .45$)	33.35	14.649	.8
SA($\kappa = .95, \delta = .965$)	241.06	14.733	.98
SA($\kappa = .35, \delta = .45$)	33.82	14.420	3.08
SA($\kappa = .05, \delta = .05$)	26.50	13.95	6.24
Simultaneous heuristic SA	1500	14.879	0

Table 1: Average reduction (%), and optimality gap (%) for the sequential and simultaneous addition after the addition of $n = 30$ points.

The combination of SA with a greedy start allows improvement upon a first very good solution. Since the starting solution is relatively good, SA may experience difficulties to improve upon that incumbent. Figure 4A shows the first 15 dynamic moves, with SA parameters $\kappa = .875, \beta = .9$, yielding the sequential optimal in 106.76 minutes. The location exhibiting the highest weighted kriging variance (point $\mathbf{a} = \mathbf{s}_{m+1}^+$) is selected and serves as a starting point for SA, yet the latter is unable to locate a better point, hence $\mathbf{a} = 1 = \mathbf{s}_{m+1}^+$. That point is added to the set \mathbf{M} and the weighted kriging variance is re-computed accordingly. Location $\mathbf{b} = \mathbf{s}_{m+2}^+$ is the point with the highest kriging variance and is selected as the starting point. SA finds a better sample at location 2-symbolized by a white dot, and that point

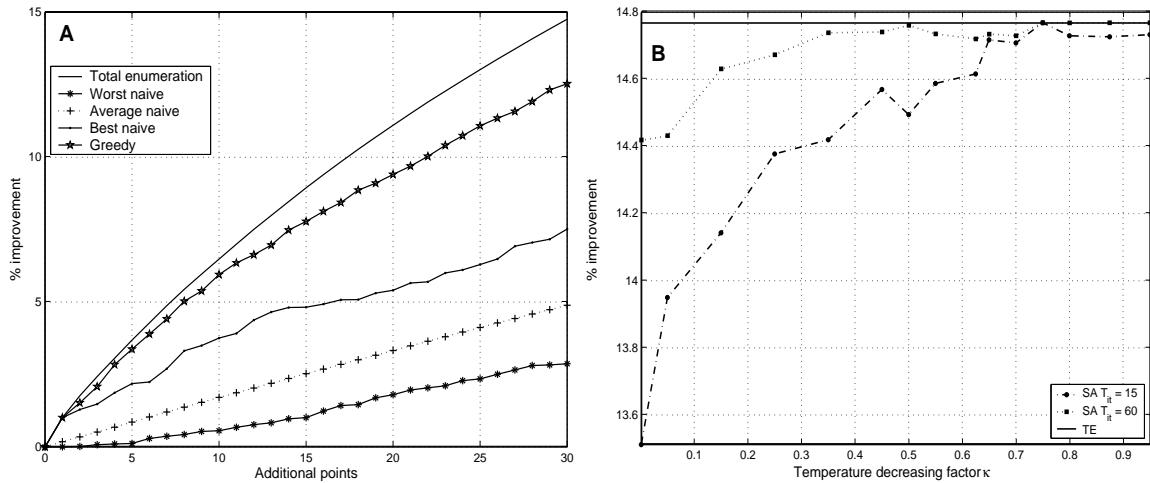


Figure 3: Percentage reduction in weighted kriging variance using a naïve approach versus total enumeration. The sensitivity of the sequential SA coupled with greedy to the cooling factor κ is illustrated in B. Notice for $T_{it} = 60$ how near-optimal solutions are obtained even if the temperature drops quickly ($\kappa = [1; .25]$).

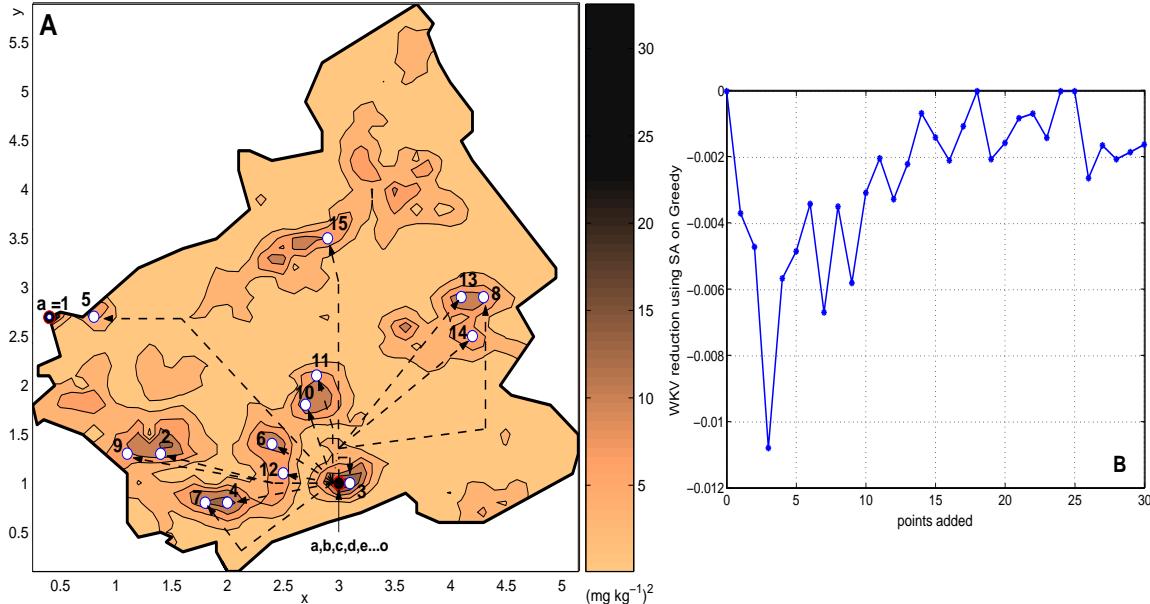


Figure 4: Illustration of the simulated annealing algorithm using a greedy start approach for the first 15 points (A), in the best-case scenario ($\kappa = .875$, $\beta = .9$). Black dots denote initial points obtained using greedy. The arrows point to the locations obtained using SA. Graph B illustrates the reduction between successive steps using SA on greedy for $n = 30$ points.

is added to \mathbf{M} . The weighted kriging variance is computed with the set \mathbf{M} that contains now two new samples, namely points **1** and **2**. In the following 17 additions, SA will ameliorate the incumbent greedy solution (see Figure 4B). Notice how often SA discovers a better solution from the initial greedy sample, yet the magnitude of that improvement decreases as new samples are being added.

4 Conclusions:

In this paper, we have addressed the second-phase spatial sampling problem based on two main criteria; the change in kriging variance, and the spatial variation of the primary variable. Results of our numerical testing showed that total enumeration outperformed all other heuristics in the sequential case, but at the cost of an extended running time. The greedy approach, which locates new samples points where the weighted kriging variance is the highest, returns near-optimal results in a short time-frame. Simulated annealing is very sensitive to the choice of the cooling factor, that governs the search procedure. The combination of simulated annealing with a greedy start performed remarkably well considering the optimality gap and the computational time.

References

- Armstrong M. (1994). Is research in mining geostats as dead as dodo? In: Dimitrakopoulos R. (Ed.) *Geostatistics for the Next Century*. Kluwer Academic Publisher. Dordrecht: 303-312.
- Burgess T.M., Webster R. and A.B. McBratney (1981). Optimal interpolation and isarithmic mapping of soil properties: IV. Sampling strategy. *Journal of Soil Science*, vol. **32**: 643-659.
- Cressie, N., 1991. Statistics for Spatial Data. Wiley, New York, USA, 900p.
- De Gruijter, J., Brus, D.J., Bierkens, M.F.P. and Knotters M., 2006. Sampling for Natural Resource Monitoring. Springer, 332p
- Delmelle E. and P. Goovaerts (2009). Second-phase sampling designs for non-stationary spatial variables. *Geoderma* 153: 205-216
- Deutsch C.V. and A.G. Journel (1997) *Gslib: Geostatistical Software Library and User's Guide*. Oxford University Press, 2nd edition, 369p.
- Goovaerts P., 1997. Geostatistics for natural resources evaluation. 483p.
- Muller, W., 1998. Collecting Spatial Data: Optimal Design of Experiments for Random Fields. Heidelberg: Physica-Verlag.
- Van Groenigen, J.W. and Stein, A., 1998. Constrained optimization of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality*, vol. 27: 1078-1086.

Spatial Determinants of Quality of Life in Urban Areas: Does Metropolitan contiguity effect?

Ali Goli

Assistant Professor of Regional Planning,
 Social Science faculty
 Shiraz University, Shiraz, Iran
 Telephone: +98711-6134407
 Fax: +98711-6289661
 Email: goli_ali@yahoo.com
goli@shirazu.ac.ir

Abstract

Studies related to the quality of life (QoL) and investigating it in human societies have been of great importance in recent days as an axis for government efficiency. Such studies have always been tried to be increased in human residences by investigating factors affecting the improvement of quality of life both in objective (availability of material facilities), and subjective aspects (satisfaction).

This research has used spatial analysis based on the results of a 20-percent count of statistics by the Iranian Statics Center in 2006 to investigate objective aspects of the objective aspects of QoL and metropolitan affect in urban areas that have contiguity with metropolitan in Iran. Furthermore, there was a great deal of spatial variation in QoL and deprivation index which is not explained by the global regression framework. Geographically Weighted Regression (GWR) analysis was undertaken using an adaptively defined kernel with a bi-square function. The kernel bandwidth was determined by minimisation of the Akaike Information Criterion (AIC) value.

Indicators being used in this research include availability of facilities to families and their home quality. Results show a direct relationship between urban population size and city size rank and contiguity with metropolitan on the availability of facilities to families and their home quality.

The GWR outputs showed that some areas with high QoL were also areas that have contiguity with metropolitan and high city growth rate. Therefore, the GWR results highlighted 'hot spot' areas .On the other hand; cities around metropolitan have better home quality (resistant settlements) and more available facilities to families than other cities.

Keywords: Quality of Life, objective quality, subjective quality, Geographically Weighted Regression (GWR)

Fractal perspectives of GIScience based on the leaf shape analysis

P. Tuček^{1,2}, L. Marek¹, V. Paszto¹, Z. Janoška¹ and M. Dančák³

¹Department of Geoinformatics, Faculty of Science, Palacký University in Olomouc, Tř. Svobody 26, 771 46 Olomouc, Czech Republic
 Telephone: +420 585 634 513
 Fax: +420 585 225 737
 Email: vit.vozenilek@upol.cz

²Regional Centre of Advanced Technologies and Materials, Palacky University in Olomouc, Department of Mathematical Analysis and Applications of Mathematics, 17. Listopadu 12, 771 46, Olomouc, Czech Republic
 Telephone: +420 585 634 521
 Fax: +420 585 225 737
 Email: pavel.tucek@upol.cz

³Department of Botany, Faculty of Science, Palacký University in Olomouc, Šlechtitelů 11, 783 71 Olomouc, Czech Republic
 Telephone: +420 585 634 805
 Fax: +420 585 634 824
 Email: martin.dancak@upol.cz

1. Introduction

Since Mandelbrot (1967) published its basics, fractal geometry and fractal dimension (non-integer dimension) is well known as a valuable tool for describing the shape of objects. It gained large popularity in many fields of natural sciences (Batty and Longley 1994, Goodchild 1980, Hastings and Sugihara 1994, Kitchin and Thrift 2009, Peitgen et al. 1992), including e.g. ecology, geography, GIScience, where the measures of object's shape are essential.

One of the major principles in fractal geometry is self-similarity and self-affinity. The most theoretical fractal objects, such as Barnsley's fern, are self-similar (any part of the object is exactly similar to the whole) and self-affine (transformed self-similar objects). And typical fractal objects like leaves are very suitable to test our methods for possible use on geodata. And for geospatial fractal-based analysis, the various drainage systems were acquired and examined.

For leaves, we show that discrimination based on only two fractal features has more than 90% accuracy. This notion is important, because it proves that automated classification can be based also on complexity of shapes and not only on their qualitative measurements. Fischer discriminant analysis is used to distinguish between families and species with satisfactory results. Leaf skeleton is very similar especially to river and road network and thus we examine different types of river drainage network and show that their complexity differs significantly.

2. Methods

There exist a number of methods for estimating fractal dimension and as e.g. Reynoso (2005) shows, results obtained by different methods often differ significantly. Also not only the method itself, but the software, which calculates the fractal dimension, may contribute to the differences (Reynoso 2005). All the calculations were accomplished in free software Fractalyse, easily downloadable from www.fractalyse.org.

2.1 Box-counting method

The box-counting method was used for modified data – binary pictures. Box-counting dimension of a subset X of the plain is defined by counting number of unit boxes which intersects X : for any $\Delta s > 0$, let $N(\Delta s)$ denote the minimum number of n -dimensional cubes of linear scale Δs (side length) needed to cover X . Then X has box dimension D if $N(\Delta s)$ satisfies (according to Hastings and Sugihara 1994, Theiler 1990):

$$N(\Delta s) \approx c(1/\Delta s)^D, \quad (1)$$

where $\Delta s \rightarrow 0$, c is a constant and box-counting dimension of X is D . Formula (1) is called power law. Dimension D is then be computed by:

$$D = \lim_{\Delta s \rightarrow 0} [-\log N(\Delta s) / \log \Delta s], \quad (2)$$

According to formula (2), calculation of box-counting dimension is simple. For a sequence of cell size $\Delta s > 0$, the number of cells $N(\Delta s)$ needed to cover the set S is calculated.

2.2 Linear Discriminant Analysis (LDA)

Discriminant analysis is used in situations where the clusters are known a priori. The aim of discriminant analysis is to classify an observation, or several observations, into these known groups (Härdle and Simar, 2007). The classification rule is often a linear function of measurements that maximizes the separation between groups relative to their within-group variability (Johnson and Wichern 2007). Discriminant scores are results of the LDA.

3. Data processing

Unique dataset of leaves was available thanks to Department of Botany of Faculty of Science in Olomouc. At first, possibilities of automated data classification were tested on dataset of leaves. A unique dataset containing 133 samples of leaves from 7 different species belonging to 3 families was available for scanning into raster digital format (fig. 1 left). Examined plants can be divided to two groups (tab. 1).

First group (Angiosperms, Eudicots)	Second group (Pteridophytes, Polypodiopsida)	
Roseaceae family:	<i>Dryopteridaceae</i> family:	<i>Blechnaceae</i> family:
<i>Alchemilla vulgaris</i> (simple lobed leaves), <i>Rubus wimmerianus</i> (compound 5-foliate leaves), <i>Fragaria moschata</i> (compound trifoliolate leaves)	<i>Polystichum aculeatum</i> (compound 2-pinnate leaves), <i>Dryopteris filix-mas</i> (compound 1-2 pinnate leaves), <i>Dryopteris carthusiana</i> (compound 2-3-pinnate leaves)	<i>Blechnum spicant</i> (simple pinnatifid sterile leaves and compound 1-pinnate fertile leaves)

Table 1. Two main groups of examined plants (with particular species).

Scanned leaves were transformed into two datasets. First examined data-set was acquired by transforming raster pictures into binary raster picture in order to perform fractal analysis of leaf area (fig. 1 middle). Then the leaf skeleton was digitalized and extracted from the raster pictures for further fractal dimension analysis (fig. 1 right).



Figure 1. Fern leaves and particular steps in their processing.

Fractal dimensions of both, the leaf area and skeleton, were calculated. These computed fractal properties together with an affiliation to family (or species) served as basis attributes for linear discriminant analysis. R project was used as a computational environment where LDA was applied, visualized and compared.

The LDA was applied using the relation $group \sim area.FD + skeleton.FD$ as the basic formula. Two-dimensional space in which single points belonged to one of three families (or 7 species) was the result of the analysis.

Predicted affiliations to single groups (based on both above mentioned methods) was visualized and compared with the real belonging to the group (fig. 2).

River drainage network was obtained from free Internet source. Ten examples from each selected drainage system were selected and examined. Data are available from DIVA-GIS website (<http://www.diva-gis.org/Data>) and were used to select appropriate areas with typical drainage systems. Drainage system is governed by many factors, most importantly by topography, geology and preceding (and contemporary) geomorphologic processes (Zernitz 1932). There are several types of drainage system, each of them typical for a certain type of relief (Knighton 1998).

Type of drainage system	Description	Example
Dendritic	Most common, river follows the slope, in V-shaped valleys	Mississippi
Parallel	Steep, uniformly sloping relief	Angola (Moxico, Cuando Cubango)
Trellis	Folded mountains with strike valleys	Appalachian Mountains
Rectangular	Rocks with uniform resistance to erosion, but with two directions of jointing at approximately right	Western Iran

	angles	
Radial	Typical for volcanoes, craters, radial depressions	Kauai Island
Deranged	No coherent pattern of rivers and lakes,	Canadian Shield

Table 2. Examined river drainage systems with descriptions and examples (according to Knighton 1998, Lambert 1998, Ritter 2006).

4. Results

In the case of LDA, the result of analysis is two-dimensional space, where points belong to the predicted groups (fig. 2). The success of the classification to the family was 93.2 % and 64.7 % in the classification to the species. Overview of classification into the families is in the table 3.

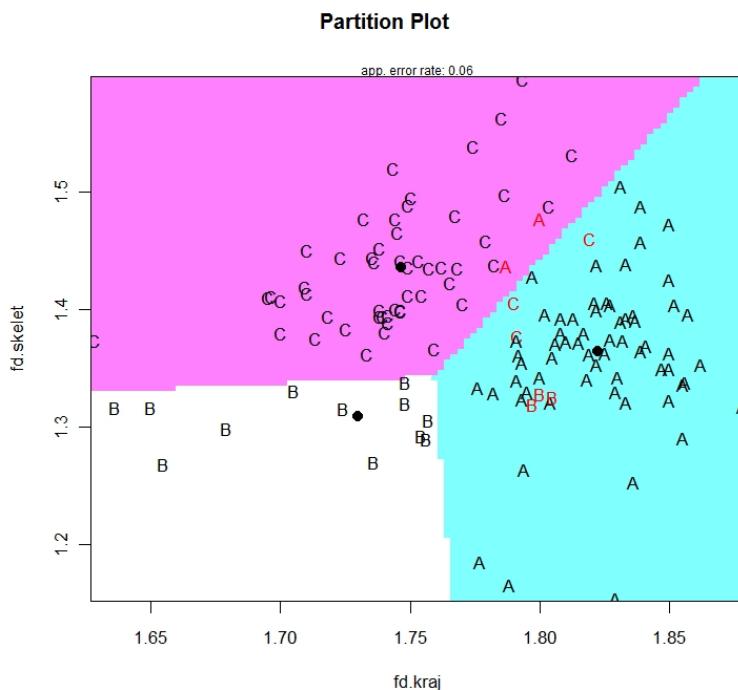


Figure 2. Classification to families using LDA.

	A	B	C	Sum
	LDA	LDA	LDA	
A	62	0	2	64
B	3	11	1	15
C	3	0	51	54

Table 3. Comparison of classification to particular families.

Cluster analysis showed that species from the same family are more similar than species from different families (fig. 3) and shows unsupervised classification of species, which corresponds with actual taxonomical classification.

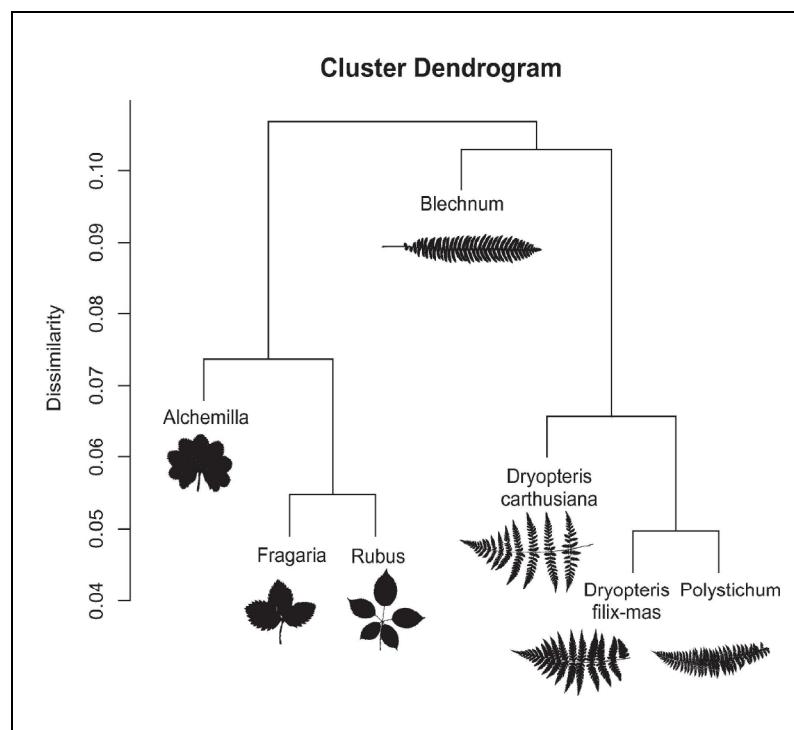


Figure 3. Dissimilarity of different species based on the cluster analysis.

Then, for each selected area of different river drainage system, fractal dimension of river network was computed. Results are summarized in figure 4.

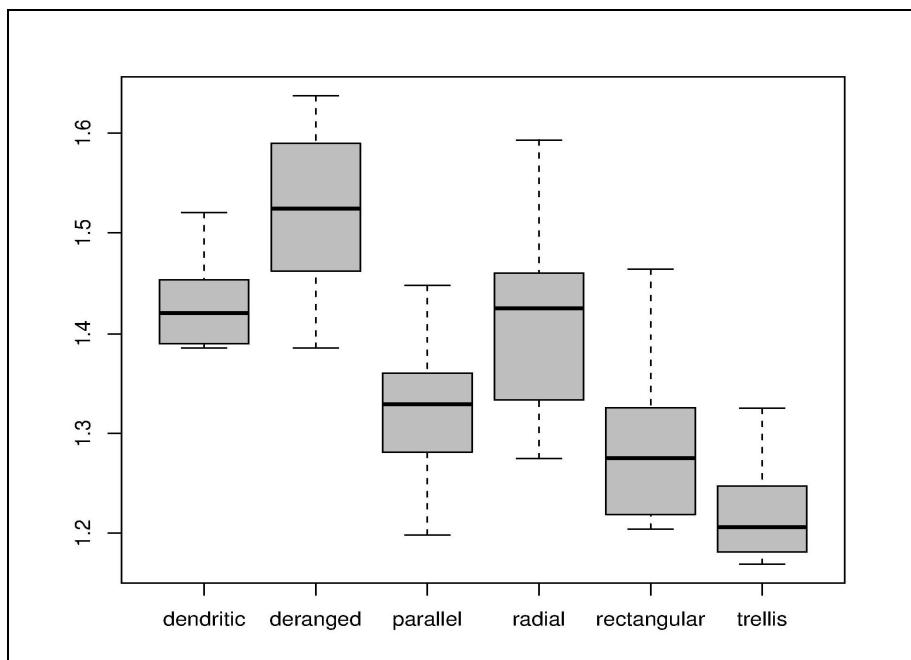


Figure 4. Graphical overview of fractal dimension values of different drainage system types.

Analysis of variance (ANOVA) was used to verify that mean fractal dimension for each drainage system differs. Bartlett's test, however, did not refuse the equality of variances. Since the only one characteristic of the drainage system was measured, data were not suitable for classification purposes. Despite of this fact, cluster analysis was conducted based on mean and standard deviation of values for each drainage network in order to examine the similarity of drainage systems.

The most alike are parallel and rectangular drainage systems, which both embody quite similar patterns. Both of them (and also trellis drainage system) show regular patterns and their fractal dimension is the lowest. Radial and dendritic drainage systems embody more irregular shapes and therefore their fractal dimension is higher. The most irregular patterns are observed in deranged drainage system. Meanders, lakes and flood plane lobes are typical for deranged drainage systems and it is difficult to observe any distinguished structure in the shape of river network. The complexity of this type of drainage system is the greatest and analysis shows that the fractal dimension is also the highest. Cluster analysis shows that this type of drainage network is also the most dissimilar from any other drainage system types.

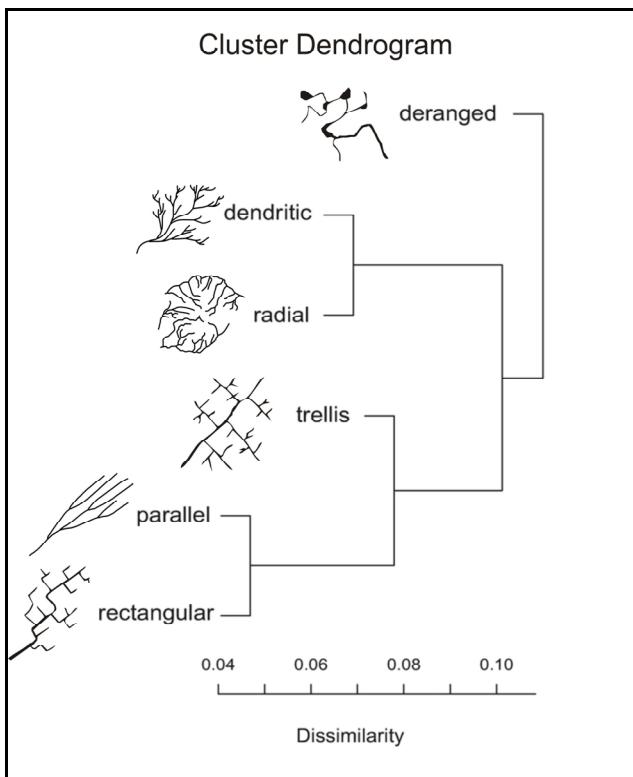


Figure 5. Dissimilarity of different drainage systems based on their fractal dimension.

5. Conclusion

We showed that just only two characteristics based on fractal dimension measurement (without any additional geometric features) are sufficient for accurate object classification. Although the analyses were applied to leaves, it is possible to use the methods for various geographical analyses based on geodatasets.

Although leaves characteristics were used in this study, the perspectives of using analogical methods in GIScience are bidding themselves.

We examined different types of drainage systems by means of their fractal dimension. Results show that mean fractal dimension for the six most common drainage systems differ significantly, and that the more regular drainage system, the lower fractal dimension (and vice versa). Complexity of fluvial network can point to different origin of the rivers and therefore this could be useful for geologist to semi-automatically evaluate the drainage systems. Also river complexity can be of vital importance in precipitation-runoff evaluation of the water basin and could be correlated to different processes, e.g. during flood event.

Further research concerning fractal measurements upon geospatial data is nowadays intensively conducted on Department of Geoinformatics, Palacky University in Olomouc, Czech Republic.

6. Acknowledgements

This work has been supported by the Operational Program Education for competitiveness – European Social Fund (CZ.1.07/2.2.00/15.0276) and the by the Ministry of Education, Youth and Sports of the Czech Republic.

7. References

- Batty M. and Longley P., 1994, Fractal Cities: A Geometry of Form and Function, Academic Press Ltd., London, San Diego, 1994, 394 s.
- Goodchild M.F., 1980, Fractals and the accuracy of geographical measures. *Math. Geol.*, Vol 12, pp 85–98.
- Härdle W., Simar L., 2007, Applied Multivariate Statistical Analysis. Springer Berlin, 2nd edition.
- Hastings H. M., Sugihara, G., 1994, Fractals: A User's Guide for the Natural Sciences. Oxford : Oxford University Press. 235 s.
- Hothorn T., Hornik K., Zeileis A. , 2006, Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3), 651–674.
- Johnson R., Wichern D., 1982, Applied Multivariate Statistical Analysis, Prentice-Hall, Englewood Cliffs, NJ.
- Kitchin R., Thrift N., 2009, International Encyclopedia of Human Geography. United Kingdom : Elsevier Science. 8250 s. (hardcover).
- Knighton, D., 1998, Fluvial forms and processes: a new perspective. Oxford University Press, Inc., New York.
- Lambert D., 1998, The Field Guide to Geology. Checkmark Books. pp. 130–131.
- Mandelbrot B. B., 1967, How long is the coast of Britain? Statistical self-similarity and fractional dimension, *Science* 155, str. 636-638.
- Peitgen H.-O., Jürgens H., Saupe D., 1992, Chaos and Fractals : New Frontiers of Science. New York : Springer. 984 s.
- R Development Core Team, 2010, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Reynoso C., 2005, The impact of chaos and complexity theories on spatial analysis - problems and perspectives. 24th Research Symposium: Reading Historical Spatial Information from around the World: Studies of Culture and Civilization Based on GIS Data, Kyoto Japan, 7-11 February.
, 63(2):1037-1068.
- Ritter M. E., 2006 The Physical Environment: an Introduction to Physical Geography. Date visited.
http://www.uwsp.edu/geo/faculty/ritter/geog101/textbook/title_page.html
- Zernitz, E. R., 1932, Drainage patterns and their significance, *J. Geol.*, 40, 498-521.

Building a Web-based Cancer Atlas for Saudi Arabia

K. Al-Ahmadi¹, A.J. Heppenstall², L. See³ and A. Al-Zahrani⁴

¹Space Research Institute, King Abdulaziz City for Science and Technology, P.O. Box 231353, Riyadh 11321, Saudi Arabia
Telephone: (+966 1) 4814542
Fax: (+966 1) 4813845
Email: alahmadi@kacst.edu.sa

²School of Geography, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT, UK
Telephone: (+44 113) 3433392
Fax: (+44 113) 3433308
Email: aj.heppenstall@leeds.ac.uk

³International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, A-2361 Laxenburg, Austria
Telephone: (+43 2236) 807423
Fax: (+43 2236) 807599
Email: see@iiasa.ac.at

⁴King Faisal Specialist Hospital and Research Centre, P.O.Box 3354, Riyadh 11211, Saudi Arabia
Telephone: (+966 1) 4647272
Fax: (+966 1) 4414839
E-mail: alisaz@kfshrc.edu.sa

1. Introduction

Cancer is the second most frequent cause of death in developed countries and the fourth in the Eastern Mediterranean Region (EMR). The estimated number of new incidents of cancer each year is expected to rise from 11 million in 2002 to 16 million by 2020, with more than half of these occurring in developing countries. In the EMR, cancer is forecast to rise by a factor of 1.8 times over the next 10 years (WHO, 2005). Saudi Arabia is located in the EMR where more than 45,500 cancer incidences have been registered between 1998 and 2004 (SCR, 2010). These incidences and a range of socio-economic variables have been compiled into a rich spatial-temporal database. GIS and spatial analysis provide opportunities for epidemiologists and cancer researchers to investigate spatial patterns within this dataset and to understand relationships between cancer and other health, socioeconomic and environmental variables (Brewer, 2006). To date, GIS has not been used extensively in Saudi Arabia for this purpose. Interpretation, assimilation and analysis of cancer incidence maps are valuable for identifying low, average and high concentrations of cancer incidence. This can be a preliminary step for research into the causes or aetiology of particular types of

cancers and for setting priorities for public health awareness campaigns, educational activities, improving methods of early detection, screening and cancer prevention and control.

The aim of this paper is to describe a new interactive web-based tool aimed at both researchers and the public for analysing cancer data in Saudi Arabia. This statistical and spatial cancer atlas (SSCA) was designed and implemented using a client-server architecture. The atlas uses data from the Saudi Cancer Registry (SCR) at four spatial levels: national, regional, sub-regional and cities. The SSCA contains maps of the spatial distribution of cancer incidence over time and trends in the incidence of different types of cancers at the four spatial scales. It was designed for planning and resource allocation of health care resources and facilities and to highlight areas for further epidemiologic investigations into the causes of cancer. The architecture of the system and the main statistical and spatial features are described in the next section. This is followed by an example of the types of results that can be generated using the atlas. A brief discussion of further research areas is then provided.

2. Design of the SSCA

2.1 Architecture

There are four main components that make up the architecture of the SSCA (fig. 1):

- **The Client:** Flex was used to build the web client for the atlas, where MXML and ActionScript have been used to define the layout, appearance and behaviour of the application. These were then compiled into a single SWF file that makes up the Flex client SSCA application. The ArcGIS Server API for Flex was also used (ESRI, 2010), which allows maps and analysis from ArcGIS Server to be displayed in the client.
- **The Web and Application Server:** responds to client requests, which can involve linking to other application servers, e.g. the database or map server. For the SSCA, ASP.NET was used to build the web application server since it is a powerful tool for creating dynamic and interactive web applications. The cancer database is maintained in a Microsoft SQL Server and the operating system is Windows Server, so the application server interfaces with these other components seamlessly.

- **The Map Server:** fulfills spatial queries, conducts spatial analysis, and generates and delivers maps to the client based upon the users' requests. The output from the map server can be a simple map image in a graphic format or map elements served by ArcGIS Server.
- **The Database Server:** houses the cancer data in a relational database structure stored in Windows SQL Server 2008 (enterprise edition). The Flex client application accesses the database through SQL. For each individual cancer case, 14 variables are recorded including: gender, age, birth date, marital status, region, city, diagnosis date, site, topography, morphology, behaviour and stage of diagnosis.

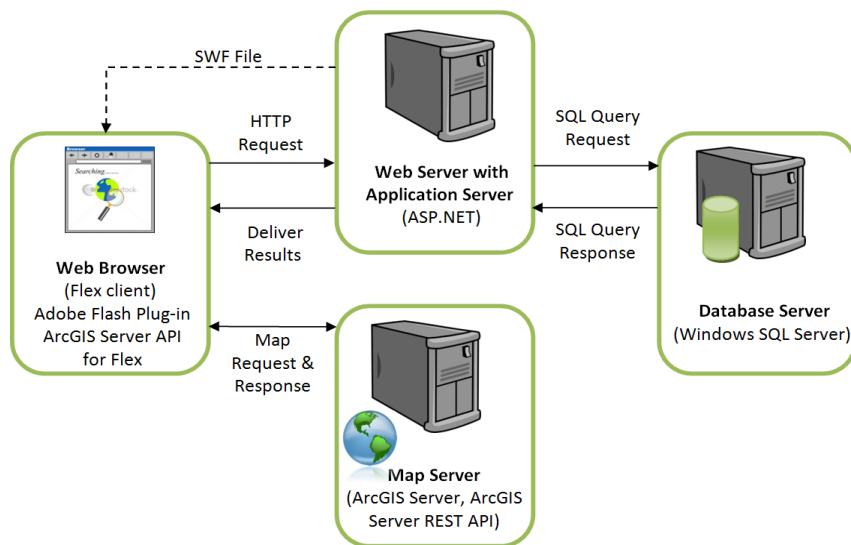


Figure 1. Architecture of the cancer atlas.

A more detailed explanation of the architecture and the database structure is provided in Al-Ahmadi (2010).

2.2 Statistical and Spatial Features of the Atlas

The statistical and spatial functionality is embedded within two different Graphical User Interfaces (GUI). The *statistical analysis interface* is divided into different dynamically-linked panels as shown in fig. 2. The function of the Analysis panel is to allow users to select the spatial level, whether analysis is to be undertaken on all types of cancers or only the ten most common types, and the type of analysis. The type of analysis is standard or advanced where standard consists of 35 pre-defined analyses

that allow users to explore the distribution of cancer cases according to different criteria: age group, gender, stage distribution, morphology, time period, etc., as well as animations of population and cancer incidence pyramids over time. Advanced analysis allows the user the option to adjust the parameters of the 35 pre-defined analyses. More details of the types of analyses available are presented in Al-Ahmadi (2010). The *spatial analysis interface* as shown in fig. 3 is similar to the statistical analysis interface except that the results are displayed in map form. However, users can also display figures and tables through the legend panel. Different maps can be generated such as graduated choropleth maps, density maps, symbol maps, pie chart maps and bar chart maps dynamically.

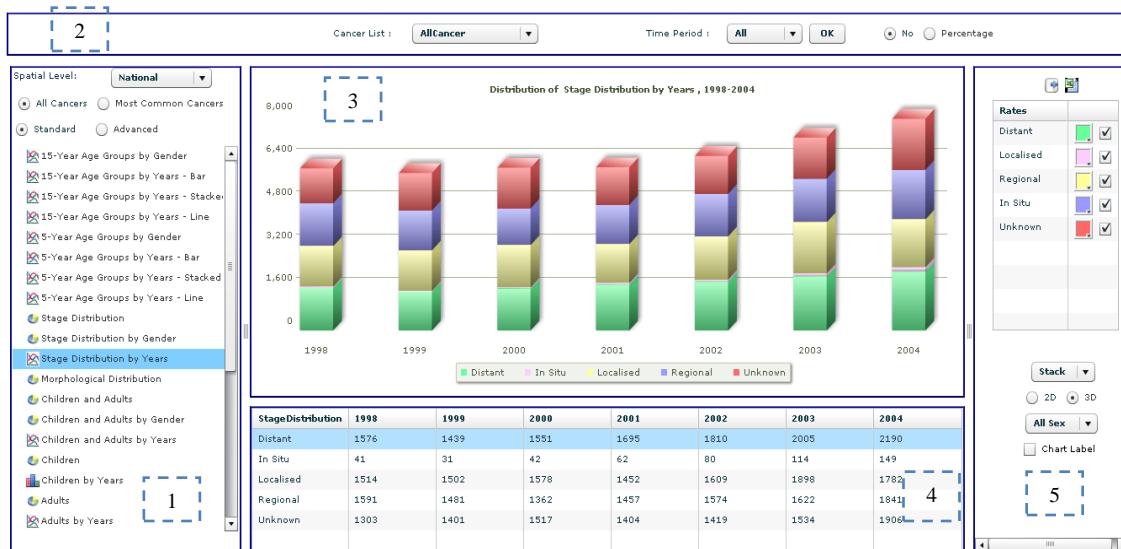


Figure 2: Statistical analysis interface panels for displaying the (1) analysis; (2) cancer type & time period; (3) figures; (4) tables; (5) legend.

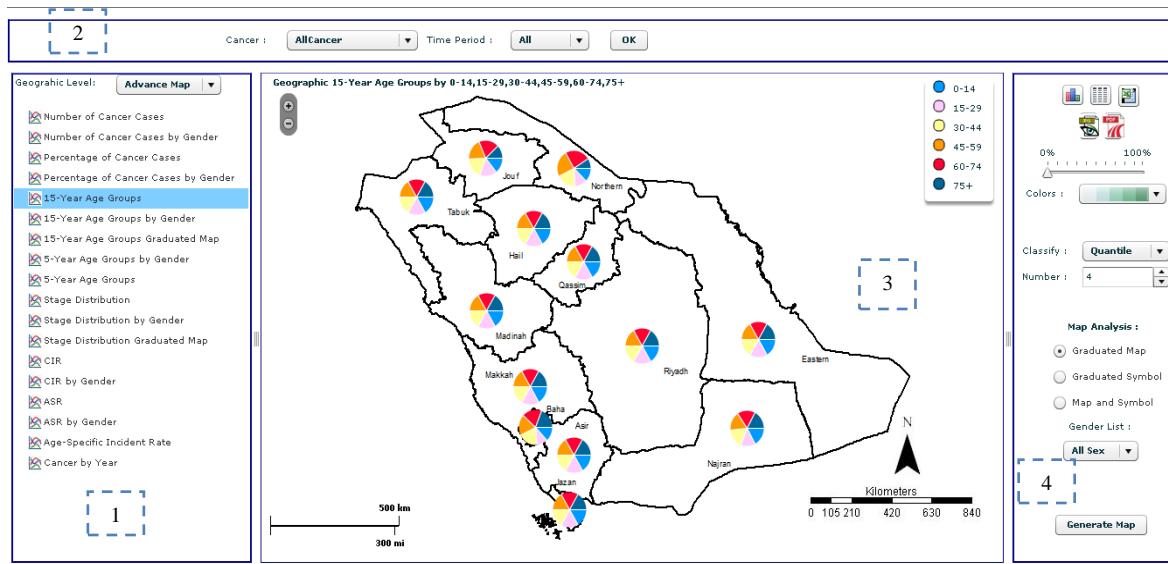


Figure 3: Spatial analysis interface panels for displaying the (1) analysis; (2) cancer type & time period; (3) maps; (4) legend.

3. Further Developments

A web-based interactive Statistical Spatial Cancer Atlas has been developed for Saudi Arabia. The atlas will be used to determine whether observed geographic variation in the cancer incidence rates for the most common cancers such as breast, liver, thyroid, and colorectal cancers are random or statistically significant. Where there are statistically significant clusters, research questions of interest are whether these are temporary or time-persistent, whether they are specific to geographic areas, whether they are consistent across all diagnostic stages and whether they can be attributed to covariates such as age, sex, and urban/rural status. The atlas will be made available to researchers in the spring of 2011 and to the public in late 2011. Based on feedback from the researchers, the application will be improved and new advanced features will be added, e.g. space-time clustering, autocorrelation statistics, logistic regression, etc.

A full demonstration of the Atlas and its capabilities will be given at the conference.

4. Acknowledgements

This research was funded by Grant 29-286 from King Abdulaziz City for Science and Technology (KACST). Cancer data was acquired from Saudi Cancer Registry (SCR). The views stated in this publication are those of the authors and do not necessarily represent the official views of KACST and SCR.

5. References

- Al-Ahmadi, K. (2010). Statistical and Spatial Cancer Atlas Web-Based Application. Riyadh, Saudi Arabia.
- Brewer, C. (2006). Basic Mapping Principles for Visualizing Cancer Data Using Geographic Information Systems (GIS), American Journal of Preventive Medicine, 30, (2S).
- SCR (2010). Saudi Cancer Registry, Saudi Arabia.
- WHO (2005). World Health Organization [<http://www.who.int/en/>]

Modelling the Humanitarian Relief through Crowdsourcing, Volunteered Geographical Information and Agent-based modelling: A test Case - Haiti

A. T. Crooks¹, S. Wise²

¹George Mason University, Room 379, Research 1 Building, MS 6B2, Fairfax, VA 22030, USA
 Telephone: (001) 703 993 4640
 Fax: (011) 703 993 9290
 Email: acrooks2@gmu.edu

² George Mason University Center for Social Complexity, Research 1 Building, MS 6B2, Fairfax, VA 22030, USA
 Telephone: (001) 703 993 1402
 Fax: (011) 703 993 9290
 Email: Sarah Wise swise5@gmu.edu

1. Introduction

Natural disasters such as earthquakes and tsunamis occur all around the world but the exact timing of such events are difficult to predict. A commonality of all natural disasters is that they alter the physical landscape and can cause severe disruption to peoples daily lives. To aid humanitarian efforts in such instance one needs spatial data but, more often than not, in less developed counties spatial data is lacking. Even in cases where spatial data is available, it often lags behind what has changed on the ground. Over recent years there has been a growth of *bottom-up* campaigns to crowdsource (Howe 2006) spatial data, using volunteers to map entire counties which some term volunteered geographic information (VGI, Goodchild, 2007). Recently attention has focused on using the crowd to help map the infrastructure and devastation caused by natural disasters, such as in Haiti and Pakistan (e.g. Biewald and Janah 2010).

While the use of GIS for emergency management is not new (see Cova 2005) applications often focus evacuation (e.g. Cova and Johnson 2003). Agent-based modellers have also attempted “agentize” such models (e.g. Thorp et al. 2006) thus adding more realistic behaviours but essentially such models are just evacuation models. There are few agent-based models that explore humanitarian assistance and those that do tend not to be overtly spatial (e.g. Salgado et al. 2010). There is a great potential for the use of agent-based modelling (ABM) and GIS to assist first responders and logistic support to understanding the complexities of people affected by such natural disasters (Fiedrich and Burghardt 2007).

This paper explores a prototype spatially explicit agent-based model where people search for food after an earthquake. The model is created from crowdsourced geographic information, coupled with other sources of publically available data, and explores how aid might be distributed to relieve the suffering of the people affected. We focus on the devastating magnitude 7.0 earthquake that struck Haiti on the 12th of January 2010 which is estimated to have killed 230,000 people and left more than 1.6 million people homeless (BBC 2010). Fig. 1, provides an idea of the population distribution of Haiti, with the greatest density in and around Port-au-Prince.

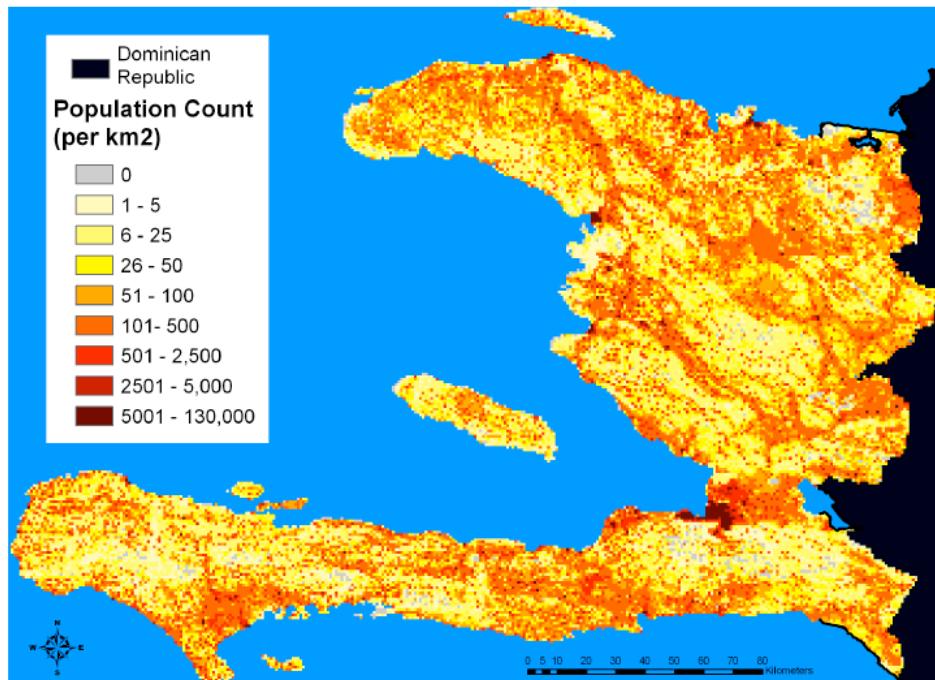


Figure 1. The Republic of Haiti and its population distribution in 2009.

2. Methodology

To demonstrate how such data can be utilized, we have created a basic agent-based model programmed in Java utilizing the MASON simulation toolkit (Luke et al. 2005) and its GIS extension, GeoMASON (Sullivan et al. 2010). One of the novelties of this model is that it combines both raster and vector data structures into a single simulation. The simulation area measures 8km by 6km around Haiti's capital, Port-au-Prince, as shown in Fig. 2.



Figure 2. Data on the devastation focused on Port-au-Prince. A: original data, B: geo-referenced image with roads shown which were used to locate the map.

Raster data comes from several sources, Fig. 3, summarises the data used in the simulation. To initialize the agent population, we use population counts from the 2009 LandScan (2011) dataset. The agents need are based on information about the devastation from G-Mosaic (2010). This data assesses damage at a number of different levels from

totally destroyed to intact structures. The assumption the model makes is that agents in the areas of greatest devastation have the greatest needs. The data was edited and geo-referenced using vector road lines sourced from Geocommons (2010). The road layer is also used for defining paths via an A* algorithm from the agents homes to the aid points. The spatial resolution of the model is set at 100m² however, multiple agents can be in one cell and agents can move a maximum of 100m per iteration (tick) of the model.

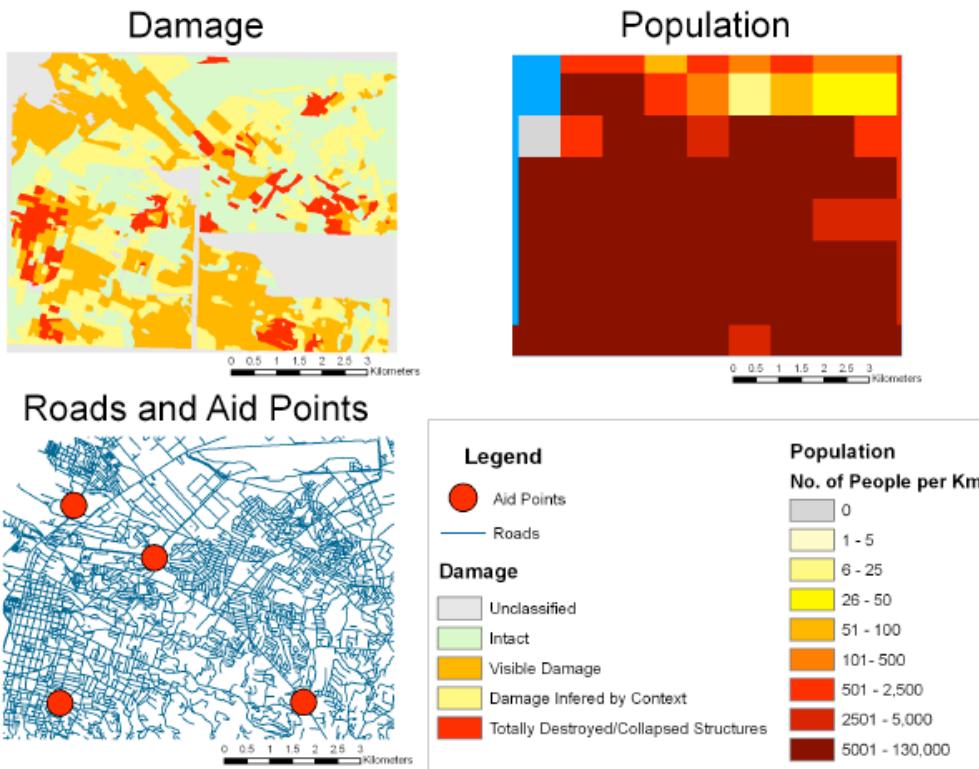


Figure 3. Model inputs.

2.1 Agent Decision Making Process

Data alone tells us little to how the people in such areas will react to the devastation or the supply of food. For this, we turn to ABM. The people (agents) within the simulation have a goal to maximize their energy, in the sense that no agent wants to starve as shown in Fig. 4. At model initialization agents around the food distribution points know of its location, agents then inform other agents about the distribution point via a diffusion mechanism. Over time more and more agents become aware of their nearest, but also other distribution points, as information is spread throughout the system. Agents then evaluate if it is worthwhile for them to go and get food. They do this by planning the shortest path to the food via the road network. Within the simulation agents have a certain amount of energy depending on the level of destruction of where they are initialised, when their energy level reaches 0 they die.

3. Simulation Results

Fig. 5, shows some simulation results. Initially, at $T=1$, few agents know about the food distribution points. Over time, such as at $T=200$, more agents become aware of the

distribution points, as agents share information about the location of distribution points either by passing on the information to their neighbours (in the sense of a rumour model) or to agents they pass while moving towards the food source.

Currently we are in the stages of calibrating and validating the agents behaviours, along with exploring the optimal placement of distribution points, which we will report at the conference.

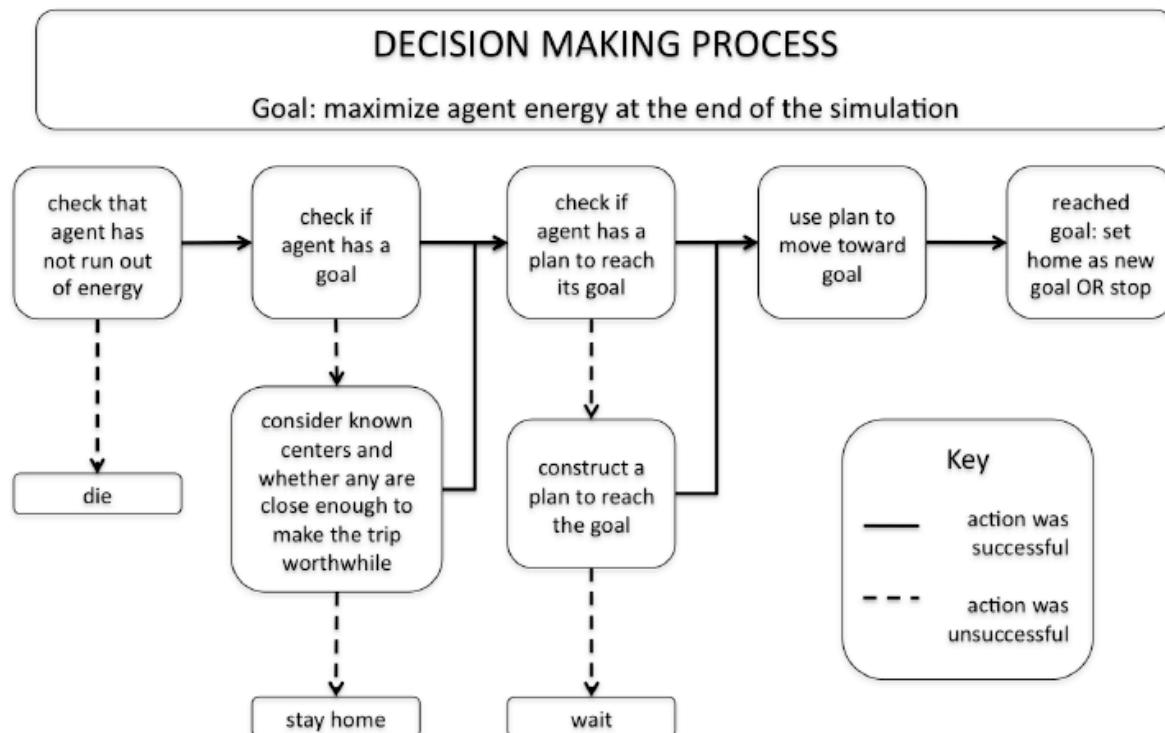


Figure 4. Agents decision making process.



Figure 5. The spread of information and movement of agents over time.

4. Summary

This paper attempts to demonstrate how GIS and ABM can be utilized to explore humanitarian relief after an earthquake. Such a model harnesses crowdsourced and other publicly accessible data. The model moves away from the more traditional disaster models that explore evacuation scenarios associated with catastrophic events, to map the consequences of such an event on the native population. It demonstrates how data can be used to initialize agents, their needs and their environments and how through a simple decision making process, people learn about and search for food. We consider this an

important aspect for humanitarian relief as natural disasters are times of great uncertainty, and it is difficult to predict beforehand how people will react to such events. By using agent-based models we can explicitly explore potential agent behaviour. Such a model, once thoroughly developed could act as a decision support tool for humanitarian relief.

5. Acknowledgements

The authors would like to acknowledge the Department of Computational Social Science at George Mason for providing support for this research.

6. References

- BBC, 2010, *Haiti quake death toll rises to 230,000*. Available at <http://news.bbc.co.uk/2/hi/8507531.stm> [Accessed on Jan, 26th, 2011].
- Biewald L, and Janah L, 2010, *TechCrunch: Crowdsourcing disaster relief*. Available at <http://techcrunch.com/2010/08/21/crowdsourcing-disaster-relief/> [Accessed on Jan, 26th, 2011].
- Cova T, 2005, GIS in emergency management. In Longley PA, Maguire DJ, Goodchild MF and Rhind D, (eds.), *Geographical information systems: Principles, techniques, applications, and management (Abridged Edition)*, John Wiley & Sons, New York, NY, 845-858.
- Cova T and Johnson JP, 2003, A network flow model for lane-based evacuation routing. *Transportation Research Part A*, 37(7): 579–604.
- Fiedrich F and Burghardt P, 2007, Agent-based systems for disaster management. *Communications of the ACM*, 50(3): 41-42.
- G-Mosaic, 2010, *Haiti data*. Available at <http://www.gmes-gmosaic.eu/haiti.html> [Accessed on Jan, 26th, 2011].
- Geocommons, 2010, *Haiti road data*, Available at <http://finder.geocommons.com/overlays/20302> [Accessed on Jan, 26th, 2011].
- Goodchild MF, 2007, Citizens as voluntary sensors: Spatial data infrastructure in the world of web 2.0. *International Journal of Spatial Data Infrastructures Research*, 2(1): 24-32.
- Howe J, 2006, The rise of crowdsourcing. *Wired*, 14.06: 161-165, Available at <http://www.wired.com/wired/archive/14.06/crowds.html> [Accessed on September 25th, 2008].
- LandScan, 2011, *Haiti data*. Available at <http://www.ornl.gov/sci/landscan/> [Accessed on Jan, 26th, 2011].
- Luke S, Cioffi-Revilla C, Panait L, Sullivan K and Balan G, 2005, MASON: A multi-agent simulation environment. *Simulation*, 81(7): 517-527.
- Salgado M, Marchione E and Gill A, 2010, The calm after the storm? Looting in the context of disasters. *Proceedings of the 3rd World Congress on Social Simulation: Scientific Advances in Understanding Societal Processes and Dynamics*, Kassel, Germany.
- Sullivan K, Coletti M and Luke S, 2010, GeoMason: GeoSpatial support for MASON. *Department of Computer Science, George Mason University, Technical Report Series*, Fairfax, VA.
- Thorp J, Guerin S, Wimberly F, Rossbach M, Densmore O, Agar M. and Roberts D, 2006, Santa Fe on fire: agent-based modelling of wildfire evacuation. In Sallach D, Macal CM and North MJ, (eds.), *Proceedings of the Agent 2006 Conference on Social Agents: Results and Prospects*, Chicago, IL.

Vector-based Mathematical Morphology

Huayi Wu, Wenxiu Gao

State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing,
Wuhan University, 129 Luoyu Road, Wuhan, 430079, China
Telephone: (+86) 27 68778311
Fax: (+86) 68778969
Email: wuhuayi@lmars.whu.edu.cn, wxgao@lmars.whu.edu.cn

1. Introduction

Mathematical Morphology (MM) is a mature theory and technique originally developed for processing raster-based binary images. Over time MM was extended to include grayscale functions and images. Today, the generalization of MM to complete lattices is widely accepted as MM's theoretical foundation.

Dilation and erosion are the basic operations of MM; dilation expands a figure uniformly, while erosion shrinks a figure uniformly. Expanding and shrinking can be altered in different directions by employing a non-circular structure element such as a diamond shape to constrain MM operations. A sequential combination of dilation and erosion operations can generate various outcomes from a unique figure. It is very useful in some cases to transform the shape of a figure or extract graphic information from figures. For example, a dilation operation plus an erosion operation may remove small holes and the same combination of operations also can determine if two figures are separate or detached. Analogous combinations of operations are widely used in image processing.

Such transformation and extraction of potential information from figures is also required in vector-based datasets and applications. MM, however, cannot be directly applied to vector data. Thus, vector data must be transformed into raster data before MM processing operations can be executed. After these operations, the vector to raster

transformation must be reversed if vector is the final data type required. Unfortunately, precision and information are lost during the two transformations.

In the past, vector-based data processing was considered as computationally intensive. Therefore, raster data were used as an intermediate format to implement some complex algorithms. Nowadays, enhanced computing power supports efficient vector data processing capabilities. Now it is possible to implement the same figure transformation and information extraction from vector data as described above by developing vector-based MM (VMM). Vector-based MM (VMM) is the focus of this paper. Vector-based dilation and erosion operations are defined so as to directly transform vector figures and extract potential information from vector data. The primary experiment proves that such definitions are of some interesting features and may potentially grow to a systematic methodology.

2. Vector-based operations and structure elements

2.1 Dilation and Erosion

The two basic operations of VMM, dilation \oplus and erosion \ominus , are defined as outward buffer and inward buffer of a vector figure. Figure 1 gives an example of dilation operation.



Figure 1. An example of dilation operation (Shi and Wu, 2003)

The buffer operation is now an ordinary operation in commercial GIS software. However, as a fundamental VMM operation, the basic algorithm must be more efficient; to implement some actions constraints may be set for a buffer operation.

Due to the differences between raster and vector data sets, VMM displays some new dilation and erosion operation features. For example, in VMM, $(A \oplus s) \oplus s = A \oplus (s \oplus s)$ and $(A \ominus s) \ominus s = A \ominus (s \oplus s)$ holds true, even if s is a circle. In MM, errors may occur since the dilation and erosion operations are based on the pixel unit, and not continuous in the 2D plane.

2.2 Structure Element

A structure element is an atomic figure used to generate various different shapes from a single figure. It is at the core of operations in raster-based MM. Likewise, for the vector-based operations defined above in section 2.1, outward and inward buffers can be considered as circle structure elements. We also can use non-circle structure elements. For example, a group of points can generate a buffer with an eclipse as a structure element to indicate the point-source pollution area. Figure 2 shows two examples of different results of a shape dilated with different structure elements.

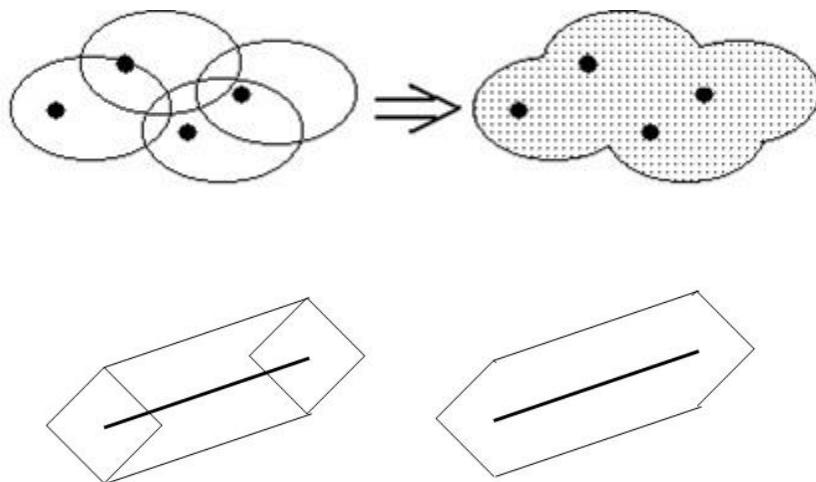


Figure 2. Examples of various dilations of a shape with various structure elements

3. Composite operations

Through combinations of dilation and erosion operations, a series of composite operations can be generated. The two well-known operations are open and close. They are defined as:

$$\text{Open operation: } S \ominus B = (S \ominus B) \oplus B$$

$$\text{Close operation: } S \bullet B = (S \oplus B) \ominus B$$

Some other interesting operations may be transplanted from MM to VMM, but special algorithms must be developed to realize them. For example, a skeleton operation generates the most simplified output shape from an input shape. The algorithm is quite intuitive in MM, but is not developed in VMM.

4. Application of generalization of polygonal map

Based on the dilation and erosion operations of VMM, we designed different sequential combinations of the two operations to extract possible collision information within a single polygon and to simplify polygons. The method provides a feasible means for polygonal map generalization (e.g. a landuse map) to detect geometric conflicts and generalize polygons.

4.1 To extract possible collision information

The black polygon shown in Figure 3 is the original polygon selected from a polygonal map. We implemented inward-outward-buffering to detect collision possibly existing within the left polygon. The blue parts shown in diagram (figure 3) represent the outcome of an inward buffer and the red parts are the outcome of an outward buffer on the blue parts. The areas between the red parts cannot be identified visually on a small-scale map. That is to say, collision may happen at these areas. The performance of the inward-outward-buffering operation is analogous to the open operation of the raster-based MM. Therefore, we define the inward-outward-buffering operation is the vector-based open operation.

Besides collision, there are some small areas between red boundaries and black boundary. Some of these areas are caused by small curves of the polygon boundary. Such details may not be necessary on a smaller-scale map and may interfere the understanding of the whole features. In addition to the small curves along the boundary, some small areas are elongating along a direction but they also may not be clearly visible on a smaller-scale map. These two cases are typical geometric conflicts occurring in polygonal maps during generalization. They are typically processed in different resolutions during map generalization in MM.

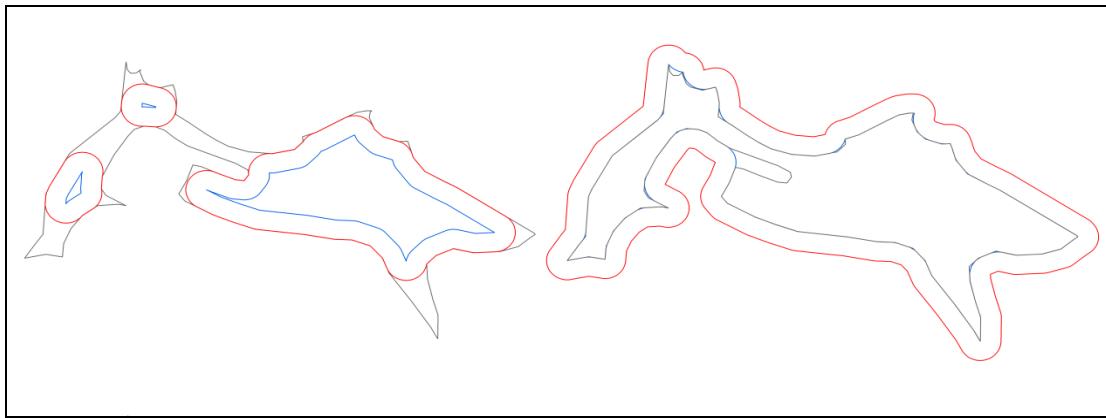


Figure 3 Performance of different sequential combination of dilation and erosion operations

4.2 To simply polygons

In the right polygon (Figure 3), outward-inward-buffering is implemented on the right polygon. The red polygon is the outcome of the outward buffering; the blue inward buffer is generated from the red polygon. The blue polygon matches closely to the black polygon except the small elongated part in the middle and along the small curves in the black boundary. The whole shape of the blue polygon is smoother and simpler than the black polygon. The performance of the outward-inward-buffering operation is analogous to the close operation of the raster-based MM. Therefore, we define the inward-outward-buffering operation as the vector-based close operation.

It is well-known that the performance of raster-based dilation and erosion operations are strongly related to the size of structure element. Likewise, the performance of vector-based operations is also impacted by the size of structure element, i.e. in this study, buffer width. For the open operation, with decreasing buffer width, fewer collision areas will exist as shown in the left polygon (Figure 4). For a close operation, with increasing buffer width, more details along the polygon boundary will be removed (Figure 4). The leftmost polygon in Figure 4 is the original polygon, and Figures 4 (b)-(e) are the outcomes of the close operation with the buffer width, 0.1, 0.2, 0.3, and 0.4mm respectively. The red polygons are the final polygon after the close operation. From the left to the right, the red polygons become more and more generalized. It provides an efficient solution to derive polygonal maps on different map scales.

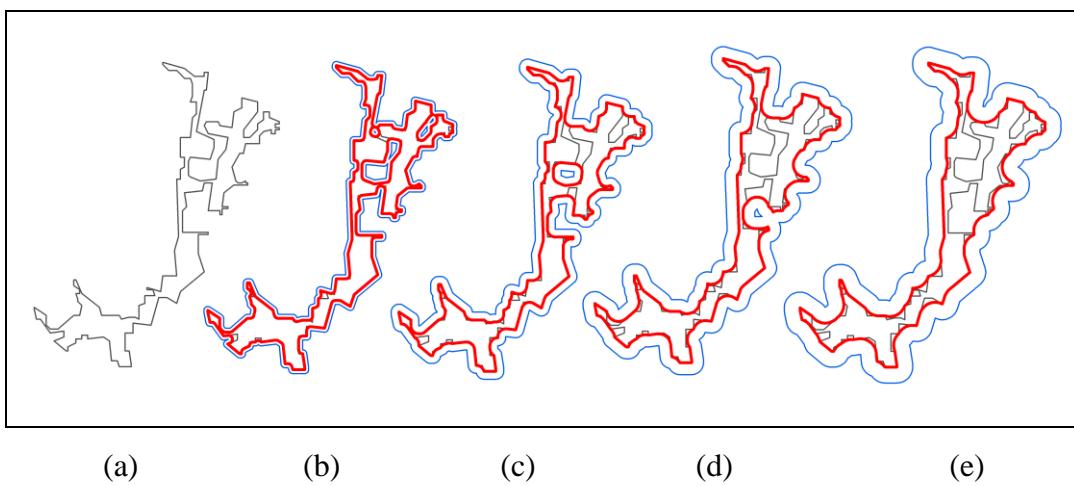


Figure 4. Performance of the close operations with different buffer widths

5. Conclusions

Through outward and inward buffers, a Vector-based Mathematical Morphology (VMM) can be constructed in a systematic way. VMM adds some additional features to those of traditional raster-based Mathematical Morphology. As illustrated by the examples given in this paper, VMM can be applied in map generalization. Many more examples are expected to result from future studies. The work in this paper is a starting point for a new framework for processing map data.

6. Acknowledgements

The work described in this paper is supported by Natural Science Foundation (40971211 and 41023001).

7. References

- Ghosh, P.K., 1991, Algebra of polygons through the notion of negative shapes, CVGIP: Image Understanding, 54(1):119
- Serra, J., 1982, Image Analysis and Mathematical Morphology, London. New York: Academic Press
- Shi, W., Wu, H., 2003, A probabilistic paradigm for handling uncertain objects in GIS by randomized graph algebra, Progress in Natural Science, 13(9):648-657
- Wu, H., 1999, Quasi-triangular Network: Data Structure and Algorithms, Ph.D. Thesis, Wuhan University

Towards Using Geovisual Analytics to Interpret the Output of Geographically Weighted Discriminant Analysis

P. Foley¹, U. Demšar²

¹National Centre for Geocomputation, NUI Maynooth, Co. Kildare, Ireland
 Telephone: ++353 01 708 6731
 Fax: ++353 01 708 6456
 Email: peter.f.foley@nuim.ie

²National Centre for Geocomputation, NUI Maynooth, Co. Kildare, Ireland
 Telephone: ++353 01 708 6178
 Fax: ++353 01 708 6456
 Email: urska.demesar@nuim.ie

1. Introduction

Geographically Weighted Methods are statistical techniques developed to model spatially varying (non-stationary) processes (Fotheringham et al. 2002). Their outputs are spatial datasets which are highly dimensional, complex and large. Interpreting these datasets is a significant challenge. One way to help with this is to use Geovisual Analytics methods.

Geovisual Analytics is the sub-discipline of Visual Analytics that deals with data with a spatial and possibly temporal extent. Visual Analytics combines “automated analysis techniques with interactive visualisations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets” (Keim et al. 2010).

This research aims to use Geovisual Analytics methods to transform the information contained in the output of a specific Geographically Weighed method: Geographically Weighted Discriminant Analysis (GWDA) into new knowledge about the underlying spatial process. This has been done before for other Geographically Weighted methods (Demšar et al. 2008a,b; 2010), but here we extend the principle to GWDA. This abstract describes progress to date and outlines a plan for the remainder of the research.

1.1 Geographically Weighted Discriminant Analysis

Discriminant Analysis is a supervised classification technique used to assign objects in a dataset to distinct classes. Training data are used to estimate the class means, covariance matrices and prior probabilities in attribute space and this information is used to calibrate classification functions of the attributes. Objects are assigned to the class with the maximum classification score. Linear Discriminant Analysis (LDA) outputs include; classification functions that are linear combinations of the attributes, the assigned class and the posterior probabilities which represent the probability that an object belongs to a particular class.

GWDA (Brunsdon et al. 2007) models spatial non-stationarity in the relationship between class membership and the attributes by allowing the parameters of the classification functions to vary spatially. GWDA outputs include; spatially varying classification functions, the assigned class and the posterior probabilities. The GWDA classification functions require analysis to understand the causes of spatial non-

stationarity but this is complex. Not only is there a cognitive difficulty comparing the values of multiple parameters for a single variable (Brunsdon et al. 2007) but in addition, the values of the classification functions are not absolute (Klecka 1980) which means that parameter values cannot be compared directly.

2. Combining Linear Discriminant Analysis with Geovisual Analytics

In this abstract we present the use of tools from the GeoViz Toolkit (Hardisty and Robinson 2010) to interpret the output of LDA and GWDA. Later, we will develop new visualizations specifically suited to exploring the output of GWDA.

2.1 Implementation of Discriminant Analysis in the GeoViz Toolkit

The GeoViz Toolkit is a free and open-source collection of visual and computational tools for exploring geographical datasets. These tools can be used in tandem so that multiple dynamically linked visualizations of the data are possible. Since one of the goals of Geovisual Analytics is to integrate visualization methods and spatial analysis techniques (Hardisty and Robinson 2010), we implemented LDA and GWDA in the GeoViz Toolkit as a first step.

2.2 Data

A data requirement for GWDA to work is that the classes are relatively evenly mixed spatially. In addition, the relationship between the classes and the attributes should vary spatially. We use a simulated dataset to ensure that both of these conditions are met. An advantage of this approach is that the non-stationary spatial patterns are already known so we are able to test the ability of different visualizations to detect them.

We used an existing well-known non-spatial Iris dataset and spatialised it to meet the requirements for GWDA (fig. 1). This dataset was first used by Fisher (1936) and comprises 150 Iris plants of three different species: 50 Iris Setosa, 50 Iris Versicolor and 50 Iris Virginica. Each plant has four associated measurements: sepal length, sepal width, petal length and petal width. To spatialise these data, we assigned the plants to cells on a rectangular grid with 10 rows and 15 columns using the following rules:

1. To ensure an even spatial mix of species, we reserved a random selection of 50 grid cells for each species.
2. To incorporate spatial non-stationarity, plants of each species were assigned to the reserved set of grid cells in a manner that created local patterns. Plants with the shortest petal length were assigned to cells in the bottom left corner of the grid and plants with the longest petal length were assigned to cells in the upper right corner. The ordering is equivalent to the height of an oblique plane over the study area such that the height at the bottom left corner is minimized and the height at the upper right corner is maximized.

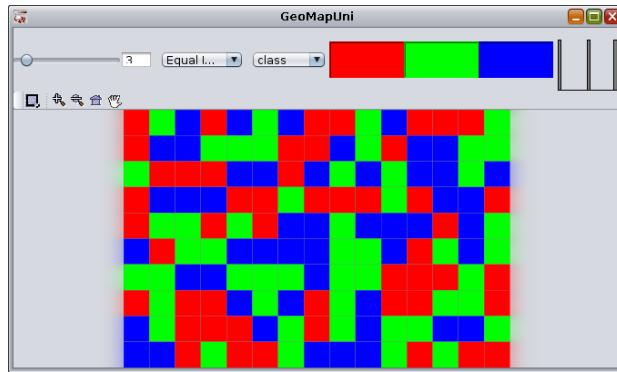


Figure 1. The spatialised Iris dataset showing Iris Setosa cells as red, Iris Versicolor cells as green and Iris Virginica cells as blue.

2.3 Experiment: Visualising LDA Results

Using our implementation of LDA, we classified the simulated spatial dataset using all four Iris measurements as predictor variables and used tools from the GeoViz Toolkit to visualise the output.

The confusion matrix for the classification is shown in table 1. The classification accuracy is 98% and only 3 out of 150 plants were misclassified.

	Iris Setosa	Iris Versicolor	Iris Virginica	Class Total
Iris Setosa	50	0	0	50
Iris Versicolor	0	48	2	50
Iris Virginica	0	1	49	50
LDA Total	50	49	51	150

Table 1. Confusion Matrix from an LDA classification of the spatialized Iris Dataset.

The following tools were found to be useful in visualizing the output of LDA:

1. GeoMapUni is a classified univariate choropleth map. It shows the spatial distribution of a single variable, in our case the 3 species of Iris (fig. 1).
2. GeoMap is a classified bivariate choropleth map. It shows the spatial distribution of two variables with a bivariate colour scheme. We used a bivariate map with a complementary colour scheme (Eyton 1984) to visualize the spatial distribution of the misclassified plants (fig. 2).

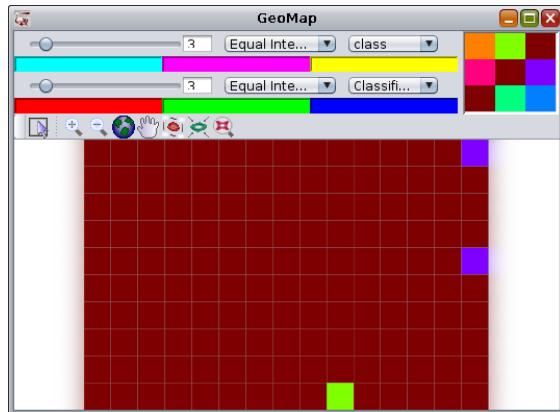


Figure 2. Location of misclassified Iris plants. Dark red cells contain correctly classified plants. The two purple cells contain Iris Versicolor plants misclassified as Iris Virginica and the single green cell contains an Iris Virginica plant misclassified Iris Versicolor.

3. ParallelPlot is a Parallel Coordinates Plot (PCP) to visualize a dataset in attribute space. We used a PCP to visualize the relationship between the three species of Iris and the predictor variables (fig. 3) and to visualize the relationship between the predictor variables, the posterior probabilities of the classification and the species of Iris for the 3 misclassified plants (fig. 4).

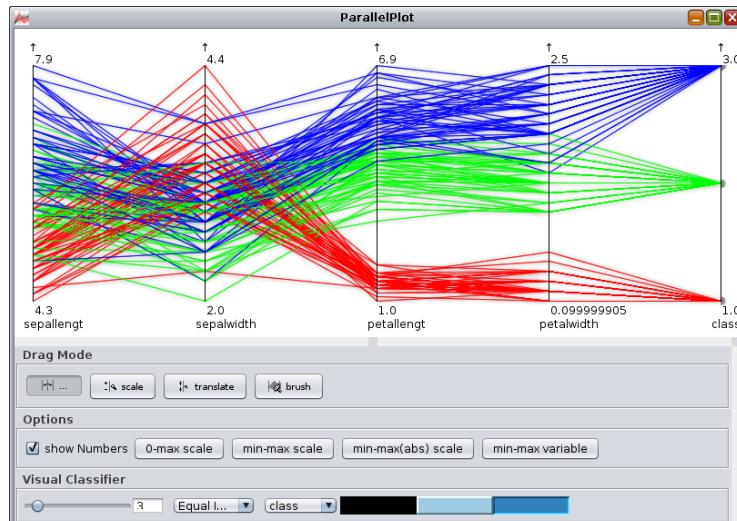


Figure 3. Relationship of the 4 Iris measurements to the species of Iris. Iris Setosa plants are in red, Iris Versicolor plants are in green and Iris Virginica plants are in blue.

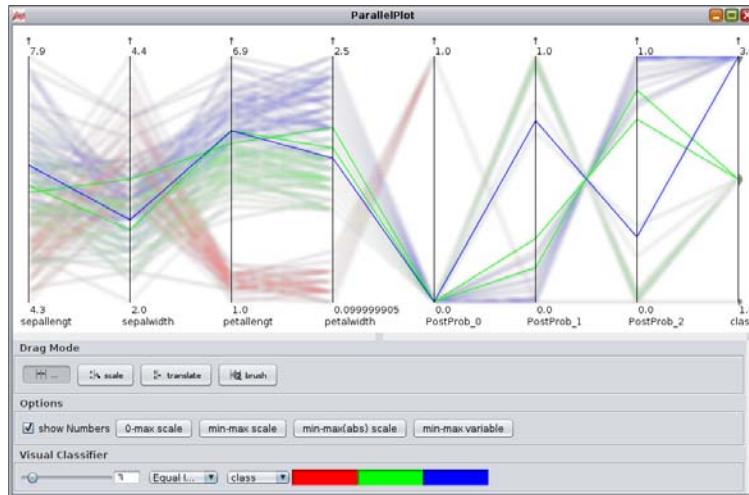


Figure 4. Highlights the relationship of the 3 misclassified plants to the 4 Iris measurements and the LDA posterior probabilities. Iris Setosa plants are in red, Iris Versicolor plants are in green and Iris Virginica plants are in blue.

4. StarPlotMap. This tool shows the spatial distribution of more than two variables using Star Plot icons. We used a Star Plot map to visualize the spatial distribution of the LDA posterior probabilities (fig. 5).

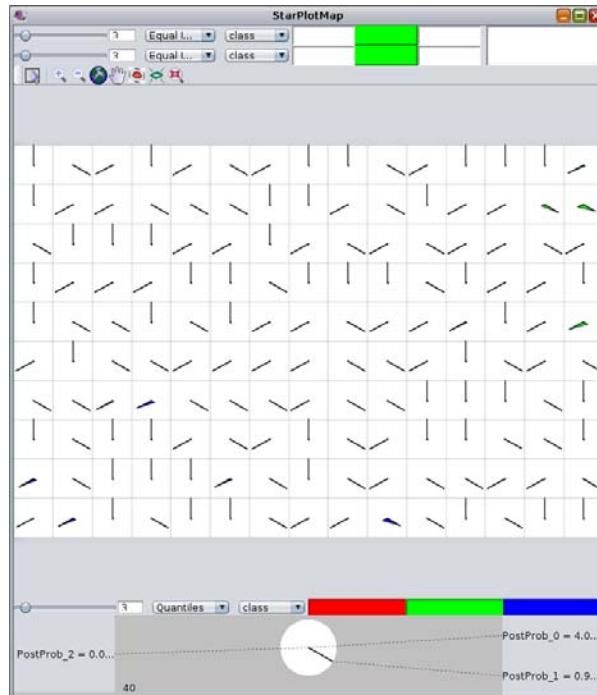


Figure 5. Spatial distribution of the posterior probabilities. The lengths of the rays are proportional to the posterior probabilities for each of the 3 species: rays pointing north for Iris Setosa, rays pointing south-east for Iris Versicolor and rays pointing south-west for Iris Virginica.

3. Next Step: Visualising the GWDA Results

The next steps are to identify the most useful tools from the GeoViz Toolkit to visualize the output of GWDA and finally, to develop new visualizations specifically for exploration of the GWDA output. These should provide additional insight into the output of GWDA and facilitate the interpretation of GWDA results. As this is work in progress, in this section we present some preliminary results from the GWDA classification of the same dataset.

Using our implementation of GWDA, we classified the simulated spatial dataset using all four Iris measurements as predictor variables. The confusion matrix for the classification is shown in Table 2. The classification accuracy is 100% and the variance in the GWDA posterior probabilities is reduced compared to the LDA posterior probabilities (fig. 6). The high classification accuracy for LDA and GWDA make it difficult to attribute the improved results to genuine spatial non-stationarity. Therefore these results should only be considered as preliminary and this experiment should be repeated for another, less ideal dataset. For example, since classification with GWDA performs so well, most of the posterior probability values are either 0 or 1 which accounts for considerable overprinting in the PCP (fig. 6). Therefore, this PCP should only be used in conjunction with other interactively connected visualisations to identify patterns. Note also the contrast between the variance of the posterior probability values in LDA (fig. 4) versus the almost binary separation in GWDA (fig. 6).

	Iris Setosa	Iris Versicolor	Iris Virginica	Class Total
Iris Setosa	50	0	0	50
Iris Versicolor	0	50	0	50
Iris Virginica	0	0	50	50
LDA Total	50	50	50	150

Table 2. Confusion Matrix from a GWDA classification of the spatialized Iris Dataset.

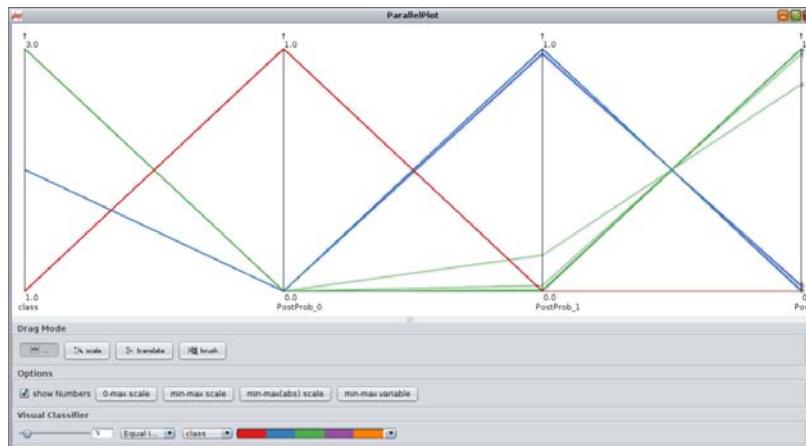


Figure 6. Relationship of species to the GWDA posterior probabilities. Iris Setosa plants are in red, Iris Versicolor plants are in blue and Iris Virginica plants are in green. This PCP shows all 150 plants, but as species are well separated by the posterior probabilities (i.e. they are either 0 or 1), there is a large amount of overprinting present in this PCP.

4. Conclusions

We have demonstrated that specific tools from the GeoViz Toolkit are useful in revealing spatial and non-spatial patterns in the output of LDA. For the remainder of the research we plan to develop new visualisations to provide insight into GWDA.

The tools for visualizing the output of LDA, described in section 2.3 can be used in exactly the same way with the GWDA output. However, the GWDA output presents additional challenges:

1. We need a method to visualize the spatially varying relationship between class membership and the predictor variables and this will require a new technique. A starting point could be to map the variation in posterior probabilities for a fixed set of predictor variables (Brunsdon et al. 2007). This could be improved by allowing the user to vary the predictor variables on the fly. We plan to visualize the posterior probabilities using a Treemap approach (Johnson and Shneiderman 1991). This should improve on the existing visualizations (StarPlot Map and PCP) which suffer from overprinting. We also plan to visualize the confusion matrix using a Mosaic Plot (Hartigan and Kleiner 1981).
2. For this particular dataset, the difference between the classification accuracy for LDA and GWDA is small. For a less ideal dataset, mapping the difference between the LDA and GWDA posterior probabilities would highlight cells where the confidence in the classification has been enhanced or reduced. To decrease the predictive accuracy of the four Iris measurements we have “confused” the dataset by perturbing them slightly (fig. 7). The contrast between the LDA and GWDA classification accuracies (~87% and ~91% respectively) has now increased.

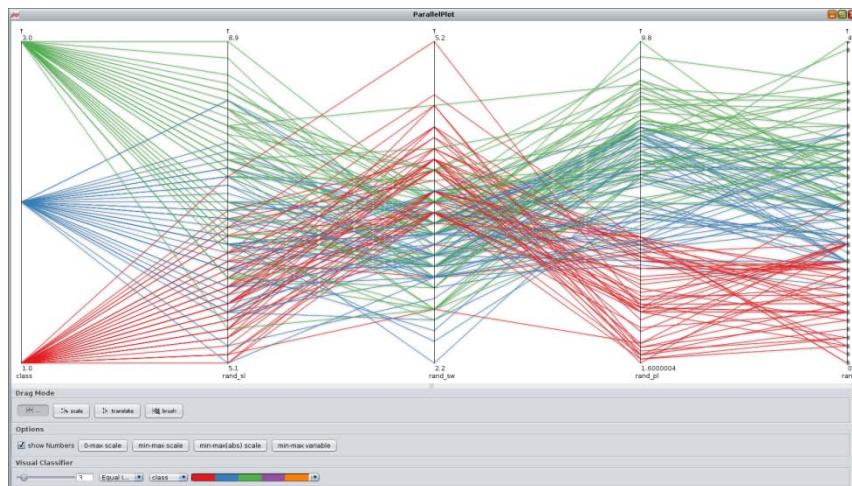


Figure 7. Relationship of the 4 randomized Iris measurements to the species of Iris. Iris Setosa plants are in red, Iris Versicolor plants are in blue and Iris Virginica plants are in green.

3. Identification of outliers in the classes is possible since the Mahalanobis Distance squared from each object to the class means follows a chi-squared distribution with m degrees of freedom where m is equal to the number of predictor variables (Manly 2005). We are investigating possible visualisations for outlier detection based on this.

5. Acknowledgements

We thank Frank Hardisty at the Pennsylvania State University for technical advice and assistance on extending the GeoViz Toolkit. The authors are supported by a Research Frontiers Grant (09/RFP/CMS2250) awarded to Urška Demšar by Science Foundation under the National Development Plan.

6. References

- Brunsdon C, Fotheringham A S and Charlton M, 2007, Geographically Weighted Discriminant Analysis. *Geographical Analysis*, 39(4):376-396.
- Demšar U and Fotheringham A S, 2010, Geographically Weighted Principal Components Analysis and the Curse of Dimensionality. *Journal of Visual Languages and Computing* (Under Review).
- Demšar U, Fotheringham A S and Charlton M, 2008a, Combining Geovisual Analytics with Spatial Statistics: the Example of Geographically Weighted Regression. *The Cartographic Journal*, 45(3):182-192.
- Demšar U, Fotheringham A S and Charlton M, 2008b, Exploring the spatio-temporal dynamics of geographical processes with geographically weighted regression and geovisual analytics. *Information Visualization*, 7(3-4):181-197.
- Eyton J, 1984, Complementary-color, two variable maps. *Annals of the Association of American Geographers*, 74(3):477-490.
- Fisher R A, 1936, The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7:179-188.
- Fotheringham A S, Brunson C and Charlton M, 2002, *Geographically Weighted Regression – the analysis of spatially varying relationships*. John Wiley & Sons, Chichester, England.
- Hardisty F and Robinson A C, 2010, The GeoViz Toolkit: Using component-oriented coordination methods for geographic visualization and analysis. *International Journal of Geographical Information Science (Forthcoming)*.
- Hartigan JA, and Kleiner B, 1981, Mosaics for Contingency Tables. *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, Springer, New York
- Johnson, B and Schneiderman, B, 1991, Treemaps: a space-filling approach to the visualization of hierarchical information structures. *Proceedings of the 2nd International IEEE Visualization Conference*.
- Keim D, Kohlhammer J, Ellis G and Mansmann F, eds, 2010 *Mastering the Information Age - Solving Problems with Visual Analytics*, http://www.vismaster.eu/wp-content/plugins/cimy-counter/cc_redirect.php?cc=full_book&fn=http://www.vismaster.eu/wp-content/uploads/2010/11/VisMaster-book-lowres.pdf
- Klecka W R, 1980, *Discriminant Analysis*. Quantitative Applications in the Social Sciences, Sage Publications, Beverly Hills, California.
- Manly B F J, 2005, Multivariate Statistical Methods: A Primer. Chapman & Hall/CRC, Boca Raton, Florida, third edition.

Discovering Different Regimes of Biodiversity Support Using Decision Tree Learning

T. F. Stepinski¹, D. White², J. Salazar³

¹Department of Geography, University of Cincinnati, Cincinnati, OH 45221-0131, USA
 Telephone: +1 513 .556.3583
 Fax: +1 513.556.3370
 Email: stepintz@uc.edu

²US Environmental Protection Agency, Corvallis, OR 97333, USA
 Email: whitede@onid.orst.edu

³Lunar and Planetary Institute, Houston, TX 77058, USA
 Email: salazar@lpi.usra.edu

1. Introduction

A pressing problem in biodiversity studies is to find the optimal strategy for protecting the species given limited resources. In order to design such a strategy it is necessary to understand associations between spatial distribution of biodiversity and environmental factors. A relationship between a response variable (a suitable measure of biodiversity) and predictor variables (measures of environmental factors) is certain to be complex as it must reflect a non-stationary character of an observed dependence. As a result one can expect an existence of several different biodiversity regimes – sets of environmental conditions *locally* associated with the levels of biodiversity measure. Multi-regime association cannot be discovered using standard methods based on linear regression; here we propose using decision tree learning methodology to discover different regimes of association between environmental variables and richness of bird species (a particular measure of biodiversity) across the contiguous United States.

Fig.1 shows a map depicting spatial distribution of richness (R) of bird species across the US. Distribution of R has a strong bimodal character effectively dividing the United States into high richness (HR) and low richness (LR) regions using a threshold value of $R=148$; this value corresponds to a location of the minimum that clearly separates the two maxima of bimodal distribution of R . The HR region is not simple-connected; instead it consists of several geographically distributed pieces. The premise is that observed distribution of R associates with locally-specific combination of values of environmental variables. We find those associations using a data mining technique based on decision tree learning. This is an expansion of a method proposed by White and Sifneos (2002).

2. Methods

We consider a set of 32 predictor variables pertaining to terrain, climate, landscape metrics, land cover, and environmental stress and hypothesized to have potential influence on bird richness. These variables constitute a subset of a larger dataset (White et al., 1999) and are given on a grid consisting of 12,337 hexagons covering the contiguous United States. A value of response variable R is the count of unique species in every hex. Breeding Bird Survey (BBS) grids (Sauer et al. 1995) representing distribution of individual bird species was used to calculate R ; the values range from $R=21$ to $R=230$.

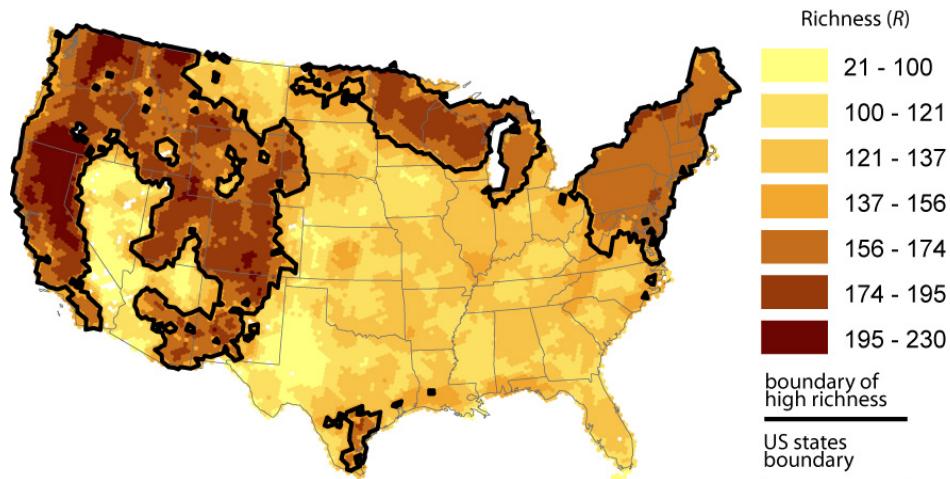


Figure 1. Map of richness, R , of bird species across the contiguous United States.

Data mining technique of decision tree learning (Loh, 2008) uses a decision tree as a predictive model. The model recursively partitions a set of predictor variables until each partition, represented as a terminal node of the tree, contains only data instances (hexes in our dataset) from which a conclusion about the response variable can be made with relatively high accuracy. A unique feature of the tree model is its interpretability; other models often possess good prediction accuracy, but they act like black boxes and do not provide insight into the roles of the predictor variables. Our focus here is *not* on a predictive accuracy of such model (after all, the values of R are known for every hex) but rather on the data partitions that we connect with different biodiversity regimes. We build two conceptually different models. First, we build a *regression tree* model which is a piecewise constant estimate of a regression function. Data is partitioned so as to increase the accuracy of linear regression in each partition. In each terminal node an average value of R serves as a predictor. Nodes are labeled as *HR* if they contain predominantly high values of R and *LR* if they contain predominantly low values of R . Second, taking advantage of a bimodal character of the distribution of R , we start by labeling hexes into *HR* and *LR* using a threshold value of $R_{\text{thres}}=148$, and then build a *classification tree*. In classification tree data is partitioned so as to increase the label purity of subdivisions. Nodes are labeled as *HR* if they contain majority of *HR* hexes and *LR* if they contain majority of *LR* hexes. We used GUIDE algorithm (Loh, 2008) to build regression and classification trees having 12 terminal nodes each. The number of terminal nodes is determined automatically by a process of cross validation.

3. Results

Results of the regression tree model (*RTM*) are shown on Fig.2. *HR* nodes and spatial regions corresponding to them are shown in warm colours while *LR* nodes and spatial regions corresponding to them are shown in cool colours. The overall accuracy of the *RTM* is $\sim 80\%$. The major split of dataset is on the value of July mean temperature. Hexes with July temperatures ≤ 21.8 C are conducive to *HR*; all but one node in the left main fork of the tree are *HR* nodes and there are no *HR* nodes in the right main fork of the tree. Surprisingly, despite a complex character of the dataset, great majority of “higher richness” hexes fulfil a single (*JulyMeanTemp* ≤ 21.82) predicate. Each *HR* node groups predominantly *HR* hexes and thus can be identified with a particular environmental regime conducive to high richness of species.

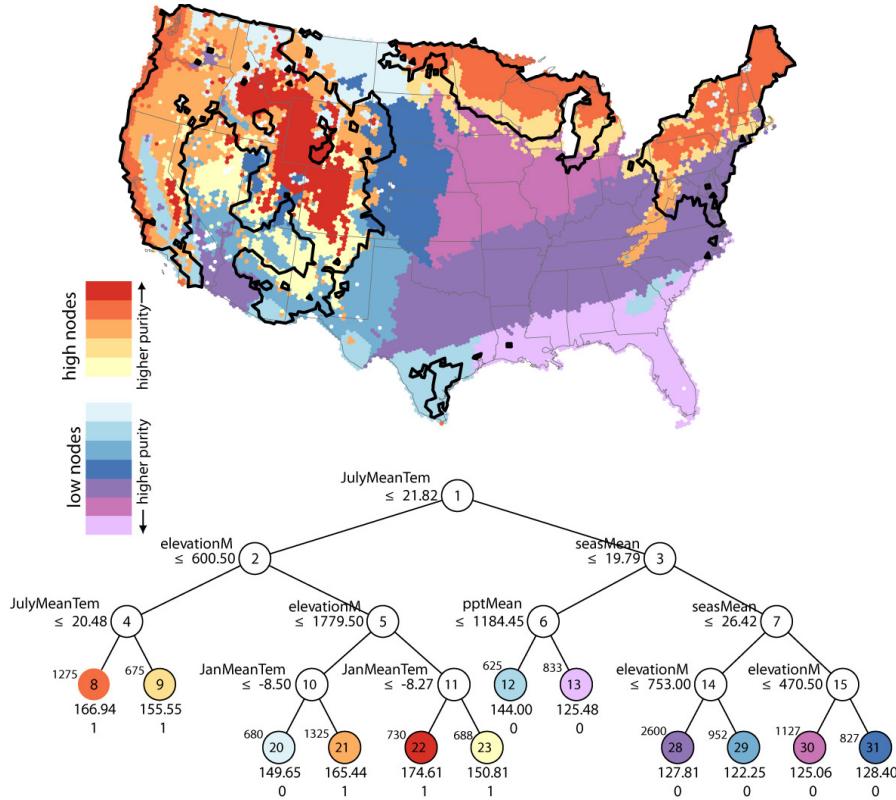


Figure 2. Map of richness of bird species environmental regimes calculated using regression tree. Tree nodes are shown as circles with ID numbers within them. Quantities on the left site of terminal nodes give the number of hexes in the node, quantities immediately below terminal nodes give an average value of R in the node, and the 0/1 labels indicate whether node represents *HR* regime or *LR* regime.

Results of classification tree model (*CTM*) are shown on Fig.3. The overall accuracy of the *CTM* is ~85%. The major split of dataset is on the value of January mean temperature, but *HR* and *LR* nodes are split between the two main forks of the tree. The *HR* node #9 accounts for majority of *HR* hexes.

4. Conclusions

The two models represent different means of decision tree learning and yield seemingly different partitionings of the dataset. From a prediction point of view they are equally useful although the *CTM* has a small edge in accuracy. From a point of view of discovering environmental regimes of biodiversity, each model provides what, at first glance, appears to be a different partitioning of the environmental data. However, closer examination reveals some similarities in spatial extent between a number of nodes in the two partitions. For example, spatial footprint of node #28 in the *CTM* resembles the footprint of node #12 in the *RTM*. Other examples include: *CTM* node #17 and *RTM* node #20, eastern portion of *CTM* node #9 and *RTM* node #8. These correspondences exist because a tree node is described in terms of a series of consecutive predicates, but a similar partition can be feasibly described by a different series of predicates if the predictor variables involved in the predicates are correlated.

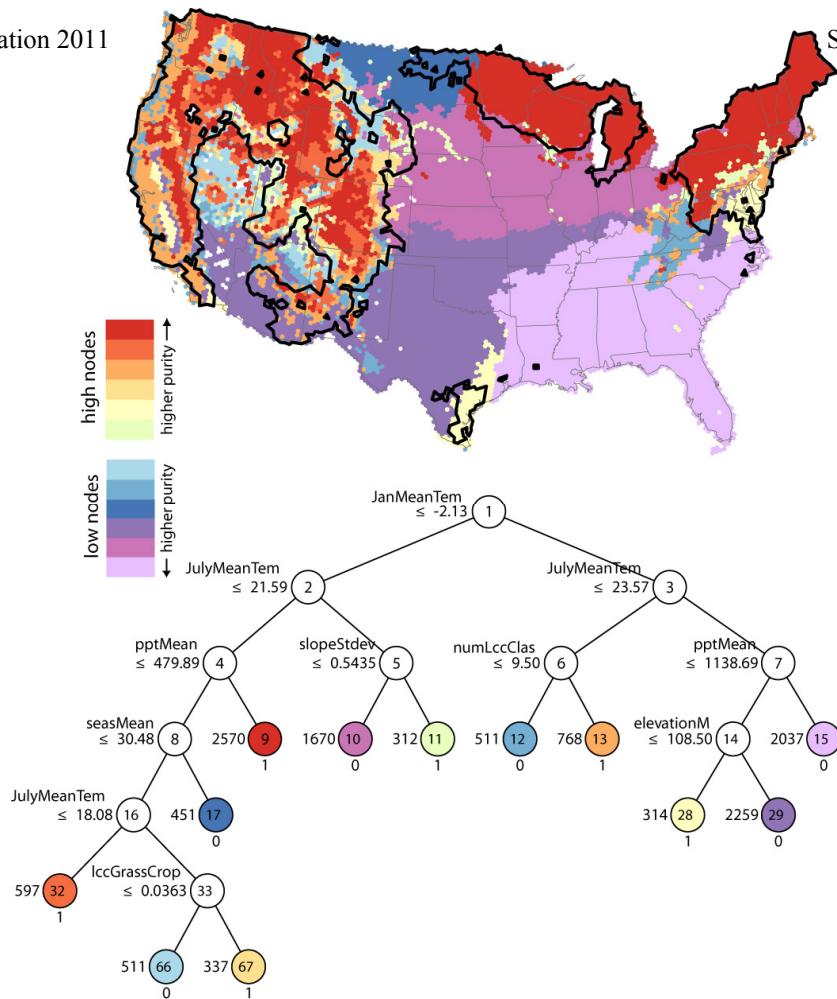


Figure 3. Map of richness of bird species environmental regimes calculated using classification tree. See also caption to Fig. 2.

Analysis of the two models reveals existence of four regimes of high richness of bird species that transcend specificity of the models. They are: (1) Southern regime (*RTM* node #12 and *CTM* node #28), (2) Northern regime (*RTM* node #9 plus portions of node #8 and *RTM* portion of node #9), (3) Mountain regime (*RTM* nodes #22, #21 and *CTM* nodes #32, portion of #9), (4) Pacific Coast regime (*RTM* portion of node #8 and *CTM* node #13). Fig.4 shows spatial extents of these regimes and their characterization in terms of predictors shown as parallel coordinates-like graphs. These characterizations provide compact but comprehensive description of each regime. For example, the Southern regime is not only characterized by climatic variables, as indicated by predicates in both regression and classification trees, but also by presence (predictor 17) and absence (predictors 18 and 19) of specific land cover classes.

Decision tree-based methodology, as presented here, can be applied to a broad range of non-stationary spatial problems where there is a need to identify different regimes of dependence between predictors and response.

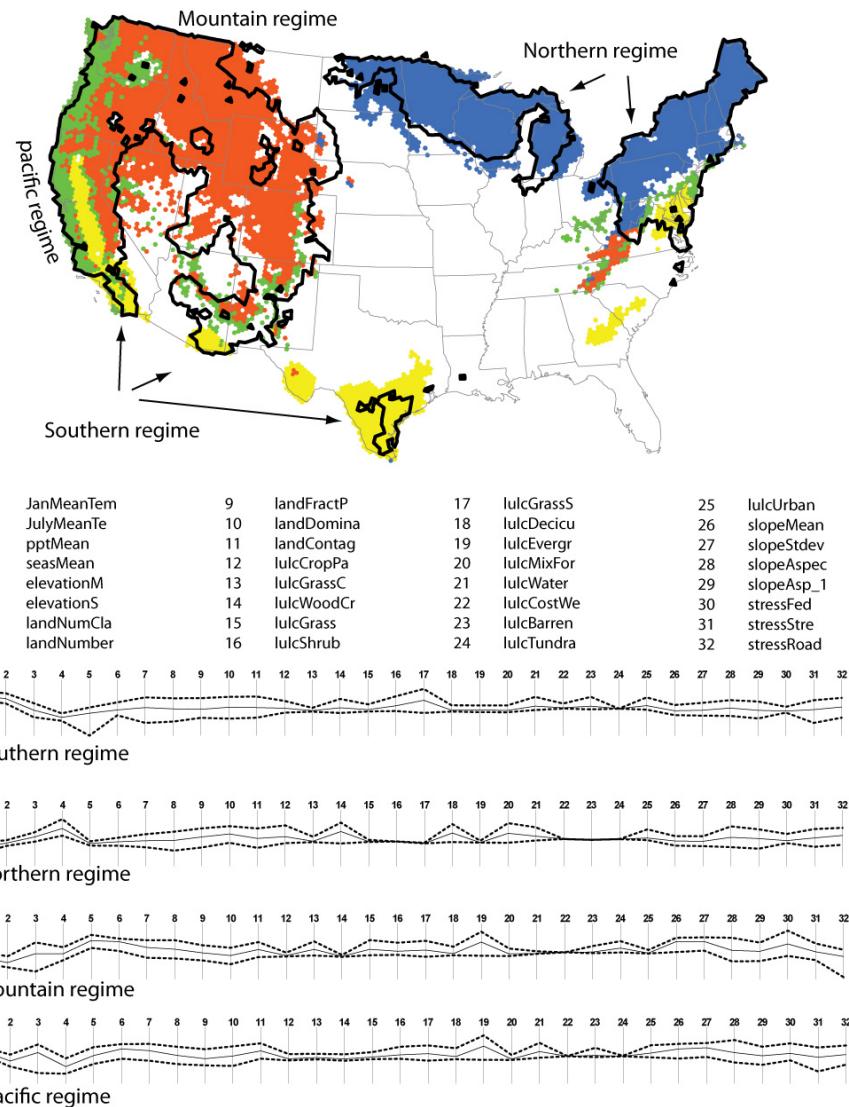


Figure 4. (Top) Map of four regimes of high diversity of bird species in the United States. (Bottom) Characterization of the four regimes in terms of all predictors; solid line – mean values, dashed lines – mean values \pm standard variation.

7. References

- Loh, W.-Y., 2008, Regression by parts: Fitting visually interpretable models with GUIDE. In: *Handbook of Computational Statistics, vol. III*, Springer-Verlag, 447-469.
- Sauer, JR., Pendleton GW, and Orsillo, S, 1995. Mapping of bird distributions from point count surveys. In: Ralph CJ, Sauer JR, and Droege S (eds) *Monitoring Bird Populations by Point Counts, USDA Forest Service, Pacific Southwest Research Station*, General Technical Report PSW-GTR-149. 151-160.
- White D, Preston B, Freemark K, and Kiester A, 1999, A hierarchical framework for conserving biodiversity. In: Klopatek J and Gardner R (eds), *Landscape ecological analysis: issues and applications*, New York: Springer-Verlag, 127-153.
- White D and Sifneos J.C., 2002, Regression tree cartography. *J. Computational and Graphical Statistics*, 11(3):600-614.

Knowledge Discovery for Exploring the Relations between Climate Change and Population Dynamics

Budhendra Bhaduri, Xiaohui Cui, Cheng Liu, Jennifer Santos-Hernandez, Benjamin Preston, Jack Schryver, James Nutaro, Stan Hadley, Richard Medina, Hoe Kyoung Kim

Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

Introduction

Climate and human mobility are essentially interconnected and interdependent. Our mobility through ground transportation system powered by fossil fuel is one of the primary forces behind the two major global crises of today's society, namely energy scarcity and climate change. On the other hand, long term change in climate and frequency of climate extreme events, such as hurricanes, floods, snow and ice storms can have both short term mobility challenges and cause long term human migrations as an adaptation phenomenon. Human settlements develop around stable environments where shelter and sustenance are found, and on a higher level, economies can be built. Climate change is expected to shift these stable regions and consequently induce large scale migrations, which in turn can result in famine, cultural conflict, disease propagation, and stress on natural resources and critical infrastructures. In the near term, to reduce oil dependence, environmental impacts, and congestion, a number of alternative energy supply, distribution, and end-use transportation systems, technologies and policies are presently being explored. However, it is still unclear when and in what precise combination these sources and technologies will emerge as successful and sustainable solutions. Ideally, future plausible development and implementation strategies for alternative energy resources and technologies will secure and support a societal system in which energy, environment, and mobility interests are simultaneously optimized. In the longer term, under climate change scenarios it is plausible to expect displacement, migration, and resettlement as an interactive consequence of climate change and its impacts on water resources, land cover and land use. Given the intertwined nature of such a system across wide geographic scales, assessing the effectiveness of possible planning strategies and discovering their unanticipated consequences require data collection, modeling, and simulation at the finest data, process, and societal response levels *coupled* with the system's behavior over large spatial and temporal scales.

Knowledge Discovery from High Resolution Data Driven Simulations

The process of knowledge discovery often extends beyond common data mining techniques on volumes of disparate data to detect patterns, to a new level whereby high resolution data are coupled with modeling and simulations of physical systems to test hypotheses and discover "evolving or emerging" behaviors and trends. In the latter case, such emerging behaviors often reflect unforeseen and undesired consequences of system design. For example, using a high-resolution climate model, researchers have showed that unique spatial pattern of land surface heterogeneity (due to clear cutting along roads) can trigger mesoscale circulations leading to more clouds and rain over the cleared patches [1]. Subsequent studies with satellite data have validated this hypothesis. Another interesting example that illustrates an energy policy-relevancy of similar modeling and simulation based research is the impact of large wind farms on local meteorology [2]. It was shown that turbulence generated in the

wake of the turbine rotors can significantly affect surface temperature and humidity. These effects can be minimized by reducing rotor-generated turbulence. Interestingly, low-turbulence rotors are also economically efficient because they can harness the energy that would have otherwise been lost to turbulence. These results have important implications for land use planning through siting, design and impact assessment of wind farms.

Research Challenges and Opportunity

For knowledge discovery, characterization of the interactions between the human dynamics and transportation infrastructure or climate change are essential and requires integration of three distinct components, namely, data, models and computation. In transportation, previous research has attempted to develop simulations to address scenarios regarding the relationships among energy, emissions, air quality, and transportation. These include detailed physical models of transportation engineering, including CORSIM, TRANSIMS [7, 3], VISSIM [4], PARAMICS and OREMS [8]. Very recently, few models have started addressing the human dynamics of physical and social systems, such as SEAS [5] and Repast/Mason [9] and others [6]. However, none has been able to successfully integrate both the physical as well as behavioral aspects to characterize the interdependencies within the US transportation system and can address the interplay between energy, environment, and quality of life.

For climate change impacts, many researchers have studied regional vulnerabilities. Multiple vulnerability indexes have been developed that can be applied in climate change impact models include the Environmental Sustainability Index (ESI) created by the Yale Center for Environmental Law and Policy and CIESIN at Columbia, the Environmental Vulnerability Index (EVI) from the South Pacific Applied Geosciences Commission, and the Social Vulnerability Index (SoVI) from the Hazards & Vulnerability Research Institute at the University of South Carolina. Each of these indices has its own take on vulnerability, which is a large factor in how regional populations will be affected by changing climates. The impacts of climate change will vary by region, because not all populations are as vulnerable to changes [10]. A recent collaborative effort from the UN University Institute for Environment and Human Security, CARE International and the Center for International Earth Science Information Network (CIESIN) identified many potential threats of climate change and regional vulnerabilities to climate change; however, the report is very clear in concluding that the research does not attempt to characterize how many migrants will likely be displaced by climate change, or their probable destinations. Climate induced migration is the result of many forces that exist in a complex space of social, psychological, cultural, physical, and economic systems. While there are many theories of what will drive future migrants and what this means to global stability, there are presently no highly detailed conceptual or computational models that focus on the problem.

Both transportation and climate induced migration is the result of many forces that exist in a complex space of social, psychological, cultural, physical, and economic systems. Progress has largely been limited by data and computational challenges necessary for accommodating the required high resolution along spatial, temporal and behavioral dimensions [11]. Integration of high resolution socio-demographic data [12] and models bring much promise for capturing the social/behavioral dimension [13]. This dimension is essential in enabling us to characterize the interplay and interdependencies

between (transportation) technology and societal features or between climate change and human migration that are likely to: (i) have an impact on the success of future transportation or adaptation technologies and (ii) be overlooked by current approaches of modeling at aggregated scales.

A High Performance Modeling and Simulation Framework to Implement Scenarios

To represent the complex social phenomena, we have developed a modeling and simulation framework to utilize an Agent Based Modeling (ABM) platform as well as a discrete event modeling platform. We employed both a micro and a meso scale simulation approaches with implementation of social units (e.g., individuals, households, firms, or nations) and their interactions, and observe the global structures that these interactions produce. In this paper we describe our efforts in developing benchmark databases that enable a scalable modeling and simulating framework that can be utilized across the entire US. We illustrate the capability by simulating a transportation scenario that allows an assessment of plausible market penetration Plug-in Hybrid Vehicle (PHEV) at a sub-County scale and its potential impact on the local electric grid and reducing the carbon footprint through displacement of gasoline. Specific insights derived from the results are highlighted to illustrate the complexity of the demographic dependency for the future success of novel transportation technologies. We will also present results from our ongoing research to highlight the development of a conceptual model of climate induced migration. Specifically, we will review the current state of the art in high resolution modeling and simulation for addressing this class of spatial analysis and the associated computational challenges. Ongoing effort to develop benchmark databases for such spatially explicit modeling and strategies to extend that to a computational modeling and simulation framework will also be discussed.

Acknowledgement

Financial support for this research was provided by the Laboratory Directed Research and Development (LDRD) program at Oak Ridge National Laboratory. This manuscript has been authored by employees of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. Accordingly, the United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

1. Baidya Roy, S. and R. Avissar, Impact of land use/land cover change on regional hydrometeorology in Amazonia, *Journal of Geophysical Research.*, 2002.
2. Baidya Roy, S, et al., Can large wind farms affect local meteorology? *Journal of Geophysical Research*, 2004.
3. Barrett, C., K Birkbigler, L Smith, V Loose, R. 1995. An Operational Description of TRANSIMS, Los Alamos, New Mexico, Los Alamos National Laboratory, 1995

4. Bloomberg, Loren, and Jim Dale. 2000. Comparison of VISSIM and CORSIM traffic simulation models on a congested network. *Transportation Research Record*. (1727):52-60.
5. Chaturvedi, A., et al. (2005). Bridging Kinetic and Non-kinetic Interactions over Time and Space Continua. *Interservice/Industry Training, Simulation and Education Conference*.
6. De Almedia, C. M., et al. (2005). "GIS and Remote Sensing as Tools for the Simulation of Urban Land-Use Change." *International Journal of Remote Sensing* 26(4): 759-774.
7. Franzese, O., et al. (2001). A Methodology for the Assessment of Traffic Management Strategies for Large-scale Emergency Evacuations. *11th Annual Meeting of ITS America*.
8. Meister, K., et al. (2006). A Comprehensive Scheduler for a Large-scale Multi-agent Transportation Simulation. *International Conference on Travel Behaviour Research*.
9. North, M. J., et al. (2006). "Experiences Creating Three Implementations of the Repast Agent Modeling Toolkit." *ACM Transactions on Modeling and Computer Simulation* 16(1):1-25.
10. Smit, B. and J. Wandel. (2006). Adaptation, adaptive capacity and vulnerability. *Global Environmental Change*, 16, 282-292.
11. Perumalla, K. S. (2006). A Systems Approach to Scalable Transportation Network Modeling. *Winter Simulation Conference*, IEEE.
12. Bhaduri B., Bright, E., Coelman, P, and Urban, M. (2007). LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal* 69:103-117.
13. Perumalla, K. and B. Bhaduri. (2006). On Accounting for the Interplay of Kinetic and Non-Kinetic Aspects in Population Mobility Models. *Proceedings of the European Modeling and Simulation Symposium*, Barcelona, Spain. October 2006.

LOCATION BASED SOCIAL NETWORKS – TRACKING ACTIVITY IN AN URBAN ENVIRONMENT USING TWITTER DATA

F. Neuhaus

Centre for advanced Spatial Analysis, UCL, 1-19 Torrington Place WC1E 7HB London – fabian.neuhaus@ucl.ac.uk

KEY WORDS: Social network, Location based information, Tracking, Geographical visualisation, Urban studies, Spatio-temporal

ABSTRACT:

Increasingly people use digital or online networks to communicate and interact. This changes the socialscape of the urban area and with it the interactive hot spots change and fluctuate throughout the city as individuals follow the narrative path of their everyday routines. People leave messages, distribute news and respond to conversations not only in traditional locations anymore but potentially anywhere in the city.

This paper discusses the emerging potential of social media data used for urban area research and city planning. Working with crowd sourced data in a web 2.0 manner as described for example by Hudson-Smith et All. (2009). Specifically we look at the connections between the emerging social network, as for example described by Boccaletti et All (2006), and the local physical surrounding and conditions. Also aspects of visualisation as well as privacy and ethical implications are discussed.

The information gathered from social media networks, is gathered directly of the platform used by the network participants as for example already employed by Eagle et all (2009) in their study of social networks using mobile phones. The twitter data however, usually can be associated with a physical location for example via the GPS of the smart phone. Research using this location based technology together with a temporal structure has been demonstrated for example by Reads et All (2009). For this virtual social network and infrastructure-mapping project, the data is derived from the Twitter micro blogging service directly via the API and aims to merge the previously listed approaches into a combined location based temporal network.

These local activity are analysed and visualised based on networks of interaction. Who knows whom and get in touch with whom? However the social networks in the sense of specific interest are these datasets in relation to place and how this location based network enable the individual to shape a distinct sense of place.

1. INTRODUCTION

Where is the city active and does it physically change over time? Urban areas are no static artefacts as they are preferably described in texts and theories. Urban areas are buzzing hot spots of human activity that, to some extend, manifest themselves as or utilise built structure, but are largely temporal and ephemeral. Meaning that no constant being of this 'artefact' is present, but merely a past aggregate is telling tales of memories and rumours.

In an attempt to listen to these stories and narrative as they unfold through the streets, alleyways, in courts, buses, on roof terraces or in swimming pools the social networking platform twitter was employed to reconstruct the cities activity hotspot as a time-frozen 'New City Landscape' drawing out the ever changing locations of people's presence and power of spatial creation through narratives and activity.

From the collected data a new landscape based on density is generated. The features of this landscape of digital activity correspond directly with the physical location of their origin but at the same time represent with hills the peaks of locations from where the activity tales are submitted. The flanks and valleys stand for areas with lesser activity and vast plains and deserts of no twitter tales stretch across the townscapes that lay dormant. These New City Landscape (See for example Figure 1) maps don't represent any physical features, but the interaction with physical features on a temporal basis. The digital realm has become as much part of the urban environment as the physical features and with these tweetography maps they are made visible for the first time. The maps allow us to make a direct comparison between real word activity, physical location and digital message. In a globalised world this local reference

develops an increased importance as a sense of place, a source of identity and memory. The digital social media data allows us to investigate into this realm of peer groups' social location interaction, combining the global scale with its local source.

Some of the physical features of the city that are shining through are the major infrastructure installations. The airports on one hand are the examples of quite intense activity and the parks on the other hand manifest themselves through the absence of activity, virtual social networking activity. Where as at airports users might be bored waiting or excited to just have landed, people in the park are engaged in physical activities other than tweeting and these locations are left virtually empty. A great example is the Central Park in New York, a virtual twitter activity desert, where around it and Manhattan as such is a very high tweet area.

2. METHODS AND TECHNOLOGY

The technology to collection twitter data is based on the technology developed for the Tweet-O-Meter (tom). This service was developed at CASA by Steven Gray. Similar to the tom service the data is collected using the twitter API. Twitter offers two different services through the API. One is the Streaming API and the other one is the Search API. For NCL we are using the Search API because of the built in spatial search function.

2.1 Process

With this spatial search we can filter the incoming messages as for a specific urban area. For the NCL maps we have defined the urban areas consistently as an areas with a 30km radius around an urban centre.

The search query will pass down from the twitter feed all messages to fit this criteria. The software will store all these results in a database continuously.

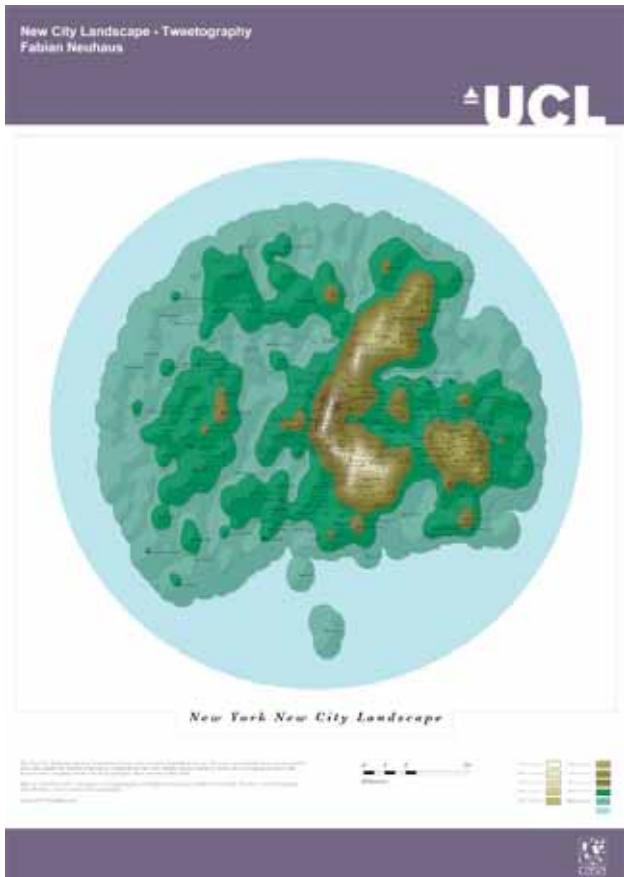


Figure 1. - New York New City Landscape Map showing the twitter activities over the period of one week as a density surface within a radius of 30 km of the NY urban area.

Due to IP limitations imposed by twitter and infrastructure limitations, we are only able to run four parallel search and collect queries at the time. Depending on the search location the resulting amount of data can be quite large, putting quite some pressure on the infrastructure. In order not to miss out on messages, the responding times of the system cannot be compromised.

The data collection per location as been limited to one week, seven days, of consecutive logging of messages sent using the twitter service. One weeks provides good comparison data over a number of days but also shows the different patterns between weekdays and weekends as well as within 24 hours.

The data collected as such therefore is already a spatially defined subset of the total number of tweets sent. However, the collected material needs to be reprocessed because the location information quality is not the same in all messages. Some messages are reverse geocoded from profile information, which generates generic place information.

The resulting data set holds all messages containing real GPS information as Latitude and Longitude coordinates. With this information a more accurate mapping is possible. It is assumes that the accuracy of this information lies within the normal range of GPS accuracy of some 5 to 15 metre.

In a second step a social network is computed, based on the interactions of users in the dataset. To do this especially retweets (RT), twitter messages that have been resent by other users and at-tweets (@), messages specifically addressed at selected twitter users are employed to establish links between individual tweeters as well as a direction of interaction.

Together with the emerging social network and the location as well as the temporal information contained in the data a location based temporal social network can be visualised.

2.2 Data

The amount of data collected varies dramatically between the different locations. There are clearly the very actively tweeting cities such as New York and London with more than 800'000 location based messages sent over the course of one week. On the other hand there are a lot of places especially non-English speaking countries with far less activity, down to a few hundred. Additionally the total location based tweets and the actual GPS tagged messages diverge a lot. Furthermore there is not a simple, more messages result in more GPS tagged messages, equation that applies. It can well be that an very active place turns out very few Latitude/Longitude stamped tweets. As it appeared for example in the case of Sydney, Jakarta or Sao Paulo, where the percentage of geotweets is below 1 % of all location based messages.

Twitter is a relatively new service, being around some four years. The number of users is continuously growing dramatically. This fact put some constraints on the comparability of the data samples. Also the short-term usage of the service is loosely connected to large media events and it is expected that numbers fluctuate quite a bit.

3. RESULTS

The point cloud of twitter messages drawn from the database and mapped using a Mercator projection. This universal projection allows for recognition and readability of urban areas located around the globe.

For the mapping the individual point are being aggregated as a density surface.

Throughout the emerging landscape features have been renamed to reflect these conditions. The new names are fabricated using the real world names in combination with a landscape description of the virtual surface overlaid. This could be 'Mountain' or 'Peak' for high points, 'Slope' or 'Valley' for descending features or 'Desert' and 'Meadow' for average and consistent areas. Inactive areas are termed for example 'Quarry' or 'Ditch'. Together with the familiar real world element the locations become tangible and memorable points of orientation and maybe identification.

The defining landscape features in the virtual NCL map are the hot spots of twitter activity, the peaks. Here the morphology varies between the urban Areas dramatically. How the twitter traffic structures the NCL is unique for each city. There are however some characteristics that can be pointed out.

The different groups could be described as central, where one main location towers over the whole urban region, the multi, where different hotspots appear as peaks across the landscape and the featuring, where one or more features draw out as shapes, groups of peaks or ridges.

Linking this to the social network, see Figure 2, the activity pattern in the temporal sense gain of importance. The variety of

different pattern displayed by different groups is very distinct from activity pattern we normally see in everyday activities. There is more scope for the individual to jump in and out of activities, but connections on hold and reactivate others than what we know from real world interaction.

On the city side the transformation of network activity hubs through out different time periods are striking and offer a new perspective on urban area usage as well as sense of place. Application for this can be found many ranging from urban planning to transport management and modelling to health and safety as in the spatial spreading of information or infections.

4. BIBLIOGRAPHY

Boccaletti, S. et al., 2006. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5), pp.175-308.

Eagle, N., Pentland, A. & Lazer, D., Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*. Available at: <http://www.pnas.org/content/early/2009/08/14/0900282106.abstract> [Accessed January 28, 2011].

Hudson-Smith, A. et al., 2009. Mapping for the Masses. *Social Science Computer Review*, 27, pp.524–538.

Reades, J., Calabrese, F. & Ratti, C., 2009. Eigenplaces: analysing cities using the space – time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5), pp.824 – 836.

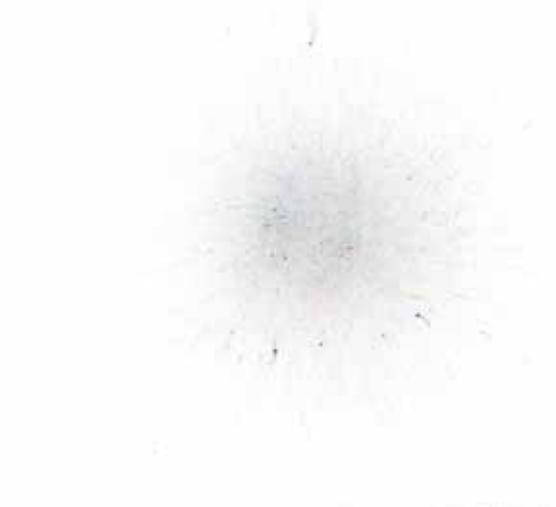
London NCL Social Network

Figure 2 - Twitter activity based social network using the London NCL map data collected in a radius of 30 km within the London urban area.

Modelling Dynamic Space-Time Autocorrelations of Urban Transport Network

T. Cheng, J. Wang, J. Harworth, B.G. Heydecker, A.H.F. Chow

Civil, Environmental and Geomatic Engineering, University College London, Gower Street, London WC1E 6BT, United Kingdom.
 Telephone: +44 207 679 2738
 Email: tao.cheng@ucl.ac.uk

1. Introduction

Various methods for modelling space-time data have been proposed over the years, including multivariate autoregressive integrated moving average (ARIMA) models and its extension space time autoregressive integrated moving average (STARIMA) models (Pfeiffer and Deutsch, 1980). In these time series models, autocorrelation is accounted for in the autoregressive and moving average terms. Parameter estimates are fixed globally both spatially and temporally. The models assume that the correlation in data can be adequately described by such globally set parameters, but this may not be the case if the correlation between data is dynamic, which it is likely to be on road transport networks (Cheng et al, 2011). For instance, traffic theories say that the current conditions on a section of road are influenced to some extent by the previous conditions of adjacent road sections along both upstream and downstream directions (see for example, Lighthill and Whitham, 1955; Richards, 1956). In congested conditions, the influence will come mainly from downstream whereas in free flowing conditions the influence will come from upstream. On a road network comprising hundreds or thousands of links, such spatio-temporal autocorrelation structure is dynamic (in time) and heterogeneous (in space). Yue and Yeh (2008) show the correlation between locations on a road network determines the forecast ability of a space-time model. This fact has been recognised in previous studies that achieve improved results by incorporating a dynamic structure in their weighting systems (Min et al, 2009; 2010; Min and Wynter, 2011). The aim of this study is to model dynamic autocorrelations of road transport network data By modifying tradition model to a generic dynamic model which capture the autocorrelation locally (heterogeneity) and dynamically (dynamic state of the network) over the traditional time series models. The proposed model is tested with traffic data collected from Central London. The result shows that the performance of estimation and prediction is improved on average through the proposed modifications.

2. A Localised Dynamic Space-Time Model - NSTARIMA

STARIMA model considers the observation at location i during time interval t to be a weighted linear combination of observations in its spatial neighbours at previous time intervals. Consider that a road network, in which measurements (e.g. speeds, journey times, etc) are collected on N links over a time period T . Let $\mathbf{z}(t)$ be an N -dimensional

column vector containing the observations $z_i(t)$ on each link i , where $i = 1, 2, \dots, N$, during each time interval t , where $t = 1, 2, \dots, T$. STARIMA model states that

$$\hat{\mathbf{z}}(t) = \sum_{k=1}^p \sum_{h=0}^{m_k} \varphi_{kh} \mathbf{W}^{(h)} \mathbf{z}(t-k) - \sum_{l=1}^q \sum_{h=0}^{n_l} \theta_{lh} \mathbf{W}^{(h)} \boldsymbol{\epsilon}(t-l), \quad (1)$$

in which $\hat{\mathbf{z}}(t)$ is a N -dimensional column vector of predictions on all links i at time t . The first term in the equation is the autoregressive (AR) component, while the second term is the moving average (MA). The term, $\boldsymbol{\epsilon}(\cdot)$, is a N -dimensional column vector of residual on each link. The spatial lag (h) represents the spatial distance between two locations. The spatial orders associated with each k^{th} or l^{th} temporally lagged term in AR and MA components are respectively m_k and n_l . The spatial order specifies the spatial extent that could have an effect on the link of interest i within the temporal lags of k and l . The notation φ_{kh} and θ_{lh} are the model parameters to be calibrated. The matrix $\mathbf{W}^{(h)}$ is an $N \times N$ spatial weight matrix for spatial lag h . This spatial weight matrix $\mathbf{W}^{(h)}$ contains the set of weights w_{ij} specifying the degree of spatial correlation between links i and j (see Kamarianakis and Prastacos, 2005; Getis, 2009).

We identify several deficiencies of the above STARIMA model for traffic modelling and propose a new dynamic time series model – which we call NSTARIMA – that includes several new features. Details are discussed below.

2.1 Spatial orders

Traditional STARIMA model considers the spatial orders to be fixed and preset for the associated temporal lag. It may not be appropriate for traffic modelling as the spatial influences vary under different traffic conditions due to different speeds encountered (Min et al., 2008). This study relaxes such assumption and considers the spatial orders to be dynamic and dependent on traffic state. Given the model updating time interval (Δt), the spatial order $m_k(t)$ at time t associated with temporal lag k is determined as

$$m_k(t) = \arg \min_m \left\{ m \left| \sum_{i_0=i}^{i-m} z_{i_0}(t) L(i_0) > k \Delta t \right. \right\}, \quad (2)$$

where $L(i_0)$ is the length of the intermediate link i_0 between the link of interest i and the spatial extent m . Essentially, $m_k(t)$ returns the number of links that traffic can proceed toward the point of interest i in a time period of $k \Delta t$.

2.2 Spatial weight matrix

The spatial weight matrix ($\mathbf{W}^{(h)}$) is usually regarded as the physical distances between the corresponding locations. In road traffic setting, the correlation of traffic at two locations does not only depend on the spatial distance, but also on the traffic conditions. We propose a novel spatial weight matrix which takes the traffic states into account. For a link pair (i, j) , the corresponding element in the spatial weight matrix is defined as

$$w_{ij} = \frac{v_j(t) - v_i(t)}{D_{ij}}, \quad (3)$$

where $v_i(t)$ and $v_j(t)$ are the respective average speeds on links i and j during time interval t ; D_{ij} is the distance between i and j . The speed $v_i(t)$ is defined to be zero if no data is observed on the link during time t . The spatial weight matrix derived using (4) is time-varying and traffic state dependent.

2.3 Model formulation

We formulate our new time series model – NSTARIMA - as

$$\hat{\mathbf{z}}(t) = \sum_{k=1}^p \sum_{h=0}^{m_k(t-k,i)} \boldsymbol{\varphi}_{kh} \mathbf{W}^{(h,t-k,i)} \mathbf{z}_i(t-k) - \sum_{l=1}^q \sum_{h=0}^{n_l(t-l,i)} \boldsymbol{\theta}_{lh} \mathbf{W}^{(h,t-l,i)} \boldsymbol{\epsilon}_i(t-l). \quad (4)$$

Original STARIMA model is specified by a single global set of parameters ($\boldsymbol{\varphi}_{kh}$, $\boldsymbol{\theta}_{kh}$) for the entire network. In this new model, the model parameters are $N \times N$ diagonal matrices ($\boldsymbol{\varphi}_{kh}$ and $\boldsymbol{\theta}_{kl}$):

$$\boldsymbol{\varphi}_{kh} = \text{diag}([\varphi_{kh}]_1, [\varphi_{kh}]_2, \dots, [\varphi_{kh}]_N) \text{ and } \boldsymbol{\theta}_{lh} = \text{diag}([\theta_{lh}]_1, [\theta_{lh}]_2, \dots, [\theta_{lh}]_N), \quad (5)$$

where $[\varphi_{kh}]_i$ and $[\theta_{lh}]_i$ are the parameters for each link i . It is noted that the NSTARIMA model covers the STARIMA and ARIMA models as special cases.

3. Case Study

The test network, which comprises 22 links in Central London, is selected for this study as shown in Figure 1 with arrows showing the directions of traffic. It has variable link lengths, ranging from 473.4m to 3.85km with an average length of 1.4km. Journey times of vehicles across the network are measured by Automatic Number Plate Recognition (ANPR) system which is operated by Transport for London (TfL). The raw journey time data are aggregated into 5-minute averages.

After discussing with TfL, data from 16 Feb 2009 to 30 Mar 2009 (43 days in total) are selected for the case study. The dataset is divided into two sets. The first 36 days are used for calibration which determines the temporal orders (p , q) and the model parameters ($\boldsymbol{\varphi}_{kh}$ and $\boldsymbol{\theta}_{kl}$). The remaining 7 days are used for validation which compares the predictions made by the calibrated model and the actual observations.

The experiment consists of three stages: identification, calibration, and validation.

- *Identification* refers to the determination of temporal orders – autoregressive (p) and moving average (q) – in the time series model.
- Given the temporal orders, the model parameters are determined in the *calibration* step by using a least square error approach. This study compares three

different time series models: ARIMA, original STARIMA and modified STARIMA.

- Finally, in *validation*, predictions made by the calibrated models are compared with the actual observations.

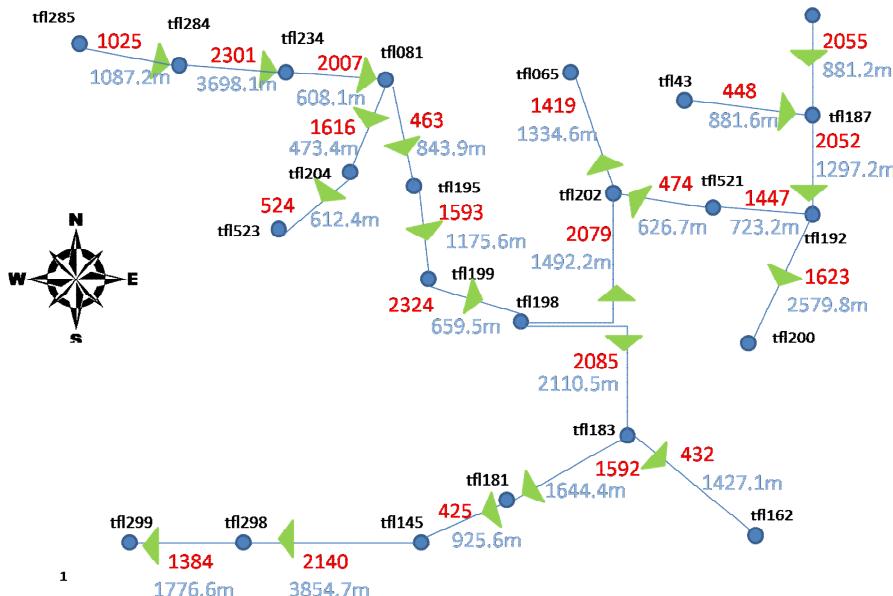


Figure 1 London test network (Cheng et al., 2011)

Figures 2 and 3 respectively show the R-square and root mean square error (RMSE) of the 22 links, which are arranged in ascending order of lengths. Results show that there is no single model dominates the others. However, if we summarise average R-square and RMSE of all links, it shows that NSTARIMA outperforms traditional STARIMA and ARIMA model.

Figure 4 shows the predictions from 12:00 to 16:00 on 30 Mar. Overall, original STARIMA has the worst average prediction results as the heterogeneity and dynamics of the urban road network cannot be well captured (Cheng et al, 2011). However, the NSTARIMA outperforms the other models on average.

4. Conclusions

This paper proposes a new space-time series model – NSTARIMA - for road traffic modelling. The proposed model is tested with journey time data obtained from the Automatic Number Plate Recognition (ANPR) system in Central London. Results show the average prediction accuracy of the NSTARIMA is better than traditional STARIMA and ARIMA model. This indicates that the new NSTARIMA can capture heterogeneity and dynamics of road traffic by modifying the original STARIMA as proposed. Given travel time is an important index for measuring transport system performance, the work reported here will contribute to the literature of traffic analysis and management.

5. Acknowledgements

The authors would like to thank Transport for London for providing the journey time data. This research is carried out under the STANDARD project, which is sponsored by the UK Engineering and Physical Sciences Research Council (EP/G023212/1).

6. References

- Cheng, T., Haworth, J., Wang, J. (2011) Some implications of complexity in network spatio-temporal autocorrelation structures, *Journal of Geographical Systems* (accepted).
- Getis, A. (2009) Spatial Weight Matrices. *Geographical Analysis*, 41(4): 404-410.
- Kamarianakis, Y. and Prastacos, P. (2005) Space-time modelling of traffic flow. *Computers and Geosciences*, 31: 119 -133.
- Lighthill, M.J., Whitham, J.B., 1955, On kinematic waves: I. Flow movement in long rivers. II. A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society (London)*, A229: 281-345.
- Min, X., Hu, J., Chen, Q. , Zhang, T. , and Zhang, Y, 2009, Short term traffic flow forecasting of urban network based on dynamic STARIMA model. *Proceedings of 12th International IEEE Conference on Intelligent Transportation Systems. (ITSC)*: 1–6.
- Min, X., Hu, J., and Zhang, Z, 2010, Urban traffic network modeling and short-term traffic flow forecasting based on GSTARIMA model. *Proceedings of 13th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 19-22 September: 1535-1540.
- Pfeifer, PE and Deutsch, SJ, 1980, A three-stage iterative procedure for space-time modelling. *Technometrics*, 22(1): 35-47.
- Richards, P.I., 1956, Shockwaves on the highway. *Operations Research* 4: 42-51.
- Yue, Y. and Gar-On Yeh, A., 2008, Spatiotemporal traffic-flow dependency and short-term traffic forecasting. *Environment and Planning B: Planning and Design*, 35: 762-771.

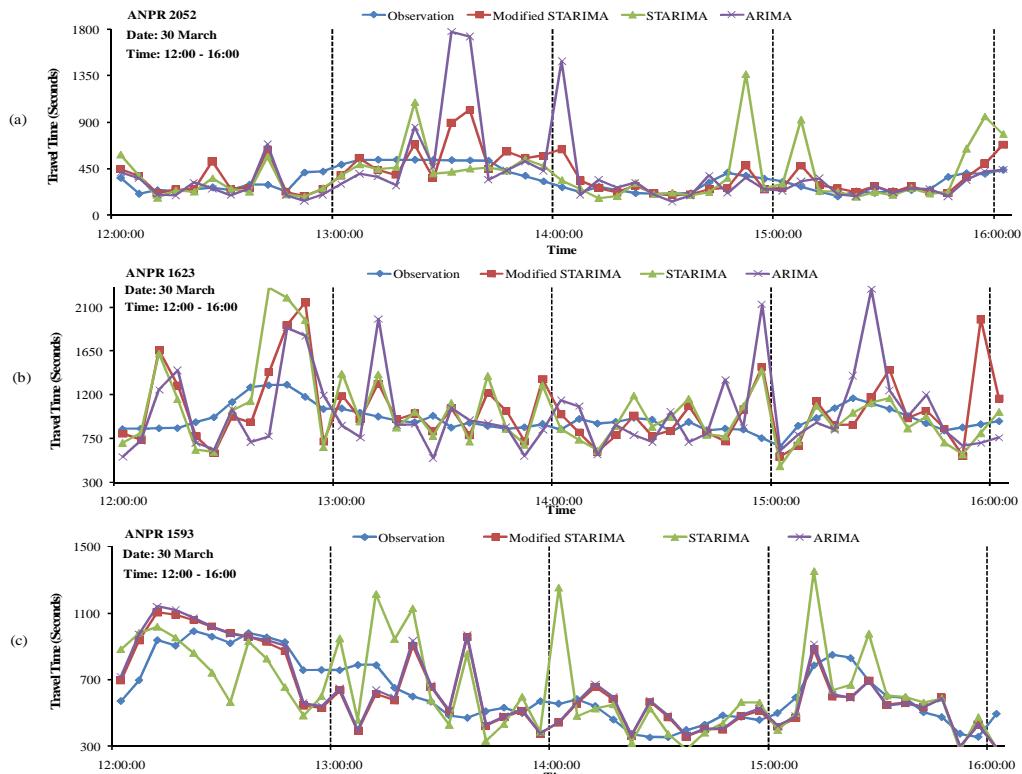


Figure 4. Prediction plots of three links 2052 (a), 1623 (b), and 1593 (c) at 12:00 - 16:00 on 30 Mar using three different models NSTARIMA, STARIMA, and ARIMA



Figure 2. R-square comparison of three different models

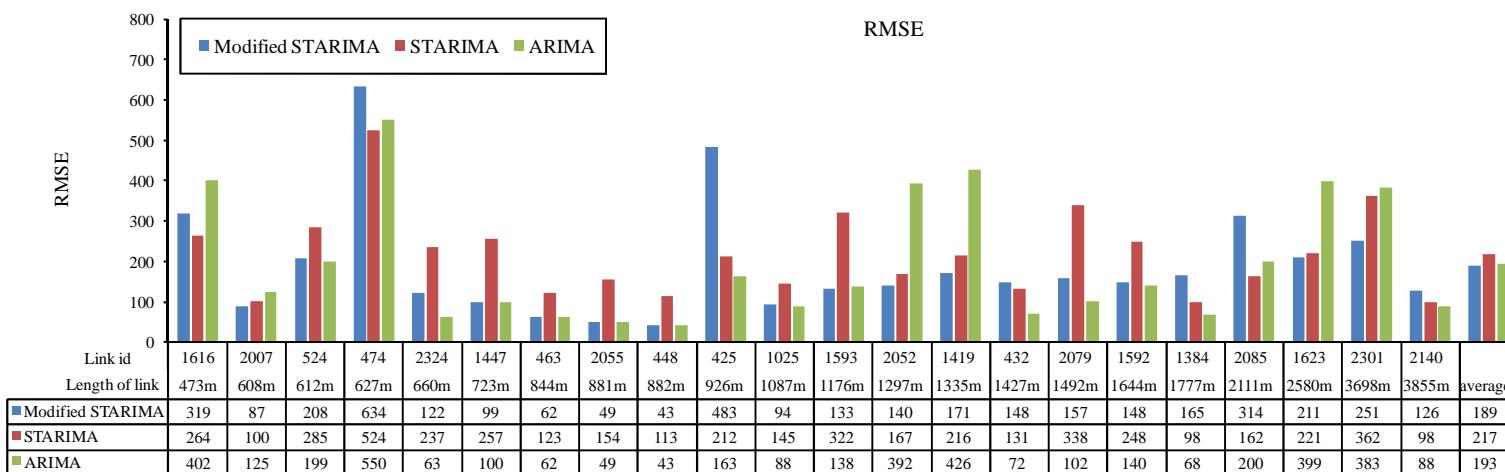


Figure 3. RMSE comparison of three different models

Road Network Analysis using Geometric Graphs of β -skeleton

T. Osaragi¹, Y. Hiraga²

¹Tokyo Institute of Technology, 2-12-1-W8-10 O-okayama, Meguro-ku, Tokyo 152-8552
 Telephone: +81-(0)3-5734-3162
 Fax: +81-(0)3-5734-2817
 Email:osaragi@mei.titech.ac.jp

²Tokyo Institute of Technology, 2-12-1-W8-10 O-okayama, Meguro-ku, Tokyo 152-8552
 Telephone: +81-(0)3-5734-3162
 Fax: +81-(0)3-5734-2817
 Email:hiraga@os.mei.titech.ac.jp

1. Introduction

There exists a numerical analysis of a road network from various viewpoints: the morphological proximity of road networks to typical geometric graphs (Tanimura and Furuyama 2002, Watanabe 2005), the efficiency of travel in a road network (Koshizuka and Kobayashi 1983), the street hierarchies from the multiple perspectives of topology and geometry (Jiang 2009) and so on. In the present study, we employ geometric graphs based on β -skeleton, which change in response to variations in parameter values, and attempt to analyze road networks by considering the morphological proximity (topological perspective) and the efficiency of travel (geometric perspective).

2. Road Network Analysis from Topological Perspective

2.1 Concept of β -skeleton

Given a spatial distribution of points p_i ($i = 1, 2, \dots, n$) in two-dimensional space, let us consider various ways of creating geometric graphs that connect the points to each other. As shown in fig. 1, let us assume that two circular arcs pass through the arbitrary points p_1 and p_2 . The size of the closed region E enclosed by the arcs (the crosshatched portions in fig. 1) varies with the parameter β (≥ 0), such that the area of E increases as β increases. Then, if some third point is included within E , then the segment with endpoints p_1 and p_2 is not an edge in the graph, whereas if no such third point is included, the graph contains this segment as an edge.

A geometric graph created according to this rule is called the β -skeleton (Wang et al. 2003, Bose et al. 2009). It is well established that the case $\beta = 0$ corresponds to the Delaunay triangulation of the set of points, $\beta = 1$ corresponds to the Gabriel graph, and $\beta = 2$ corresponds to the relative neighbourhood graph.

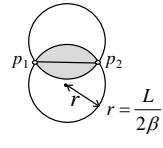
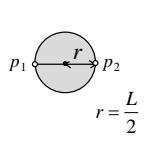
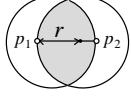
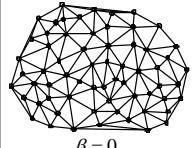
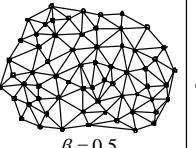
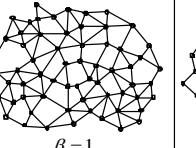
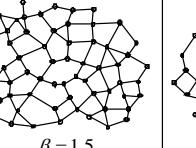
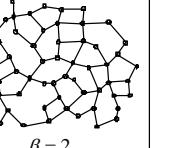
Value of β	$\beta = 0$	$0 < \beta < 1$	$\beta = 1$	$1 < \beta$	
Definition of β -skeleton					
Example of neighborhood graph	 $\beta = 0$ Delaunay triangulation	 $\beta = 0.5$	 $\beta = 1$ Gabriel graph	 $\beta = 1.5$	 $\beta = 2$ Relative neighborhood graph

Figure 1. Definition of β -skeleton.

2.2 Definition of agreement rate

Let us define an “agreement rate” as an index expressing how closely the morphology of an actual road network resembles that of a geometric graph. The set of edges making up the road network is denoted by R and that of the geometric graph is denoted by G . The number of elements in the set of edges is written as the function $n()$. Then, we define the agreement rate (C -ratio) as the number of elements in $R \cap G$ divided by the number of elements in $R \cup G$, that is, $n(R \cap G)/n(R \cup G)$.

2.3 Maximum agreement rate and value of β

The greater Tokyo metropolitan region was chosen for the study area, and subdivided into eight sub-regions shown in fig. 2.

Geometric graphs were created for various values of β , and the resulting agreement ratios with respect to the actual road network were calculated (fig. 3). The value of β yielding the maximum agreement rate is labelled β_1 . Table 1 shows the maximum agreement rate and the corresponding β_1 . As shown, the values of β_1 for the sub-regions lie between 1.0 and 1.5.

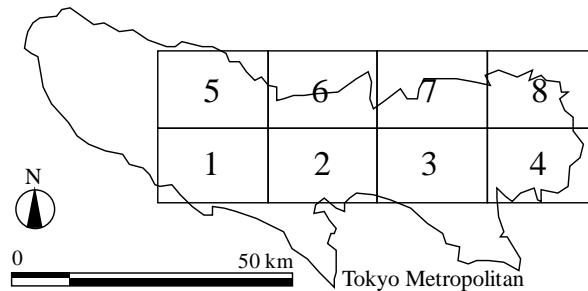
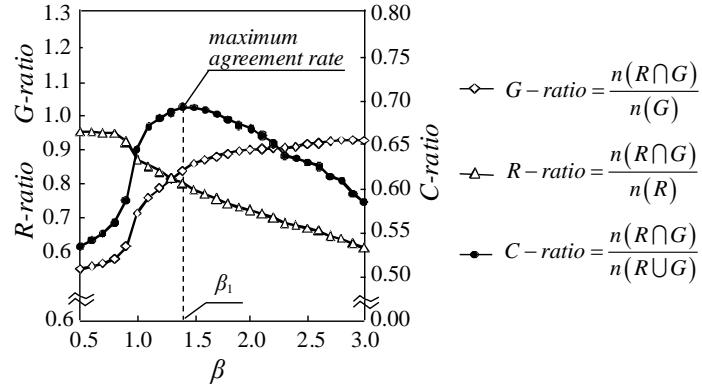


Figure 2. Study area.

Figure 3. Agreement rate as function of β (sub-region 4).

Sub-region	Maximum agreement rate	β_1
1	0.610	1.40
2	0.643	1.45
3	0.639	1.15
4	0.693	1.40
5	0.623	1.20
6	0.614	1.20
7	0.637	1.30
8	0.656	1.25

Table 1. Maximum agreement rate and the corresponding value of β_1 .

3. Road Network Analysis from Geometric Perspective

3.1 Concept of spanning ratio

The spanning ratio (SR) has been suggested as an index expressing the travel efficiency through a network (Wang et al., 2003). SR is defined as the value of the distance L between two points on the network paths divided by the Euclidian distance D between the points. In other words, the greater the values SR , the lower the travel efficiency in the network.

3.2 Spanning ratio of road network and geometric graph

The intersection points in the road networks R in the previous section were used to create Geometric graphs for various values of β ($1.0 \leq \beta \leq 2.0$). Next, two intersections at a time were extracted at random and the value of SR was calculated for that pair. The mean m and standard deviation σ were calculated for the SR of 1,000 point pairs for each graph. The results showed that m is an increasing linear function of β ($m = a\beta + b$; a and b are unknown parameters). The increase in m is due to Geometric graphs with higher values of β having lower numbers of edges, decreasing the efficiency of spatial motion in the graphs.

Also, the results showed that the value of σ grows with the value of β . The growth of σ indicates that there is high variation in the travel efficiency between point pairs, that is, that there is a large difference between the Euclidian distance and the distance in the network between point pairs. Therefore, it is preferable to conduct analysis of spatial motion in regions with low road densities on the basis of distance in the network rather than on the basis of Euclidian distance.

The mean m of SR for 1,000 point pairs was calculated for the actual road network of each sub-region. The values of β (β_2) were then inversely estimated using m by the equations ($\beta_2 = (m - b)/a$). Specifically, the values of β for the geometric graph indicating the mean values of SR equivalent to that of the actual road network were calculated. These values are shown in table 2 along with the corresponding values for parameters of regression equations. As shown, in all the sub-regions analyzed here, β_2 remains within the range 1.0 to 1.5, the same as β_1 .

Sub-region	m	a	b	R^2	β_2
1	1.224	0.217	0.913	0.993	1.440
2	1.196	0.184	0.934	0.993	1.432
3	1.155	0.184	0.946	0.981	1.146
4	1.166	0.145	0.968	0.993	1.363
5	1.184	0.213	0.906	0.998	1.310
6	1.194	0.238	0.874	0.994	1.350
7	1.178	0.202	0.914	0.989	1.310
8	1.210	0.207	0.918	0.995	1.374

Table 2. Value of β_2 for the geometric graph whose travel efficiency is equivalent to that of road network.

3.3 Relation between β_1 and β_2

Figure 4 shows relationships between the β_1 (value of β for morphological proximity) and the β_2 (value of β for similar travel efficiencies). In sub-regions 1, 5, and 6, suburban areas with low densities of roads, $\beta_1 < \beta_2$ holds. In these areas, there is a risk that using Geometric graphs (the geometric graph for β_1), which have been created on the basis of morphological proximity, will provide erroneous predictions of travel efficiency. Specifically, the travel efficiency in the actual road network is likely to be lower than that in the geometric graph created on the basis of morphological proximity. On the other hand, β_1 and β_2 are roughly similar in sub-regions 2, 3, 4, and 7, the downtown Tokyo area, where the density of roads is high.

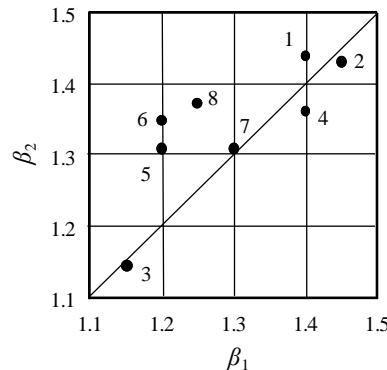


Figure 4. β_1 versus β_2 (numerals are sub-region numbers).

4. Summary and Conclusions

We carried out an analysis of a road network from each of two viewpoints, network morphology and travel efficiency, by using the concept of β -skeletons. The following findings were identified:

- (1) The value of β in a geometric graph with a maximal morphological proximity to an actual road network is in the range 1.0 to 1.5 for the networks examined here.
- (2) The agreement rate between a road network and a geometric graph is less in mountainous suburban areas or similar areas with low densities of roads.
- (3) The travel efficiency (*SR*) between two points shows more variation in suburban areas with low densities of roads; therefore, when investigating the travel efficiency between locations, the analysis must employ the distance in the network rather than the Euclidian distance between the points.
- (4) The value of β when there is high morphological proximity between a road network and a geometric graph (β_1) was nearly equal to the value of β when there is a strong similarity between the travel efficiencies in the actual network and the graph (β_2) in the central part of Tokyo. However, $\beta_1 < \beta_2$ in the Tokyo suburbs, indicating that an analyst must take account of the higher travel efficiency in the geometric graph mostly strongly resembling the actual road network than that in the actual road network itself.

In this paper, we compared the properties of geometric graphs to real road networks. This approach can be extended for the general modelling of various numerical simulations, as well as theoretical analysis on intersections which are randomly distributed following the Poisson distribution.

5. References

- Tanimura T, Furuyama M, 2002, A Study on The Rational Network Morphology Embeped in English Historic Town, *Journal of architecture, planning and environmental engineering, Transactions of AIJ*, 563:179-186.
- Watanabe D, 2005, A Study on Analysing the Road Network Pattern using Proximity Graphs, *Journal of the City Planning Institute of Japan* 40:133-138.
- Jiang B, 2009, Street hierarchies: a minority of streets account for a majority of traffic flow, *International Journal of Geographical Information Science*, 23.8: 1033-1048.
- Koshizuka T, Kobayashi J, 1983, On the Relation between Road Distance and Euclidean Distance, *City planning review*, 18:43-48.

- Bose P, Cardinal J, Collette S, Demaine ED, Palop B, Taslakian P, Zeh N, 2009, Relaxed Gabriel Graphs, *CCCG* (Vancouver):169-172.
- Wang W, Li XY, Moaveninejad K, Wang Y, Song WZ, 2003, The Spanning Ratio of β -skeletons, *CCCG* (Halifax):35-38.

The Head/tail Division Rule for Characterizing the Scaling of Geographic Space

(Extended abstract from full paper <http://arxiv.org/abs/1009.3635>)

Bin Jiang and Xintao Liu

Department of Technology and Built Environment, Division of Geomatics

University of Gävle, SE-801 76 Gävle, Sweden

Email: bin.jiang@hig.se, xintao.liu@hig.se

Scaling of geographic space refers to the fact that for a large geographic area its small constituents or units are much more common than the large ones. This paper develops a novel perspective to the scaling of geographic space using large street networks involving both cities and countryside. This paper is motivated by the belief that geographic space essentially exhibits a heavy tailed distribution rather than a normal distribution. We attempt to investigate the scaling of geographic space from the perspective of city and field blocks. This paper is further motivated by another intriguing issue, i.e., how to delineate city boundaries. Delineating city boundaries objectively is essential for many urban studies and urban administrations. Researchers and practitioners alike usually rely on the boundaries provided by census or statistical bureaus. These imposed boundaries are considered to be subjective or even arbitrary.

Three largest European countries France, Germany and UK were adopted in our study. Given a street network of an entire country, we decompose the street network into individual blocks, each of which forms a minimum ring or cycle such as city blocks and field blocks. We adopt the street networks of three largest European countries for the computation and experiments. Before the extraction of individual blocks for scaling analysis, we need to build up topological relationships, which lead to an arc-based graphs or networks.

We compute the arc-based networks to extract individual blocks in order to investigate some scaling properties. To introduce the computation, we adopt a fictive street network shown in Figure 1, which includes forty blocks and several dangling arcs that do not constitute any part of the blocks. To extract the individual blocks, we first need to set a minimum bounding box for the network in order to select an outmost arc to start traversal processes. There are two kinds of traversal processes: left traversal process and right traversal process. The left traversal process means that when comes to a node with two or more arcs, it always chooses the most left arc. On the other hand, it always chooses the most right arc for the right traversal process. Once the traversal process (starting from the outmost arc) is over, it ends up with one cycle: either a minimum cycle (which is a block) or a maximum cycle which is the outmost border. If the maximum cycle is not generated, then the program chooses a reverse direction for the traversal process until the maximum cycle is detected, and the corresponding arcs are marked with the traversal direction (left or right).

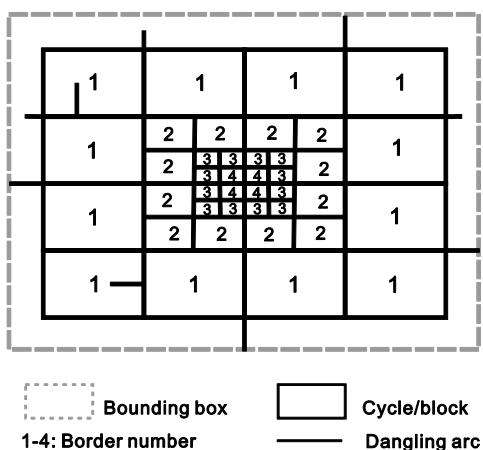


Figure 1: Illustration of the minimum cycles (or blocks) and the maximum cycle

The next step is to choose an arc on the border, and begin the traversal process along the opposite direction as previously marked, until all arcs on the border are processed. This way all the blocks on the border are detected and are assigned to border number 1. This process goes recursively for the blocks that are adjacent to the blocks with border number 1. We will get all the blocks with border number 2. The above process continues until all the blocks are exhausted and are assigned to an appropriate border number; refer to Appendix and Figure A1 for details on the algorithmic procedures. As a note on computation, it takes many hours for the server-like machine to have the process done: France and UK each about 5 hours, and Germany 63 hours. Eventually those dangling arcs are dropped out in the process of extracting the blocks. The border number is a de facto topological distance of a block far from the outmost border (Note: the border is not necessarily a country border). Every block has a border number, showing how far it is from the outmost border. The higher the border number, the farther the block is from the border, or reversely the lower the border number, the closer the block is to the border.

The block sizes demonstrate the scaling property, i.e., far more small blocks than large ones. Interestingly, we find that the mean of all the block sizes can easily separate between small and large blocks- a high percentage (e.g., 90%) of smaller ones and a low percentage (e.g., 10%) of larger ones. Statistically, the block sizes demonstrate one of the heavy tailed distributions, lognormal distribution (Figure 2). This regularity is termed as the head/tail division rule, i.e., *given a variable X , if its values x follow a heavy tailed distribution, then the mean (m) of the values can divide all the values into two parts: a high percentage in the tail, and a low percentage in the head.*

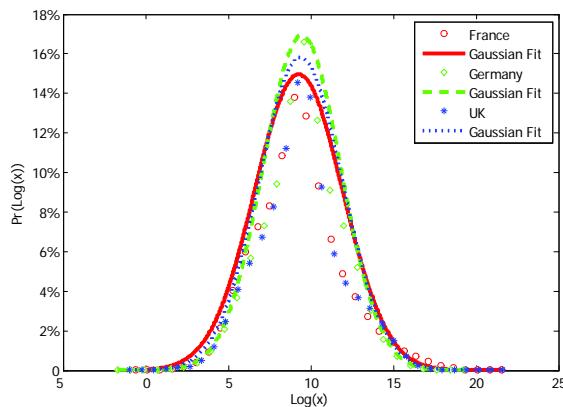


Figure 2: (Color online) Lognormal distribution of the block sizes for the three street networks

Because all blocks in one country exhibit a heavy tailed distribution (i.e., lognormal distribution), we can use the mean to divide all the blocks into smaller ones (smaller than the mean) and larger ones (larger than the mean). We then cluster the smaller blocks into individual groups. This clustering process goes like this. Starting from any smaller block whose neighboring blocks are also smaller ones, we design a program to traverse its adjacent blocks, and cluster those smaller blocks whose adjacent blocks are also smaller ones. This processing continues recursively until all the smaller ones are exhausted. We find that the sizes of the clustered groups demonstrate a heavy tailed distribution. Because of this, we then rely on the head/tail division rule to divide the groups into smaller ones (smaller than the mean) and larger ones (larger than the mean). The larger groups are de facto cities or natural cities as shown in Figure 3.

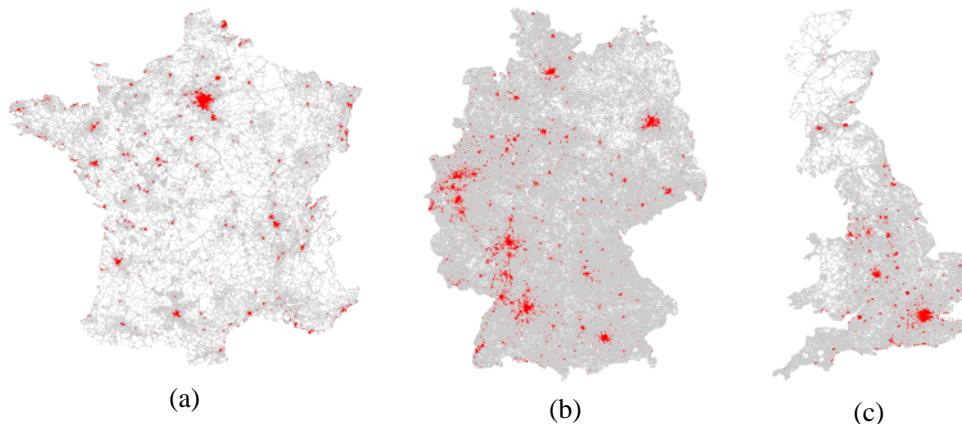


Figure 3: (Color online) All natural cities in red identified for the three networks of (a) France, (b) Germany, and (c) UK (Note: the gray background shows the extracted blocks)

We further define the concept of border number as a topological distance of a block far from the outmost border to map the center(s) of the country and the city or to characterize the scaling of geographic space (Figure 4).

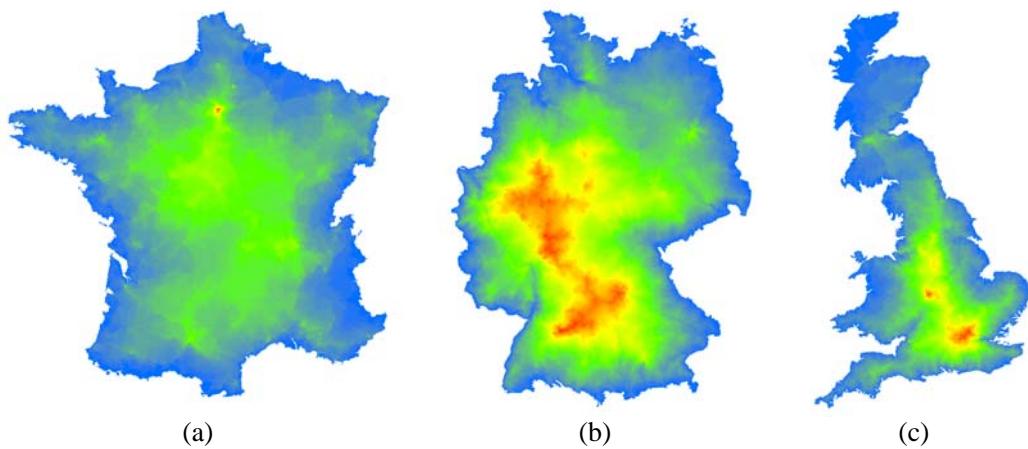


Figure 4: (Color online) Mapping the border number using a spectral color legend (Note: the higher the border number the warmer the color; red and blue represent respectively the highest and lowest border numbers)

The patterns shown in Figure 4 are illustrated from a topological perspective, which is very different from a geometric one. For example, given any country border or shape, we can partition the shape into equal sized rectangular cells (at a very fine scale, e.g., 1000 x 1000), and then compute the border number for the individual cells. Eventually, we obtain the patterns shown in Figure 5. As we can see, the centers of the countries are geometric or gravity centers that are equal distances to the corresponding edges of the borders. Essentially the country forms or shapes are viewed symmetrically. This is a distorted image of the countries, since the geometric centers are not true centers that the human minds perceive. This geometric view is the fundamental idea behind the concept of medial axis (Blum 1967), which has found a variety of applications in the real world in describing the shape of virtually all kinds of objects from the infinitely large to the infinitely small including biological entities (Leymarie and Kimia 2008). While medial axis is powerful enough in capturing a symmetric structure of a shape, it presents a distorted image of a shape as seen from Figure 5. This distortion is particularly true for France, since the true center Paris is far from the geometric or gravity center.

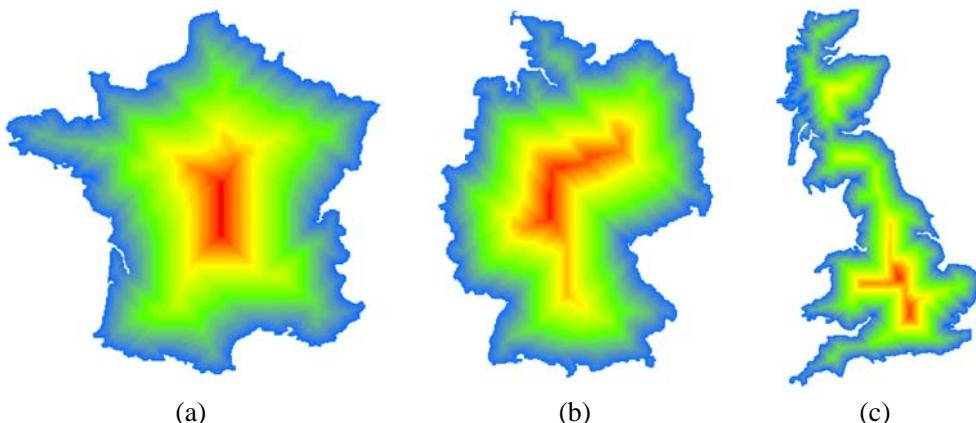


Figure 5: (Color online) Distorted images of the country based on the geometric distance far from the outmost boundaries

We add some implications to understanding the morphology of organisms. The city and field blocks can be compared to the cells of complex organisms. We believe that this kind of scaling analysis of geographic space can be applied to complex organisms and we consequently conjecture that a similar scaling structure is appeared in complex organisms like human bodies or human brains. This would reinforce our belief that cities, or geographic space in general, can be compared to a biological entity in terms of their structure and their self-organized nature in their evolution. Our future work will concentrate on the further verification of the findings and applications of the head/tail division rule.

For more details about the study, the reader is encouraged to refer to Jiang and Liu (2011) and references therein.

Jiang B. and Liu X. (2011), Scaling of geographic space from the perspective of city and field blocks and using volunteered geographic information, Preprint, arxiv.org/abs/1009.3635.

Distance dependence in the spatial structure of China aviation system: A complex network perspective

Jingyi Lin

Division of Geoinformatics, Department of Urban Planning and Environment
 Royal Institute of Technology 100 44 Stockholm, Sweden
 Email: jingyilin2008@gmail.com

1. Abstract

Aviation systems are constructed by cities and airlines, which can be analyzed under a complex network framework. In this paper, considering the unique construction mechanism of airlines, we firstly clarify that China's aviation system presents a hierarchical structure under the form of a network by detecting a scaling relation between clustering coefficient and node degree. It also indicates a more complex spatial pattern to be revealed. In the second section, in terms of the node strength and traffic flows on China's aviation network, we conclude that the distance dependence effects should be explored on different geographic distance scales. Only for medium- and long-distance travel, a gravitational law can be detected. This result paves the way as a reasonable reference for optimizing aviation systems and understanding the spatial organization of complex networks.

2. Introduction

Aviation systems, as an indispensable part for a country, have gained extensive attentions from various disciplines for a long time. In the last decade, the complex network theory introduced an innovative perspective to this field, and many real aviation systems have been studied under worldwide or national scales ([Guimerà et al 2005](#), [Liu and Zhou 2007](#), [Bagler 2008](#)). Various network characteristics have been studied and results consistently show that aviation systems present small world effects and scale-free properties. However, compared with statistical measurements, the spatial patterns of complex system have not earned enough attentions from network perspective until recently. Some researchers have began to discuss the geographical structure and distance effect of social networks and public infrastructure systems ([Lambiotte 2008](#), [Jung et al 2008](#), [Hu et al 2009](#), [Krings et al 2009](#), [Levy 2010](#)). In terms of the statistical measurements, Ravasz and Barabasi ([2003](#)) discovered some networks, whose clustering coefficient and degree satisfy $c(k) \sim k^{-1}$, should present a hierarchical architecture, while some distance-driven networks, such as power grids did not. Based on this conclusion, Liu and Zhou ([2007](#)) detected a hierarchical architecture in China's aviation network and considered that the spatial impact was negligible for it. However, it seems undeniable that any aviation system will have some spatial component. In this sense, how do aviation systems really get rid of spatial effects, if, indeed, they do? What is the underlying spatial mechanism and how does the mechanism influence the traffic flows? In this paper, we will go beyond the general topological

properties and put more focus on the unique distance dependence effect within China's aviation system from a complex network perspective.

3. Explicate the research question under a network framework

It is worth mentioning that airlines are not physically constructed as links on the ground, and this feature may endow an aviation system with distinctive network topology and spatial pattern to some extent. In this part, we will explicate the problem step by step under a complex network framework. Before this, we would like to briefly introduce the network representation pattern. Due to the hop-by-hop architecture of aviation systems, we can construct China's aviation network easily in terms of graph theory. A city is regarded as a node no matter how many airports it possesses, and a link is established as long as there is a directed flight between two cities. An aviation network with 140 nodes and 1044 edges is obtained.

As a starting point, Fig.1 depicts the correlation between node degree and clustering coefficient for 140 nodes in China aviation network. It is worth noting that the clustering coefficient is not independent of node degree but follows a scaling law in terms of degree. Such a relationship implies a hierarchical architecture for China's aviation network. It means that some vertices form lower-level communities, which are then entangled into a higher-level community. In other words, the neighbors of hub cities are not mostly linked to each other. On the contrary, geographic organizations can not display hierarchical structures due to the spatial limitations on the link length (Ravasz and Barabasi, 2003). To understand this inference, we should preliminarily introduce the unique construction of aviation systems. Compared to other ground transportation systems, such as railways or metro networks, aviation systems are less limited by geographical conditions or investment cost because physical links are not constructed. But on the other hand, every airport has its precise geographic position, so the whole aviation system is undoubtedly space-embedded. Medium-length airlines are dominant in China's aviation system, and flight flows also concentrate upon such routes. In comparison, extremely short or long routes are both rare out of cost considerations (Fig.2). In this sense, these seemly incompatible judgments actually indicate a more complex and challenged spatial pattern to be revealed. We can conjecture that China's aviation network possesses an intermediate architecture between that of a social network and geography-involved system.

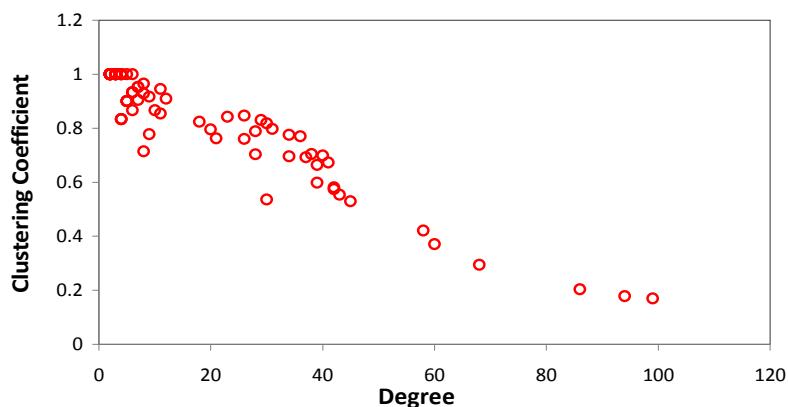


Figure 1. The scaling correlation between node degree and clustering coefficient

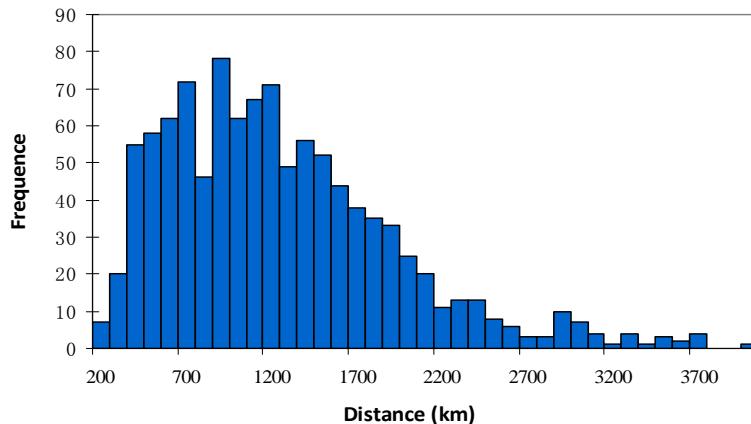


Figure 2. The histogram of geographical distance between city pairs for CAS (Most geographical distances between two cities fall into the ranges of 500-2000 km)

4. Distance dependence for different distance scales on China's aviation network

To explore the dynamics of the aviation network, we collect the weekly flight numbers between each connected city pair, and consider them as edge weights of the network. Firstly, we will depict the correlation between edge weight and spatial distance on China's aviation system in Fig.3. In order to denote the pattern more clearly, 1044 values for two variables are respectively clustered with an equidistant interval. Remarkably, the plot shows a two-regime distribution, and the critical threshold is around 500km.

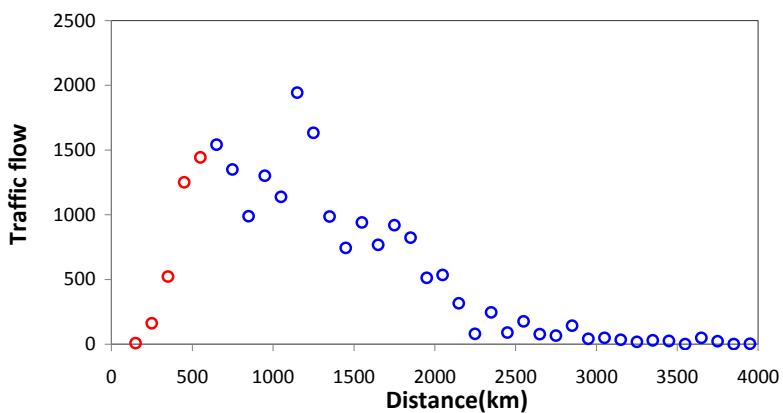


Figure 3. Dependence of the traffic flow on spatial distance (For better illustration, we set the distance interval between data points is 100 km. The result shows two-regime pattern and two different colors serve as a guide to distinguish)

Then the traffic flow is investigated in terms of node strength and spatial distance. The result is presented by Fig.4. Similarly, the correlation should be considered in terms of two separate parts, and the critical distance is 474 km. No clearly dependence effect can be found for the first part. However for the second part, a gravitational law can be detected as,

$$\frac{w_{ij}}{s_i s_j} \propto f(d_{ij}) \quad (1)$$

In which s_i is the node strength for city i , and w_{ij} is the edge weight in the aviation network, denoting the traffic flow between two cities. $f(d_{ij})$ represents the distance dependence function. In this case, it can be generalized as a scaling relation. The decay coefficient is 0.697, which is even smaller than 1. This result is reasonable. Powerful ground transportation may impose huge competitions on short-distance travel so that the advantages of aviation transportation only focus on the medium and long distances. On the other hand, due to the small-world property of the aviation network, people can transfer in some hub cities instead of constructing extremely long trips to minimize the cost.

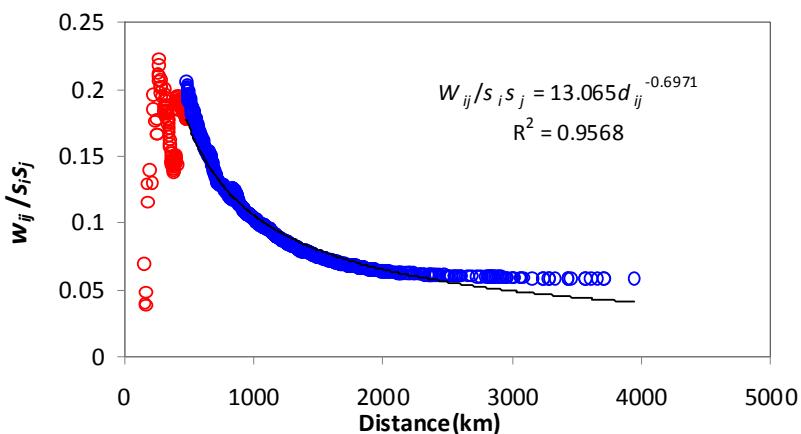


Figure 4. Regression results of dependence function based on the node strength (The red dots represent city pairs whose distances are less than 474km, while the blue dots are those with medium and long distances)

5. Conclusion

In this paper, we conclude that the unique construction of aviation system does imply a complex spatial mechanism. Although China's aviation system presents a hierarchical structure from a complex network perspective, it still displays spatial effects. This directly contradicts the proposal of Rabasz and Barabasi (2003) that geographically constrained networks would not show hierarchical organization. In part this is because geographical effects are not as simple as an elementary limitation by spatial distance. On the contrary, more complex distance dependence effects show up when edges are analyzed under separate distance ranges. We cannot find a clear

law for short-distance travels, while for the medium- and long-distance travels in the system, a gravitational law can be summarized as a distance dependence function. The dynamic simulation can be examined by introducing more detailed and real-time data in the future. This would be a significant addition to understanding how to optimize the aviation network and to exploring the geography of spatial networks.

6. Acknowledgment

The author would like to thank Dr. Bin Jiang for many valuable suggestions and discussions, and the anonymous referees for their helpful comments.

7. References

- Bagler G. 2008. Analysis of the airport network of India as a complex weighted network. *Physica A*, 387: 2972-2980.
- Guimerà R, Mossa S, Turtschi A, et al. 2005. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proc. Nati. Acad. Sci., USA*, 102: 7794-7799.
- Hu Y.Q, Wang Y.G, Di Z.R, 2009, The scaling laws of spatial structure in social networks. Preprint arXiv: 0802.0047 v2.
- Jung W-S, Wang FZ, Stanley H.E, 2008, Gravity model in the Korean highway. *Europhys.Lett*, 81, 48005.
- Krings G, Calabrese F, Ratti C and Blondel V.D, 2009, Urban gravity: a model for intercity telecommunication flows. *J.Stat.Mech*, L07003.
- Lambiotte R, Blondel V. D, Kerchove C. de, Huens E, Prieur C, Smoreda Z. and Dooren P. V, 2008, Geographical dispersal of mobile communication networks. *Physica A*, 387: 5317-5325.
- Lee K, Jung W.S, Park J.S, Choi M.Y, 2008, Statistical analysis of the Metropolitan Seoul Subway System: Network and passengers flows. *Physica A*, 387:6231-6234.
- Levy M, 2010, Scale-free human migration and the geography of social networks. *Physica A*, 389: 4913-4917.
- Liu HT, Zhou T, 2007, Empirical study of China city airline network. *Acta phys.sin.* 56(1): 106-112 (In Chinese).
- Newman M.E, 2003, Mixing patterns in networks, *Phys. Rev. E*, 67: 026126.
- Rabasz E, Barabasi A.L, 2003, Hierarchical organization in complex networks. *Phys. Rev. E*, 67: 026112.

An automated method to assess Data Completeness and Positional Accuracy of OpenStreetMap

Thomas Koukoletsos¹, Mordechai (Muki) Haklay², Claire Ellul³

^{1,2,3} University College London,
Gower Street, London, WC1E 6BT, UK
+44 20 7679 2745

¹ thomas.koukoletsos.09@ucl.ac.uk , ² m.haklay@ucl.ac.uk , ³ c.ellul@ucl.ac.uk

1. Introduction

OpenStreetMap (OSM) is an open source mapping application that is based on volunteered effort to create a free and worldwide spatial database. The increasing density, importance and acceptance of OSM increase the importance of understanding data quality, so that potential users can evaluate fitness-for-purpose. When spatial quality analysis is performed through comparison with a reference dataset, a data matching procedure is necessary for the comparison to be meaningful. This matching is usually performed manually at data preparation stage. After this, methods need to be applied to measure quality elements of completeness, positional and attribute accuracy, which should be capable of dealing with OSM's heterogeneity in accuracy, density and attribute information.

So far, research in the UK for OSM (Haklay 2010, Basiouka 2009, Ather 2009), provided valuable information on OSM for selected areas. However, all these studies include manual procedures and methods that hinder repetition of the evaluation in a different and larger area or in the future when OSM data is updated. Furthermore, they measure positional accuracy using a simplified version of the Increasing Buffer Method (Goodchild and Hunter, 1997).

We slightly modify and integrate the Increasing Buffer Method in an automated method that performs data matching and evaluates data completeness and positional accuracy of OSM data, taking into consideration heterogeneity of Volunteered Geographic Information (VGI). We apply the proposed method to the area of greater Liverpool.

2. Method

2.1. Data selection

As reference dataset, the ITN dataset of Ordnance Survey's (OS) MasterMap was used, as the most accurate official dataset covering the whole country. The method is applied in the greater area of Liverpool (1780 km^2) (fig. 1).

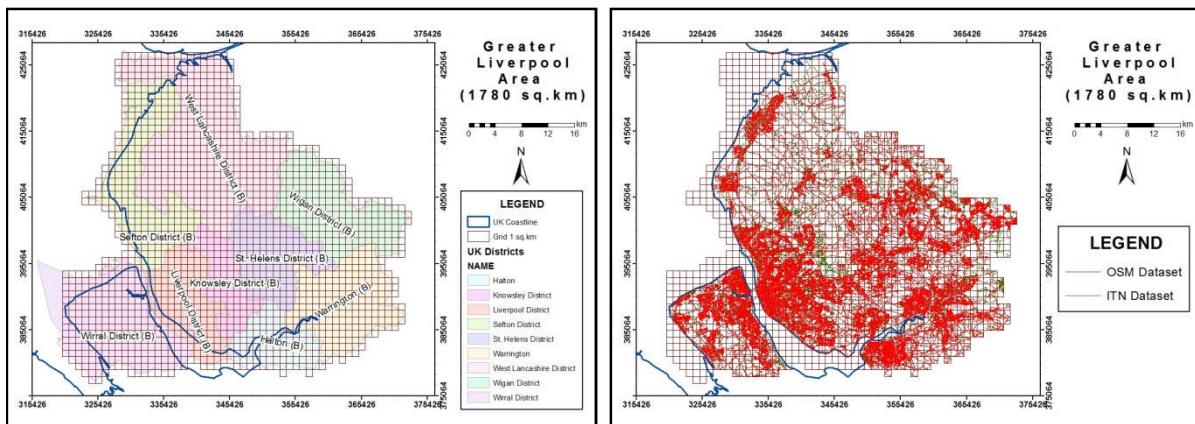


Figure 1. Area studied

2.2. Dealing with VGI heterogeneity

Data was split along the OS 1 km² National Grid and examined individually. In this way, possible variations in data density and accuracy will produce different results for each area, providing a more representative quality evaluation for VGI.

2.3. Data matching

As a first step, it is essential to remove any data that is not present in both datasets, so that any further evaluation will refer to corresponding data. The proposed data matching method combines geometric and attribute restrictions in a multi-stage approach (table 1).

Stage	Basic Unit	Constraints (in order of importance)
1	ITN Segment	Geometric (Distance,Orientation,Length)
2	ITN Segment	Attribute and geometric (name,type,Distance,Orientation)
3	ITN Segment	Attribute and geometric (name,type,Distance,Orientation)
4	ITN Segment	Geometric (Distance,Orientation)
5	OSM & ITN Feature	Geometric (Length)
6	OSM Feature	Attribute and geometric (name,type,Distance)
7	OSM Feature	Geometric and attribute (Distance,Length,type)
8	OSM & ITN Feature	Geometric (Length)

Table 1. The proposed multi-stage approach

We start by splitting features into segments. Stage 1 deals with corresponding segments based on distance, orientation and length when there is only one possible candidate. Stages 2 and 3 look for an exact and similar name matching accordingly. Stage 4 deals with segments with no name attribute. Stage 5 recomposes features and classifies them as matched or not, based on the information gathered so far. Stages 6, 7 address non matched features to solve cases not covered in previous stages. Stage 8 moves away from the tile-by-tile examination and deals with datasets as a whole, to cover matching errors in cases of corresponding features that because of their proximity to the tile border, they lie in different tiles.

A manual evaluation of data matching is performed in a randomly selected area of 80 km² (fig. 2). The lengths of the misjudged features are calculated and compared with the dataset's length for each tile and dataset. Results prove the efficiency of the data matching method (table 2).

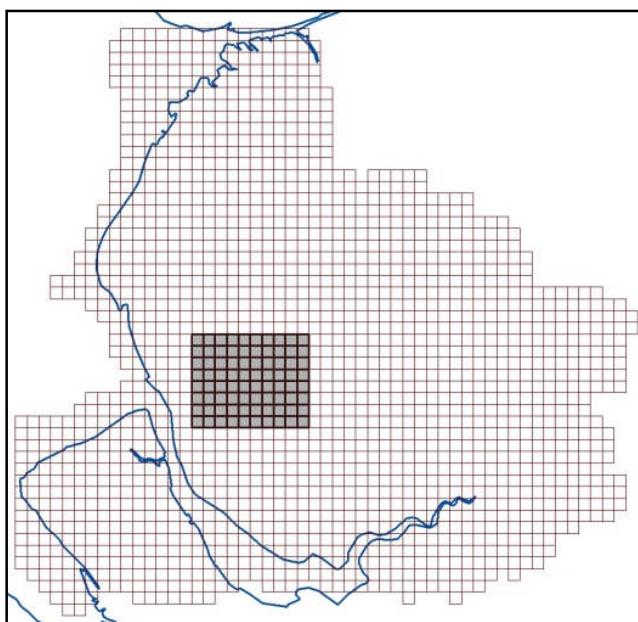


Figure 2: Data matching evaluation area

Dataset	Total length(km)	Length evaluated (km)	Missing data length (km)	Surplus data length (km)	Total matching error (km)
OSM	9042.138	694.469(7.68%)	1.575(0.23%)	2.298(0.33%)	3.873(0.56%)
ITN	10863.845	898.855(8.27%)	0.105(0.01%)	30.911(3.44%)	31.016(3.45%)

Table 2. Evaluation results: Total matching errors

2.4. Data completeness

The length of matched features is calculated and compared with the total dataset length for each tile and for each dataset, producing a data matching percentage for OSM and ITN. Table 3 provides a rough classification of the possible matching scenarios. Classification however depends on the percentages' distribution and the crisp boundaries of table 3 cannot always be appropriate for visualisation. Fuzziness due to spatial correlation may demand more classes with variable size to represent the matching percentage distribution; in the studied area for example, 90 % of the examined tiles achieved percentages above 50 % for both datasets.

Case	OSM matching percentage	ITN matching percentage	Mixed percentage	Meaning
1	High	High	High	Datasets agree with each other
2	High	Low	Low	ITN is denser
3	Low	High	Low	OSM is denser
4	Low	Low	Very Low	Datasets contain different data

Table 3. General cases of matching score for each tile

Since OSM dataset contains footpaths, steps, bridleways etc, the data matching results show the agreement rather than the completeness between the two datasets. For the results to be more representative of completeness, certain OSM road types are removed before the matching process (e.g. steps, bridleways, footpaths, tracks).

2.5. Positional accuracy

After removing data not present in both datasets, we address positional accuracy. According to Goodchild and Hunter (1997), if an increasing buffer is applied on a reference line, it will accordingly cover increasing percentages of the tested line (fig. 3). The buffer could then be considered as the accuracy of the reference dataset for the specific overlap percentage. We can either provide a buffer value to calculate the percentage, or provide a desired percentage to calculate the buffer (accuracy) using an iterative method. For the second option, which is not applied in any study so far, we use the binary search algorithm rather than the suggested formula by the authors.

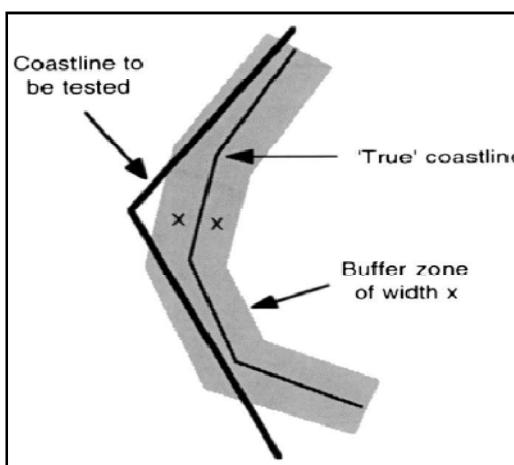


Figure 3: Increasing buffer method (from Goodchild and Hunter ,1997, p.301)

The user defines a desired overlap percentage. A first buffer of 8 m is applied on the ITN dataset and the OSM percentage falling into the buffer is calculated. If and as long as it is less than the user-defined desired overlap percentage, the buffer is doubled and calculations are repeated. When the percentage exceeds the desired one, the next buffer to be applied is half the distance between the two buffers previously used that achieved a lower and bigger percentage than the desired one correspondingly (table 4). The iteration process finishes when the percentage is within 0.1 of the desired one, or when successive buffers differ less than 0.1 m.

Tile	Iter.1	Iter.2	Iter.3	Iter.4	Iter.5	Iter.6	Iter.7
SD3612	8m- 90.9%	16m- 95.7%	12m- 93.1%	14m- 94.1%	15m- 94.8%	15.5m- 95.3%	15.25m- 95.1%

Table 4. Example of the binary search algorithm, target percentage: 95%

To decide on a suitable ‘desired percentage’, tests were carried out in an area of 25 km² in central London (where OSM is proved to be accurate by previous research). The method was applied for various percentages and the corresponding buffer values were examined. A value of 95% was chosen to be used. Above this, differences in features’ length between datasets

(due to varying data capture) as well as possible matching errors lead to unusually high buffer values.

3. Results

Fig. 4, 5 show data matching percentages for each dataset, as well as their combination. Generally ITN proves to be much more complete, as most of its data is not found in the OSM dataset (table 5).

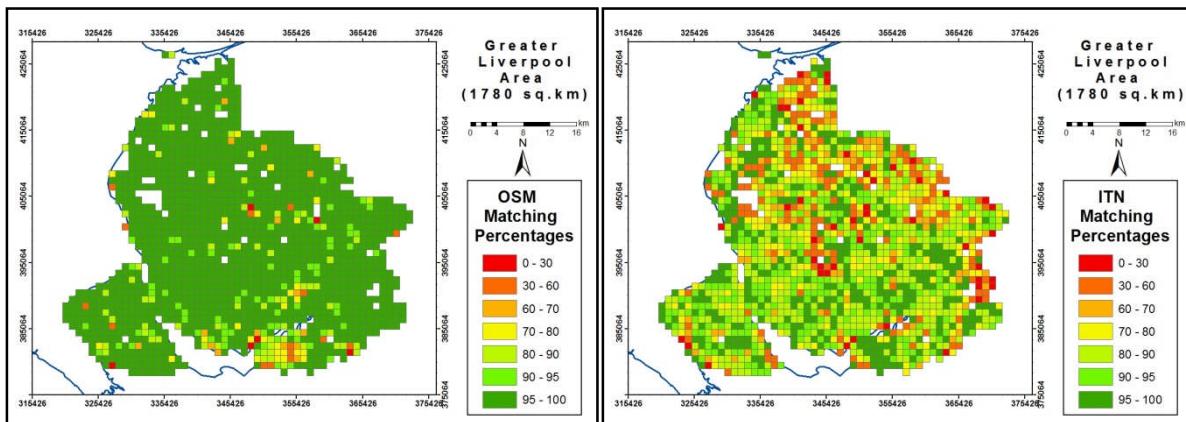


Figure 4: Data matching percentages for each dataset

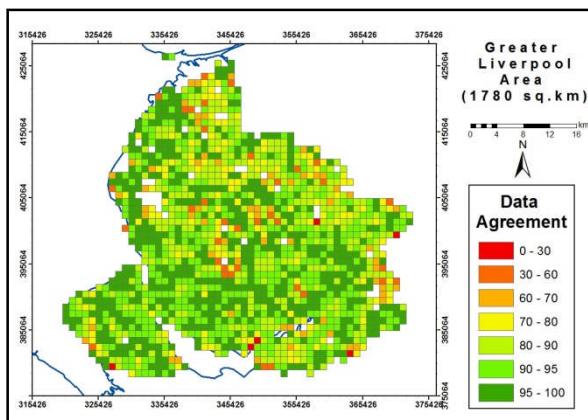


Figure 5: Data agreement between ITN and OSM

	OSM	ITN
Total length compared (km)	9175.903	10863.845
Total length matched (%)	96.62%	84.91%
Average pct matched (per tile)	96.77%	80.77%

Table 5. Data Completeness results

Fig. 6 shows the positional accuracy for 95% of OSM dataset per tile in the studied area (average accuracy 6.94 m, standard deviation 3.46 m). However, 19 tiles with buffer sizes up to 487 m had to be removed, as outliers. Due to different data capture methods, these tiles contain corresponding objects with the OSM feature extending much further than the ITN one, resulting in an increased buffer in order to reach the desired overlapping percentage, as shown in fig. 7.

The proposed method could also be used to compare other road network VGI sources and official datasets, provided that data structures include road name and road type attributes.

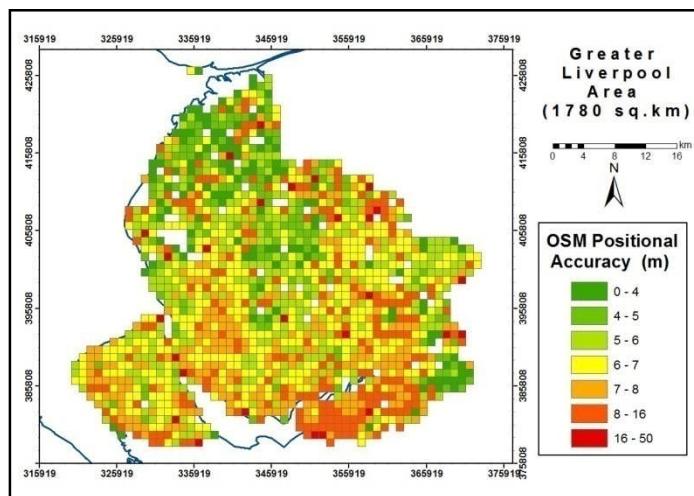


Figure 6: Positional accuracy of OSM

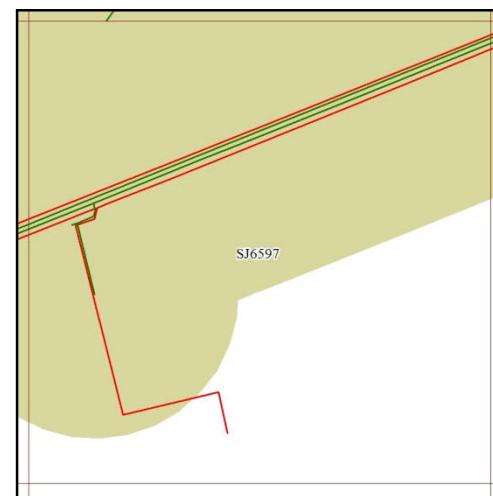


Figure 7: Buffering problems

4. Future Work

More areas need to be examined and a deeper statistical analysis of the results is necessary. Positional accuracy and data completeness results also need to be combined in search for a possible correlation. Finally, evaluation of other data quality elements needs to be integrated in the automated procedure as well.

5. Acknowledgments

We thank the Ordnance Survey and OSM for the data used in this work. All figures and tables using OS data are ©Crown Copyright/database right 2011, an Ordnance Survey/EDINA supplied service.

6. References

- Ather, A., 2009. *A Quality Analysis Of Openstreetmap Data*. MSc, University College London
- Basiouka, S., 2009. *Evaluation of the Openstreetmap Quality*. MSc, University College London
- Goodchild, M.F. and Hunter, G.J., 1997. A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, 11(3):299-306
- Haklay M (2010). How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England, In *Environment and Planning*, 37(4):682-703

Improving Global Land Cover through Crowd-sourcing and Map Integration

L. See¹, S. Fritz¹, I. McCallum¹, C. Schill², C. Perger³ and M. Obersteiner¹

¹International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, A-2361 Laxenburg, Austria

Telephone: (+43 2236) 807423

Fax: (+43 2236) 807599

Email: see@iiasa.ac.at

²University of Freiburg, Tennenbacherstr. 4, Freiburg, Germany

Telephone: (+49 761) 2038645

Fax: (+49 761) 2033701

Email: christian.schill@felis.uni-freiburg.de

³Fachhochschule Wiener Neustadt, Johannes Gutenberg-Strasse 3, A-2700 Wiener Neustadt, Austria

Telephone: (+43 2622) 89084243

Fax: (+43 2622) 8908499

Email: christoph.perger@fhwn.ac.at

1. Introduction

There are currently a large number of satellites orbiting the earth collecting vast amounts of Earth Observation (EO) data. With developments like Google Earth and Google Earth Engine, we are witnessing the democratization of EO through public access to high resolution satellite imagery via the internet. One important EO-derived product from satellites is global land cover. In the last decade, three global land cover products have been created: GLC-2000 (Fritz et al., 2003), MODIS (Friedl et al., 2002) and GlobCover (Bicheron et al., 2008). These datasets are currently used as inputs to a range of different global, regional and national scale applications, e.g. resource assessments of forest and agricultural land and inputs to global economic land use models.

There are, however, problems with land cover. A pixel-by-pixel comparison reveals areas of the world where these maps do not agree, in some cases by large amounts (Fritz and See, 2008). As a result we do not know precisely how much land is currently forested or under cultivation because the uncertainty in the estimates provided by these products is too high. This has clear implications for determining deforestation rates and how much land is available for. Users of these products also have a difficult choice, i.e. which is the best product to choose and what effect will this

choice have on a particular application? For example, Fritz et al. (2010a) have shown that comparing global land cover (GLC-2000) with the equivalent MODIS product produces large areas of disagreement when assessing the amount of agricultural land available in parts of eastern Africa. The problem with these datasets lies in their validation, as at present, there are an insufficient number of in-situ validation points, which can serve both as input data for calibration algorithms of satellite data, and to validate land cover products. The Geo-Wiki application, developed by Fritz et al. (2009), has integrated Google Earth and crowd-sourcing as a way of increasing the amount of publically-supplied in-situ validation points. The ultimate goal is to use this crowd-sourced data to create a hybrid land cover product that is better than any currently available. The aim of this paper is discuss how validation from Geo-Wiki and a rule-based map integration algorithm could be used to develop such a hybrid product.

2. The Geo-Wiki Land Cover Crowd-sourcing Application

The Geo-Wiki Project (www.geo-wiki.org) was developed to encourage a global network of volunteers to help improve the quality of global land cover maps through crowd-sourcing (Fritz et al., 2009). Geo-Wiki overlays the GLC-2000, MODIS and GlobCover onto Google Earth as well as maps of where these different land cover products disagree. Volunteers can choose any area of land on the earth or an area of high disagreement. Geo-Wiki shows them where the pixels from each land cover product overlap and the land cover types as shown in fig. 1. The light blue rectangle is GLC-2000 and has the lowest resolution of 1km. The dark blue square is one pixel from the MODIS land cover product while the red square is GlobCover at the highest resolution of 300m. Volunteers are then asked to determine whether the land cover maps at that point agree with what they see based on Google Earth. Their input is recorded in a database, along with any photos they upload. At present there are 300 users registered on the system who have contributed more than 15,000 validation points.

3. Development of a Hybrid Product through Map Integration

Fritz et al. (2010b) have developed a methodology for combining five different land cover maps to create a cropland or forest extent using expert knowledge and national and sub-national statistics. However, to create a global land cover map is more complicated because the legends of the different land cover products do not match. An aggregated and simplified legend to which the different land cover products can be

matched directly must first be created (e.g. Herold et al., 2008). This is already available on Geo-Wiki as a simplified legend with 11 classes.

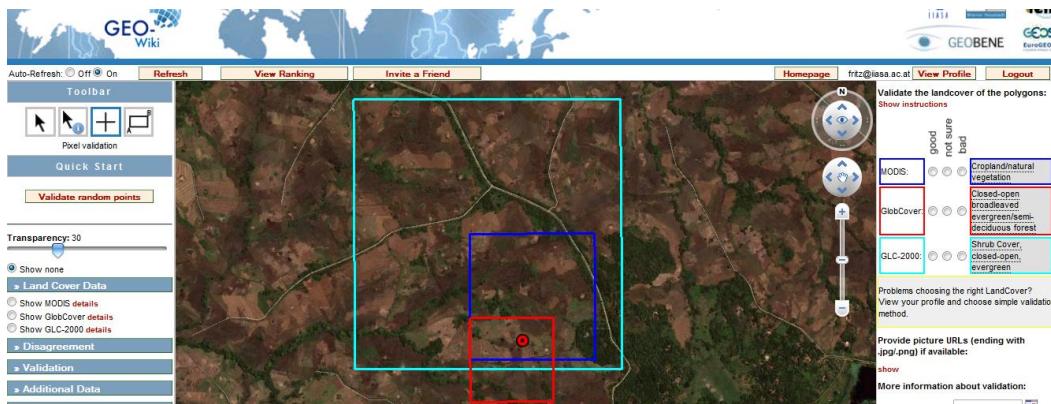


Figure 1. Illustration of Geo-wiki.org for improving land cover information

To use the validation data for creating a hybrid map, there must be sufficient confidence in the data before they are used in the new hybrid product. To gain confidence, people are directed to the same validation site so that a frequency distribution can be derived and agreement can be reached on which product is better. Once a certain threshold had been achieved, the point qualifies to be a validation point to be used in the hybrid map generation. Those validation points where the confidence is low will not be used in the hybrid map production

To create a hybrid map, a rule-based system is currently under development. For each pixel, the system will query whether there is disagreement between the land cover products as follows:

- Where no disagreement exists, the hybrid land cover map will be assigned the class from the aggregated legend unless validation data exist at that pixel which disagrees with the land cover products. If there is sufficient confidence in the validation data, then these data will be used to correct the information from the land cover products in the new hybrid product.
- Where disagreement between land cover products exists, the validation data from Geo-Wiki will be used. Where validation data of a sufficient confidence are available at that pixel, this will be used to assign the land cover class. Where no validation data exist, we will employ a search algorithm to determine if the same corrections have been applied using validation points within a certain radius,

which will be used to assign the land cover class. This will also increase our confidence in the validation sites.

This rule-based system will be implemented in Geo-Wiki in the next few months and an example of a hybrid map and the issues that have arisen from the implementation will be presented at the conference.

4. Ongoing Developments and Further Research

There are several ongoing developments with Geo-Wiki to improve the volume and coverage of data collected through the website. The first addresses the problem that there is currently little incentive for volunteers to willingly validate global land cover. One method of providing this incentive will be to develop a game that encourages users to play while simultaneously providing land cover validation information. An Austrian Funding Agency project called LandSpotting, which will begin in Feb 2011, addresses the creation and implementation of such a game. A second development is from the research community. A workshop on land cover validation (sponsored by the International Livestock Research Institute (ILRI) and to be held in June 2011 at IIASA) will bring together land cover and validation experts to discuss methods for creating hybrid products and the sharing of land cover products and validation data via Geo-Wiki. This will increase the size of the validation database and particularly the coverage across Africa and south-east Asia. The progress on these advances will also be presented at the conference.

5. Acknowledgements

This research was supported by the European Community's Framework Programme via the Project EuroGEOSS (No. 226487) and by the Austrian Research Funding Agency (FFG) via the Project LandSpotting. (No. 828332).

6. References

- Bicheron P, Defourny P, Brockman C, Schouten L, Vancutsem C, Huc M, Bontemps S, Leroy M, Achard F, Herold M, Ranera F and Arino O, 2008, GLOBCOVER, http://ionia1.esrin.esa.int/docs/GLOBCOVER_Products_Description_Validation_Report_I2.1.pdf
- Friedl MA, McIver DK, Hodges JCF, Zhang XY, Muchoney D, Strahler AH, Woodcock CE, Gopal S, Schneider A, Cooper A, Baccini A, Gao F, Schaaf C, 2002, Global land cover mapping from MODIS: algorithms and early results. *Remote Sensing of Environment*, 83:287-302.
- Fritz, S., Bartholomé, E., Belward, A., Hartley, A., Stibig H.J., Eva, H., Mayaux, P., Bartalev, S., Latifovic, R., Kolmert, S., Roy, P., Agrawal, S., Bingfang, W., Wenting, X., Ledwith, M., Pekel, F.J., Giri, C., Mücher, S., de Badts, E., Tateishi, R., Champeaux, J-L., Defourny, P., 2003.

- Harmonisation, mosaicing and production of the Global Land Cover 2000 database (Beta Version)*, Luxembourg: Office for Official Publications of the European Communities, EUR 20849 EN, 41 pp., ISBN 92-894-6332-5*.
- Fritz S, and See L, 2008, Quantifying uncertainty and spatial disagreement in the comparison of Global Land Cover for different applications, *Global Change Biology*, 14:1-23.
- Fritz S, McCallum I, Schill C, Perger C, Grillmayer R, Achard F, Kraxner F and Obersteiner M, 2009, Geo-Wiki.Org: The use of crowd-sourcing to improve global land cover. *Remote Sensing*, 1(3):345-354.
- Fritz S, See LM and Rembold F, 2010a, Comparison of global and regional land cover maps with statistical information for the agricultural domain in Africa. *International Journal of Remote Sensing*, 25(7-8):1527-1532.
- Fritz S, You L, Bun A, See LM, McCallum I, Liu J, Hansen M and Obersteiner M, 2010b, Cropland for Sub-Saharan Africa: A synergistic approach using five land cover datasets. *Geophysical Research Letters*, doi:10.1029/2010GL046231, in press.
- Herold M, Mayaux P, Woodcock CE, Baccini A and Schmullius C (2008) Some challenges in global land cover mapping: an assessment of agreement and accuracy in existing 1 km datasets. *Remote Sensing of Environment*, 112: 2538–2556.

Geospatial Service Web – Cyber Infrastructure for Service-oriented Geospatial Science

Jianya Gong, Huayi Wu, Wenxiu Gao, Peng Yue, Xinyan Zhu

State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, China
geogjy@163.net; wuhuayi@lmars.whu.edu; wxgao@lmars.whu.edu.cn;
geopyue@gmail.com; geozxy@263.net

1 Introduction

Geographic Information Systems (GIS) have evolved from desktop GIS to Web-based GIS, and then to geospatial information Web Services, so what comes next? The advancement of Web Service technologies redirects the focus of geospatial services from data services providing information and static knowledge to geospatial processing services and their combinations (Forster, 2005; Hey and Trefethen, 2005). Geospatial information services and their combinations produce value-added information and derived knowledge. Meanwhile, emerging semantic tools facilitate interaction between service and service, and between users and services (Bemers-Lee, et al., 2001; Yue, et al., 2007).

This paper envisions the next step of GIS as a widely-connected, interoperable and semantically supported Geospatial Service Web (GSW), a future framework for geospatial information technology. Data, information and knowledge services are essential bricks, but the GSW features geospatial processing services and their combinations that collaborate to simulate, deduce and predict geographic phenomena, processes and results. In the data source side, GSW extends its antenna from static databases to real-time data collecting sensors. On the application side, GSW extends capacity from mere visualization to real-time model automation for decision making. GSW embraces an intelligent mechanism for auto-gestation by combining geospatial processing services and deploying, registering these combinations in repository geographic models. This paper introduces our pioneering thoughts about the rationale, conception, framework, technologies and standards for building the GSW. A prototype system was developed to demonstrate and illustrate the initial shape of GSW.

2 Concepts of Geospatial Service Web

The Geospatial Service Web (GSW) is a virtual geospatial infrastructure which integrates various geospatial-related resources. GSW unifies the functions of a geospatial acquisition system, a data transformation system, distributed spatial data collection, a high-capability server system, a large volume storage system, remote sensing and a GIS system. These functions are implemented by web services and communicated through the standardized protocols of the Internet. Fig. 13.1 shows the logical components and structures of GSW. With GSW, users are able to effectively describe, organize, manage, manipulate, interchange, search and release the geospatial-related resources.

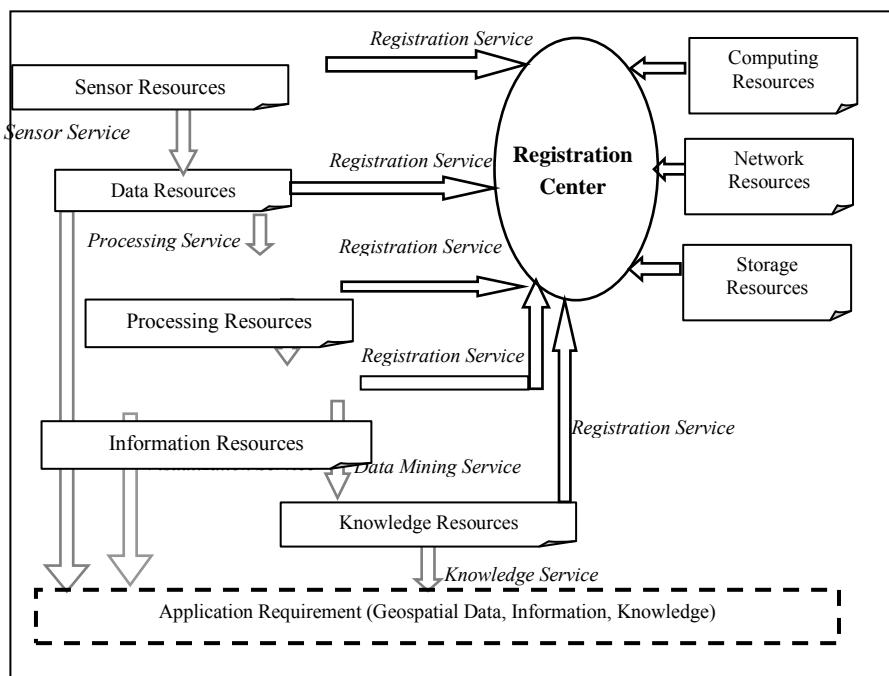


Fig. 1. The procedure of geospatial service from sensor to knowledge

A registration center is the core of the virtual infrastructure. It accepts and archives the registration information of all resources. Thus, all of the resources can be acquired and accessed through the center. Sensor resources incorporate all kinds of sensors for data acquisition including space-borne sensors, air-borne sensors and handheld devices.

The original Earth Observation data acquired by sensors comprises a huge geospatial data resource. Processing resources are the collections of theoretical models, process models and behavior models which are necessary to pre-process, transform, compact, project, generalize, visualize geospatial data for a specific application context. In many cases, these models can be composed into an orderly integrated model to implement a complex function and further to derive potential information and knowledge from datasets. Finally, all of the resources are combined by the associated web services and provider services as a whole for geospatial-related applications. In addition to the geospatial domain resources , the general resources must be recognized and considered, including computing, network and storage are also indispensable for GSW, but they are beyond the theme of this chapter.

In conclusion, the mission of GSW is to:

- acquire global geospatial data for all seasons, all days and in all directions by all kinds of sensors on satellites, aircraft and on the surface.
- chain the whole process seamlessly from sensors to application services by unified information networks, including satellite communicate, data relay network and wired or wireless computer communication networks.
- register sensors, computing resources, storage resources, internet resources, geospatial data and manipulate software, geospatial knowledge on the Internet, and process geospatial data online quantitatively, automatically, intelligently and in real-time.
- provide geospatial services, compose virtual service chains and transmit user-required information by the most effective and efficient means.

3 Framework of Geospatial Service Web

Based on the concept of a geospatial service web described in the previous section, Fig.2 illustrates the corresponding framework including five basic components: a geospatial resource component, a geospatial service component, a geospatial service applica-

tion component, a geospatial service security component, and a geospatial service standard component.

Geospatial resource components are the cornerstone that involves almost all applicable resources in the digital environment. The Geospatial service component is the bridge between geospatial resources and geospatial service applications, and provides functions with services such as accessing, processing, transporting, and visualizing data. Some geospatial services can be combined into a specific application system, or be developed into individual application tools, or be used to build a visualization environment for geospatial data. Geospatial standards are fundamental supports for communications and data sharing between heterogeneous components in the framework. The geospatial service security component protects geospatial data and services from illegal usage or attack.

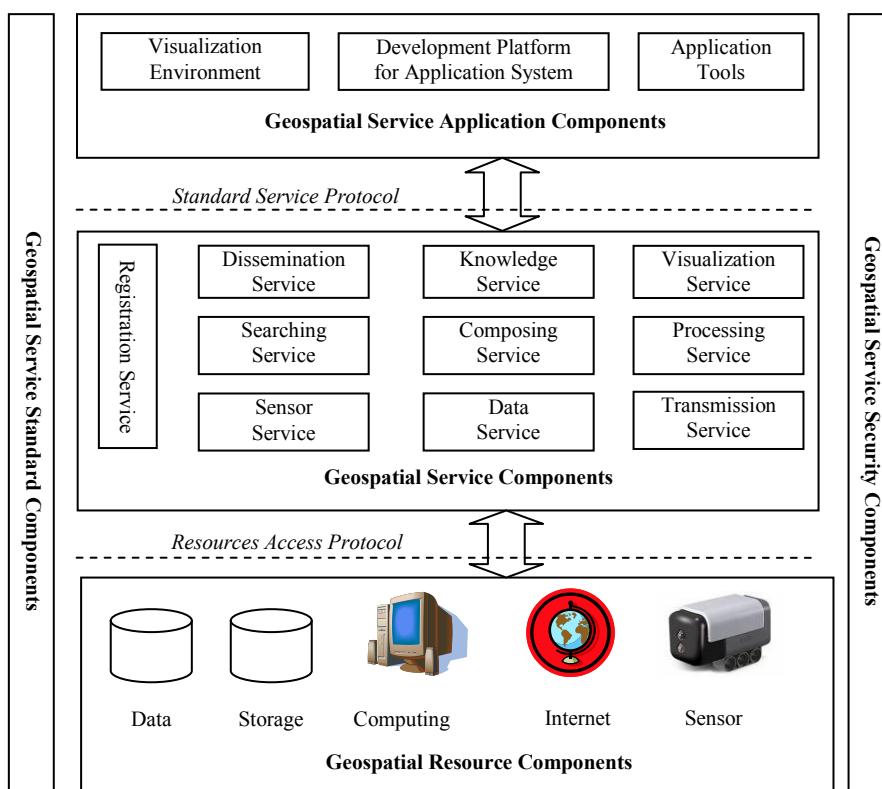


Fig.2. Framework of Geospatial Service Web

4. Implementation of Geospatial Service Web

Fig. 3 shows the system architecture towards the implementation of GSW. The implementation of GSW will integrate and communicate different types of space-earth data acquired by using various earth observation technologies such as satellite, airplane and in-situ observation. The application areas of GSW are diverse, such as meteorology, agriculture, forestry, transportation and digital city. The GSW is built upon open, consensus-based standards (i.e. specifications for geospatial information resources in Fig. 3) that will allow the “plug-and-play” of community-developed, standard-compliant components and services. The following paragraphs provide detail descriptions of the major components for the implementation of GSW.

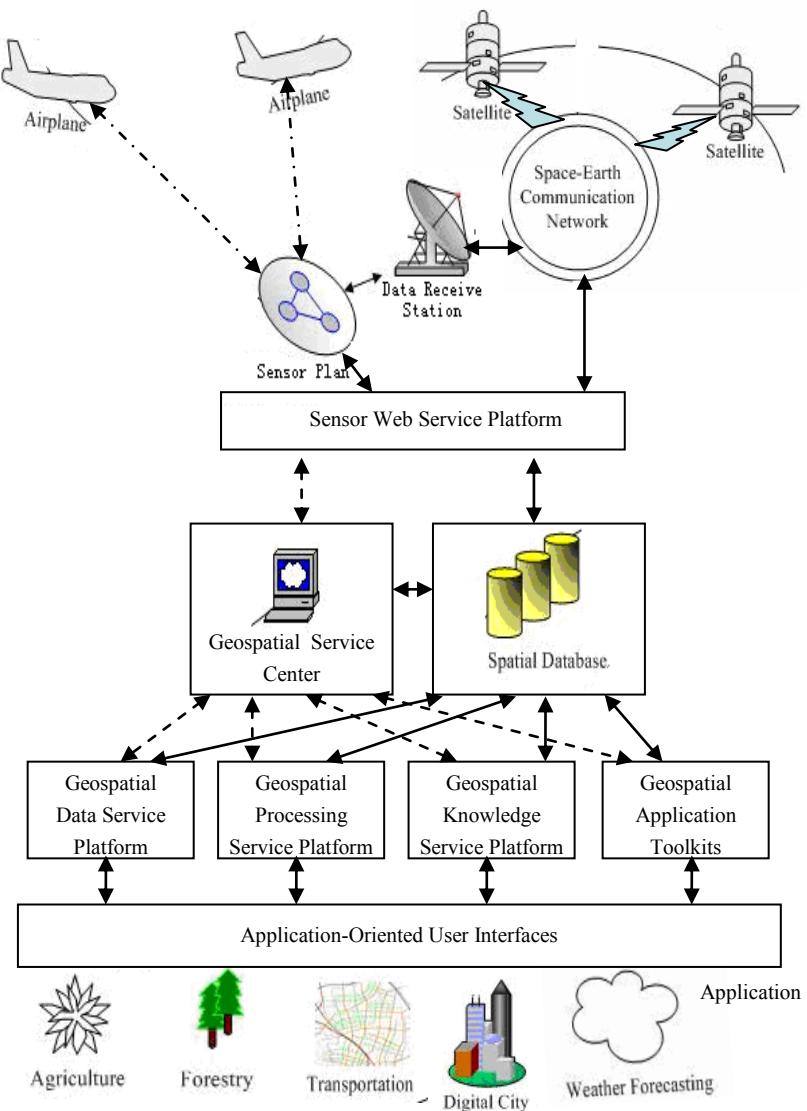


Fig. 3 System architecture towards the implementation of GSW

Fig.4 shows the user interface of a prototype for Geospatial Service Center based on the architecture described in the section above. The left tree lists the available services registered in the center, e.g. geospatial data services, processing services, and map portrayal services. This center accepts the registration of data type, data instance, service type, service instance and map symbols. Users can view the information for a service by clicking it on the tree.

The prototype provides an environment for composing service chains as application-specific workflows. Fig. 5 illustrates an example of a service chain composed for flood submergence analysis. From the palette on the left column, the useful services are chosen and dragged to the right area. In turn, these services are chained in a logical order according to the specific requirements for flood submergence analysis. Finally, an abstract chain will be built and stored as an expert workflow. The abstract chain will be transformed into a BPEL service chain and executed by the engine of service chain. This kind of abstract chain can be reused and adjusted for different applications. Fig. 6 presents the result of the service chain for flood submergence analysis. The dark part in the center of the map is the area submerged by flood.

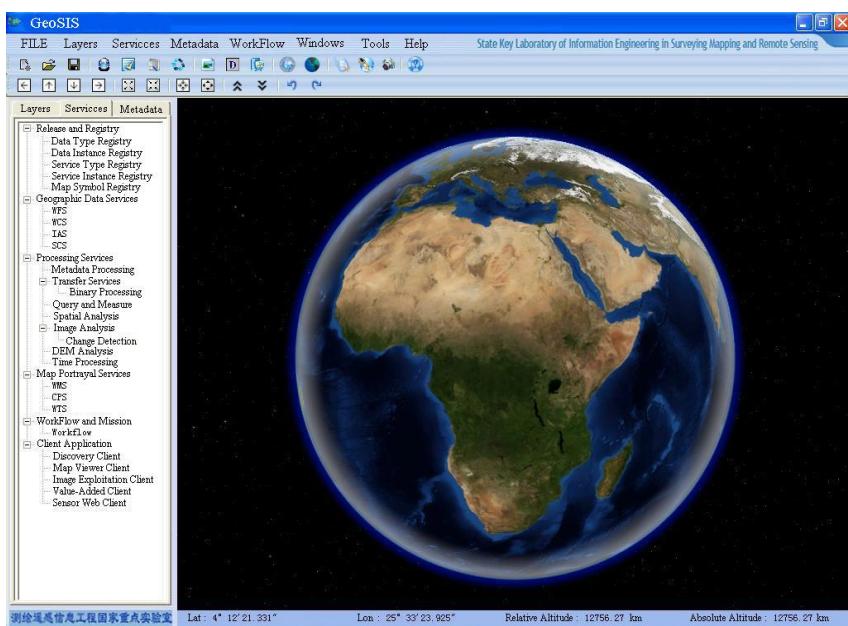


Fig.4 The user interface of a Geospatial Service Center prototype

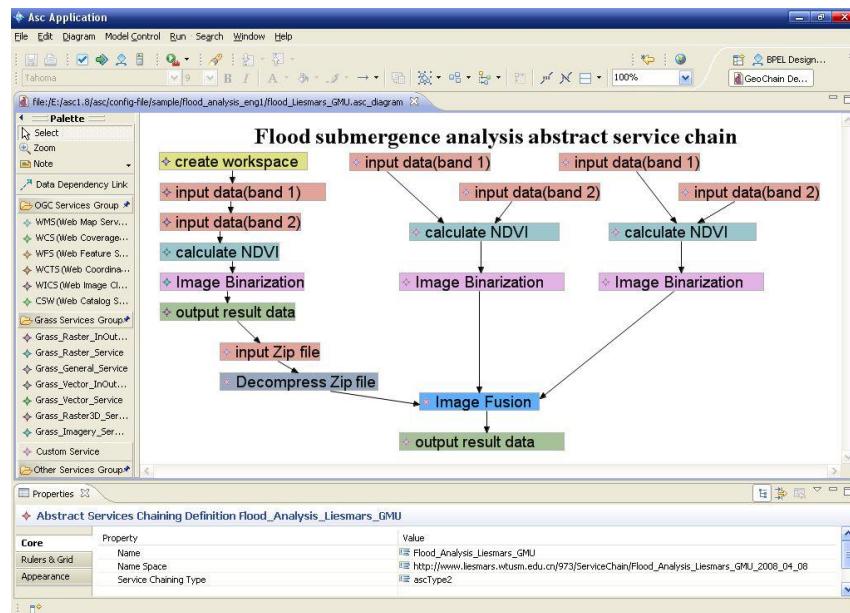


Fig. 5 A service chain for flood submergence analysis



Fig. 6 The result of the service chain for flood submergence analysis

5. Conclusion

With Web Service technology advancements, the services deployed and distributed on the Internet are not only data services comprised of implied information and static

knowledge, but also geospatial processing services, other related geospatial services and their combinations, which are generating user-induced information and dynamically growing knowledge. Semantic tools meanwhile, are aiding the interaction between service and service, and between users and services. Thus, a new concept, Geospatial Service Web (GSW), as a basic umbrella framework is the future of geospatial information technology, as proposed in this paper. Data, information and knowledge services are still the essential bricks of this web, however the focus of the GSW web is on atomic processing and processing combination services that collaborate to simulate, deduce and predict geographic phenomena, processes and results. This new concept will also expand the reach of geospatial connectivity to embrace both the spatial and temporal dimensions. In the data source rim, this web extends its antenna from static database to all data collecting sensors from satellite-based, airborne to ground and mobile. At the application end, this web realizes applications from visualization to real-time automatic decision support. What is more, this web will have a mechanism of rule-based auto-gestation. New combinations of geospatial processing services can be deployed, registered and included in the repository geographic models. As a pioneer effort, this chapter systematically preaches this new thought, outlines the concepts, framework, technologies and standards of GSW.

6. Acknowledgements

This study was supported by the Natural Science Foundation (40971211 and 41023001).

References

- Berners-Lee T, Hendler J, and Lassila O (2001). The Semantic Web. *Scientific American*, 284(5):34-43.
- Forster I (2005). Service-Oriented Science. *Science*. 308(5723):814-817.
- Hey T and Trefethen A E (2005). Cyberinfrastructure for e-Science. *Science*. 308(5723): 817-821.
- Yue P, Di L, Yang W, Yu G and Zhao P (2007). Semantics-based automatic composition of geospatial Web services chains. *Computers & Geosciences*, 33(5): 649-665.

A Universal Framework for Parallel Processing Massive Spatial Data using a Split-and-Merge Paradigm

Xuefeng Guan, Huayi Wu

State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing,
Wuhan University, 129 Luoyu Road, Wuhan 430079, P. R. China

1. Introduction

The improvement of modern remote sensing technology, e.g. LiDAR (Light Detection and Ranging), has resulted in the explosive growth of spatial datasets [1]. As the size of LiDAR point clouds increase from gigabytes to terabytes, even to petabytes, it is impossible to process them anymore within a single desktop personal computer (PC). At the same time, the multicore-enabled Central Processing Unit (CPU) are becoming ubiquitous from the single desktop PC to clusters[2]; while the costs to build a powerful computing cluster are getting lower and lower. Therefore, it is natural and necessary that typical users employ high performance clusters (HPC) to efficiently process massive LiDAR point clouds [3, 4].

Inherently different from classical compute-intensive applications, the kernel of processing massive LiDAR point cloud is not complex but rather simple; nevertheless it still requires extensive computing resources and lengthy execution time. Hence, this type of application can be characterized as a data-intensive application [5].

Data-intensive applications involve heavy I/O operations. The decomposition, scheduling, load-balance are much different from traditional compute-intensive applications. Thus, porting such a data-intensive application into a HPC context is a challenging task. This paper proposes a universal parallel framework in a HPC environment to facilitate this transition. The framework supports a Split-and-Merge programming paradigm for users/programmers, exemplified by processing massive LiDAR point clouds. Under this paradigm, our framework can automatically parallelize and schedule user's tasks.

2. A universal parallel framework

2.1 A Split-and-Merge paradigm

After evaluating many algorithms available and in use for processing LiDAR point clouds, a common characteristic of all of them is data locality. Data locality means that the kernel of these algorithms only involves the proximity data of input element. This characteristic is the basis for processing LiDAR point cloud in a split-and-merge paradigm. In this paradigm, the entire LiDAR point cloud is first decomposed into many discrete blocks; then these blocks are individually processed

by the original algorithm; and finally the intermediate results are merged into the actual output.

The complete execution graph can be divided into two categories according to the decomposition and merge patterns: two-level n-ary tree, and n-level binary tree, illustrated by Fig.1. In the first type, all the intermediate results are merged in a whole; in the second type two adjacent intermediate results are merged hierarchically.

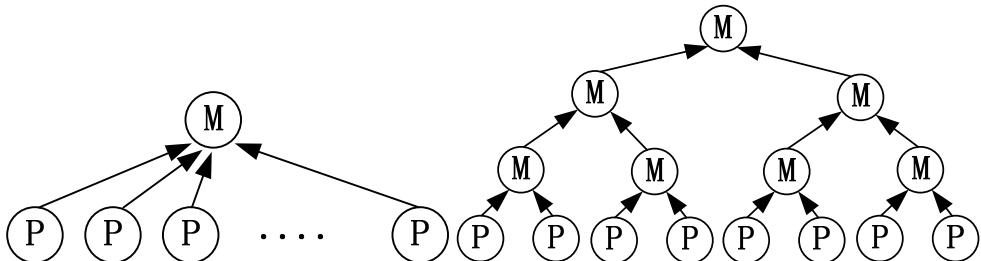


Fig.1 two types of Split-and-Merge paradigms
(Left: two-level n-ary tree; right: n-level binary tree)

For a specific LiDAR algorithm, users/programmers only focus on actual implementation of two processing steps: Split and Merge. After users implement these two steps and choose the execution pattern, the framework will automatically generate a collection of scripts to enclose these individual tasks.

2.2 Data decomposition and block organization

Discrete decomposition of the LiDAR point cloud is the prerequisite for the Split and Merge steps. A kd-tree (short for k-dimensional tree) based decomposition schema was designed to carry out this decomposition (here k is 2). The bounded extent of the entire LiDAR point cloud is divided into $n*m$ rectangle blocks which represents the parallel granularity. A 2D kd-tree is recursively constructed to cover these point blocks. Each point block is then mapped to the corresponding leaf node of the constructed kd-tree. All the points are assigned to point blocks later. Some leaf nodes with no available points are marked as null. The internal nodes also represent the intermediate results during the whole execution process. The kd-tree based decomposition schema can be illustrated by Fig.2.

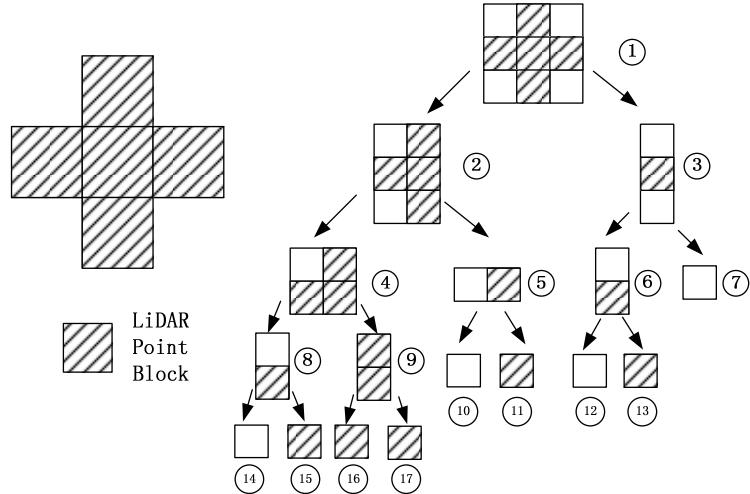


Fig.2 The kd-tree based decomposition schema

The framework also has an index module to organize the discrete blocks (Data Scheduler). This module also stores the current status of block distribution among cluster nodes and provides the real-time information for the later data-aware scheduling module (Task Scheduler). These two modules will be elaborated in section 2.3.

2.3 A universal parallel framework

Our universal parallel framework is built on a typical SMP (Symmetric Multiprocessor) cluster, illustrated in Fig.3. In a SMP cluster, each node is equipped with two or more symmetric processors. Each processor is also multicore-enabled. Thus, there are two levels of parallel computing resources available for one SMP cluster: inter-node and inner-node. A customized Torque [6] runs on the master node. A special data-aware scheduling strategy was designed for the custom Torque to schedule user's decomposed tasks.

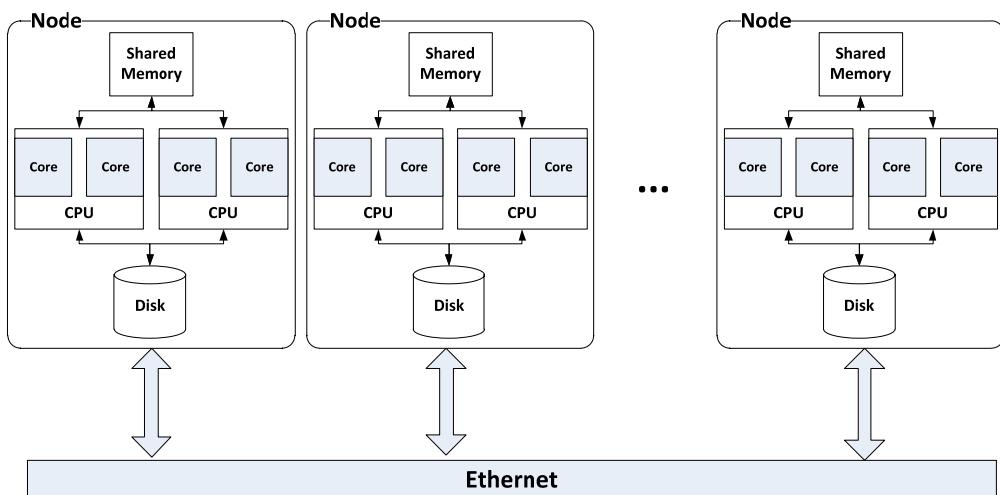


Fig.3 the illustration of one typical SMP cluster

Fig.4 shows the basic data flow for our proposed framework. The framework is

controlled by the task/data scheduler. The task scheduler creates and manages the processes that run all Split and Merge tasks. The data scheduler manages data input for task execution.

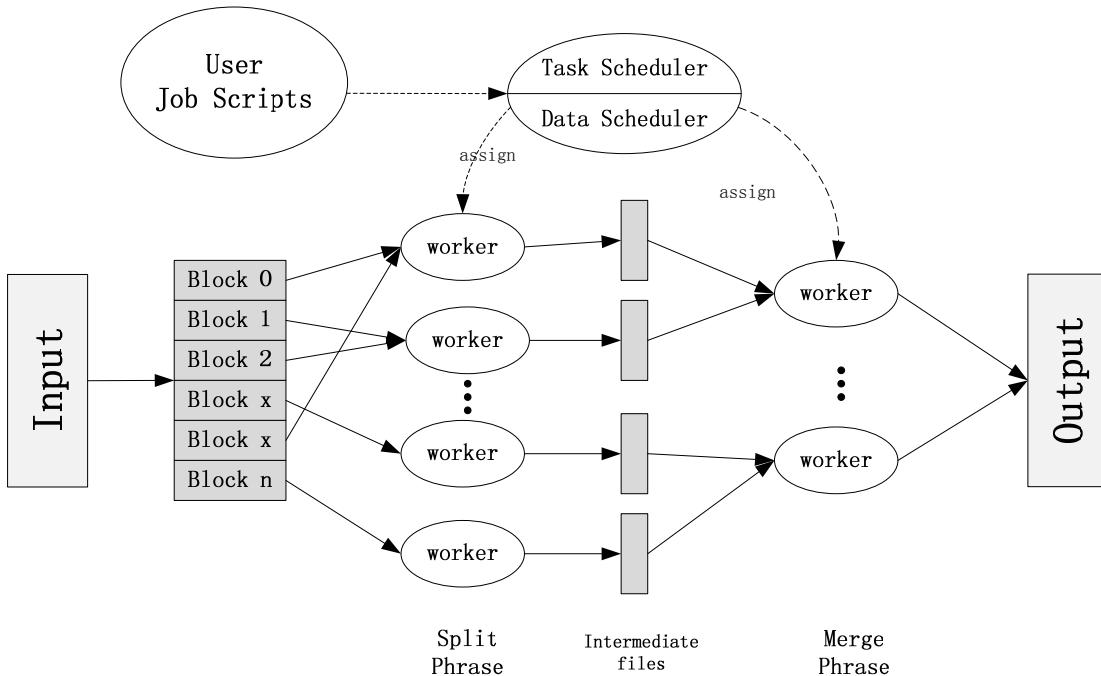


Fig.4 the basic data flow in the parallel processing framework

2.4 Performance evaluation

Here, the experimental cluster consists of 5 nodes, in which one is the master node and the other four are slave nodes. Each node is a 2-Way-Quad-Core computer running Fedora 13, equipped with two Quad-Core Intel Xeon E5405 (2GHz in each core), 8GB DDR2-667 ECC SDRAM, and 1TB hard disk (7200 rpm, 32 MB cache). The LiDAR point cloud used for this experiment was collected from Gilmer County, West Virginia in 2004, as illustrated in Fig.5. The point cloud contains 0.883 billion points and occupies 16.4 GB of external space. The entire dataset was partitioned in advance into 1,000m by 1,000m square blocks. The total number of point blocks was 2,173.

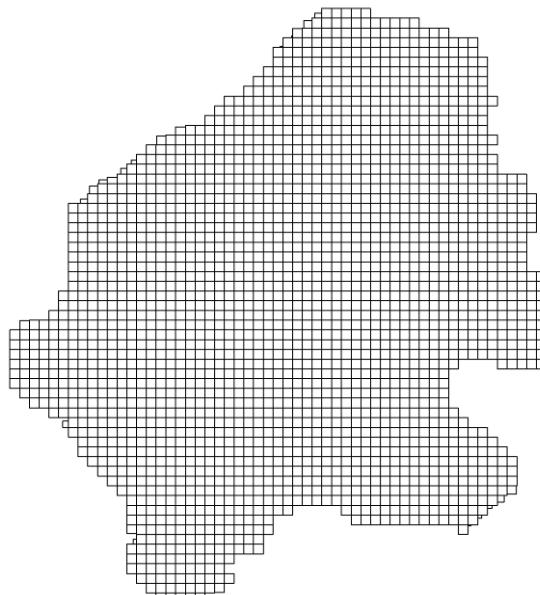


Fig.5 The Gilmer County dataset

Two common LiDAR processing algorithms, IDW interpolation and Delaunay triangulation (DT), were parallelized to examine the suitability of the proposed universal parallel framework. The detail parallelization of these two algorithms can be seen in [7, 8]. All the Split and Merge tasks were written in C++ and compiled with linux gcc 4.3. The experiment results demonstrated that significant speedup and high data-throughput are achieved, illustrated by Fig. 6. At the same time, the memory footprint was very low compared with the size of the input point data.

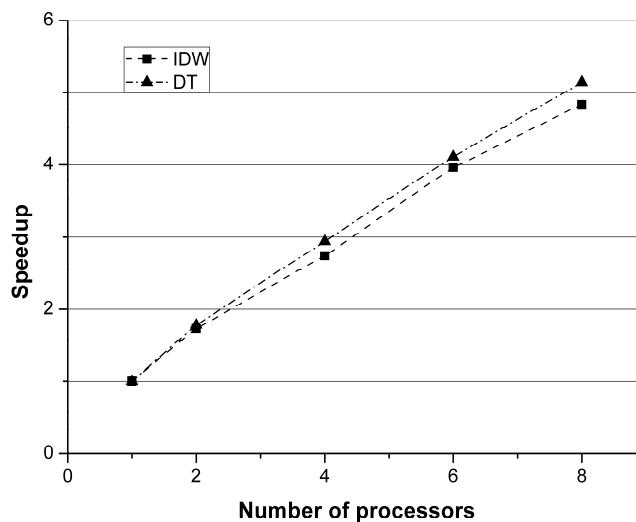


Fig.6 Speedup of parallel IDW&DT in this framework when scaling the number of processors

3. Conclusion

This paper proposed a universal parallel framework for processing massive LiDAR point clouds in a HPC environment. Within this framework, the user/programmers are provided a Split-and-Merge programming paradigm. In this way user/programmers can focus on the simple functional expression of their specific algorithm, and leave parallelization and task-scheduling to the runtime system. This framework automatically and intelligently handles key scheduling decisions for tasks and data reducing overhead related to task spawning and data communication. Two LiDAR algorithms, IDW and DT, are evaluated to prove the suitability of our proposed framework.

4. Acknowledgement

This work is supported by the Natural Science Foundation of China (Grant: 40971211).

5. References

- [1]. Leigh, Charlotte L, Kidner David B,Thomas Malcolm C. The Use of LiDAR in Digital Surface Modelling: Issues and Errors. *Transactions in GIS*, 2009, **13**(4): 345-361
- [2]. Kunle, Olukotun,Lance Hammond. The future of microprocessors. *Queue*, 2005, **3**(7): 26-29
- [3]. Hongchao, Ma,Wang Zongyue. Distributed data organization and parallel data retrieval methods for huge laser scanner point clouds. *Computers & Geosciences*, 2011, **37**(2): 193-201
- [4]. Huang, Fang, Liu Dingsheng, Tan Xicheng, Wang Jian, Chen Yunping,He Binbin. Explorations of the implementation of a parallel IDW interpolation algorithm in a Linux cluster-based parallel GIS. *Computers & Geosciences*, 2011, **37**(4): 426-434
- [5]. Cannataro, Mario, Talia Domenico,Srimani Pradip K. Parallel data intensive computing in scientific and commercial applications. *Parallel Computing*, 2002, **28**(5): 673-704
- [6]. Staples, Garrick, TORQUE resource manager, in Proceedings of the 2006 ACM/IEEE conference on Supercomputing. 2006, ACM: Tampa, Florida. p. 8.
- [7]. Guan, Xuefeng,Wu Huayi. Leveraging the power of multi-core platforms for large-scale geospatial data processing: Exemplified by generating DEM from massive LiDAR point clouds. *Computers & Geosciences*, 2010, **36**(10): 1276-1282
- [8]. Wu, Huayi, Guan Xuefeng,Gong Jianya. ParaStream: A parallel streaming Delaunay triangulation algorithm for LiDAR Points on Multicore Architectures. *Computers & Geosciences*, 2011, DOI: 10.1016/j.cageo.2011.01.008

Using a Moving Window SVMs Classification to Infer Travel Mode from GPS Data

A. Bolbol, T. Cheng, J. Haworth

Department of Civil, Environmental and Geomatic Engineering, University College London,
Gower Street, London WC1E 6BT, United Kingdom
Email: {a.bolbol; tao.cheng; j.haworth@ucl.ac.uk}

1. Introduction

Understanding travel behaviour is important for studying tourist activity, the quality of life, a strike's impact on transportation and other environmental impacts. However, it is a challenge to model travel behaviour due to its complexity and diversity. Attempts have been made to infer meaningful information about travel behaviour from positional data obtained from sensors such as GPS. Among these types of information is the travel mode (e.g. cycling, walking, bus and so forth). This inference could largely replace or complete a lot of the feedback required by users when labelling and tagging their travel diaries.

Previous machine learning (ML) approaches that attempt to derive travel modes from GPS data suffer from design decisions that limit their accuracy and flexibility. For example, Zheng et al. (2008) compares different machine learning methods such as Decision Tree and Bayesian Net to segment tracks into partitions of different travel modes. However, the process depends on real-life assumptions that could differ from one person to another. Liao et al. (2007) uses Hierarchical Conditional Random Fields to infer the travel mode from GPS fixes taking the user's context into consideration. It achieves a good accuracy; however, it relies heavily on temporal features such as the duration and time of day, which again differs from one person to another. Other methods use Neural Networks to do a similar inference (Gonzalez et al., 2008); however, Neural Networks deliver multiple solutions associated with local minima and for this reason may not be robust over different samples.

In this work we attempt to identify travel modes from sparse GPS data, without information or assumptions about the user's context which is usually needed in other approaches. We use Support Vector Machines (SVM) to perform the inference from velocity values obtained from GPS data. Due to its high quality of out-of-sample generalization and ease of training, SVMs provide far beyond the capacities of traditional ML methods used in previous research. However, SVMs depend on data with multiple attributes to work best. To overcome this, we use a moving window that classifies instances of data sequences. We complement this by using logical filters that apply a transition matrix.

2. Dataset

The training dataset used for testing is a 2 months of a multi-modal track of three users between 01/05/2010 and 28/06/2010 (figure 1). The track is collected within London at 1 minute frequency. London is chosen due to its complexity and the diversity of its transportation networks. The dataset was labelled by the users as to which travel modes were used. The dataset was then filtered for the “cycle”, “bus”, “walk” and “stationary” modes, so as to use SVMs to infer these modes. The rest of the modes, such as the “tube” and “train” modes, are excluded because they could be easily inferred using GIS to perform network matching.

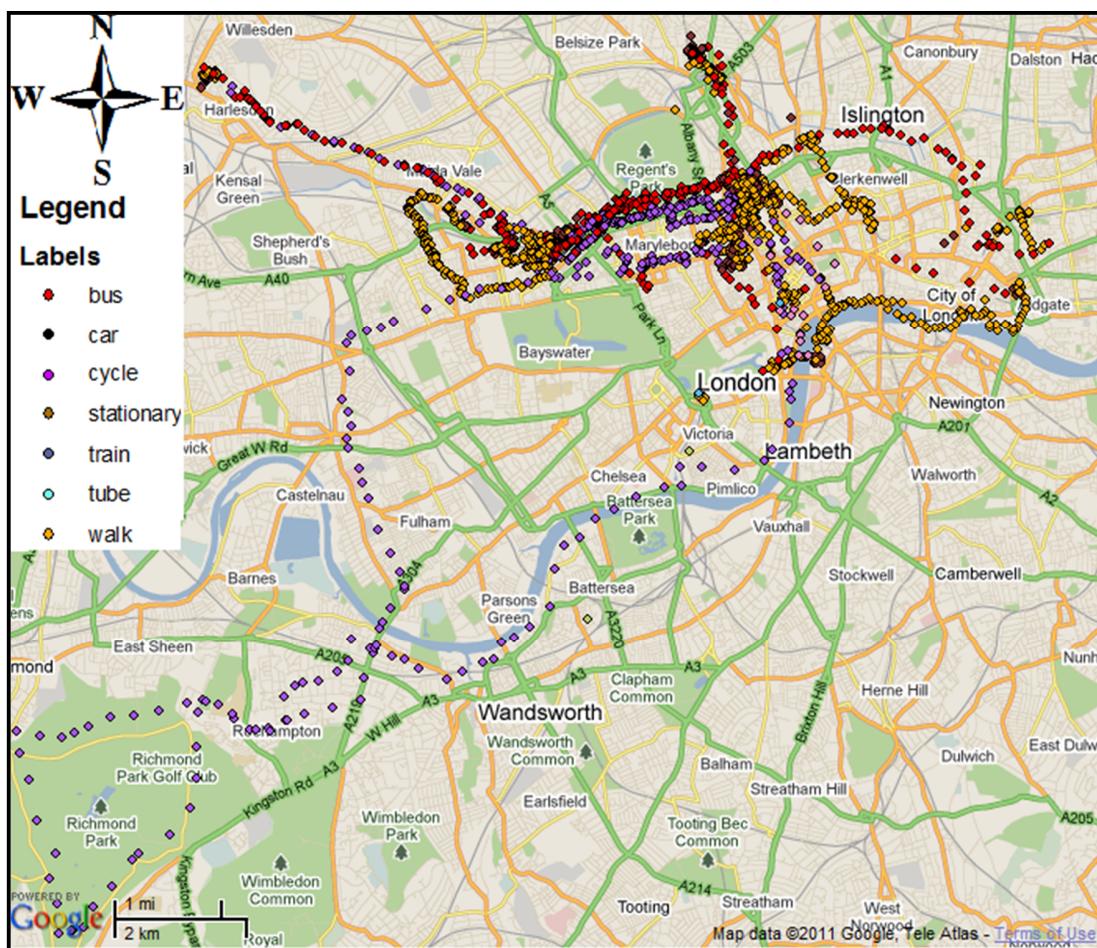


Figure 1. The Study Area in London

The number of fixes of the “walk” mode in the dataset was the highest amongst other modes and almost as double as the second highest mode (“stationary”). This demonstrates the high occurrence of walks within an individual’s daily journey. This is due to the fact that walking often occurs as an intermediate link between different modes.

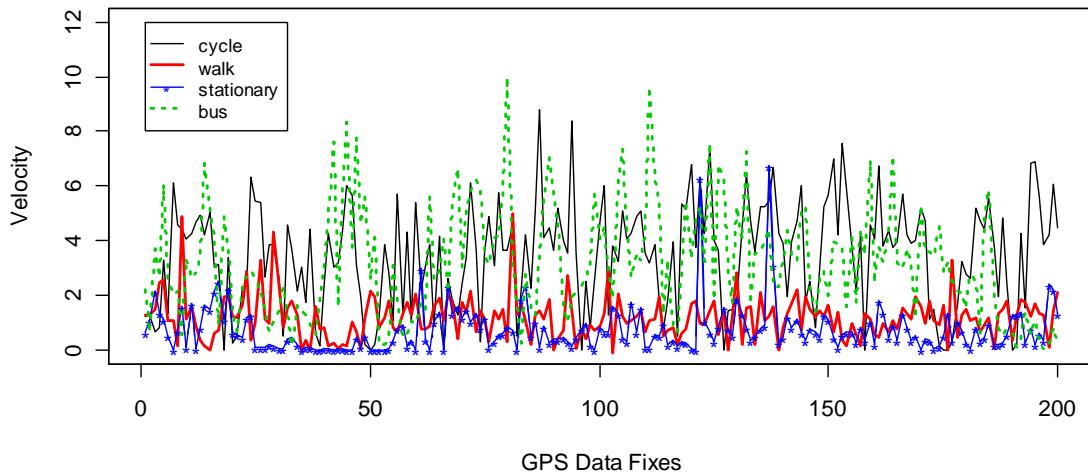


Figure 2. A Sample of Travel Modes Velocities (Outliers>15m/s Removed)

Figure 2 shows the respective velocities for the first 200 fixes of each mode of the 4 chosen modes in this dataset. Outliers due to the GPS errors are removed for values >15m/s. The figure demonstrates a clear confusion and overlap between “walk” and “stationary” modes. This is due to the existence of many small stationary segments embedded within “walk” segments. There also appears to be almost a clear overlap between “cycle” and “bus” modes. This is due to the similarity of the general speeds of these travel modes in urban areas. This emphasizes the nature of different forms of commute in the London network.

3. Using Support Vector Machines for Classification

Velocity values are calculated for every arc of the track. However, non-fix indoor activity causes false velocity calculation for the first point that follows them, therefore, these values are excluded. Data sequences of the same travel modes are then aggregated and prepared for the SVMs learning process.

3.1 Instance-based Classification

Support Vector Machines work best in a multi-attribute environment. Therefore, once the data is aggregated, it is divided into equal sized instances of several arcs as demonstrated in table 2. This simulates the multi-dimensionality of the data in the learning process which SVM is best at dealing with. Another reason for using instances is that for a more accurate classification, it is more meaningful to learn a certain period of a trip than one single value which could be misleading (e.g. bus stopping at the traffic lights could be misclassified for walk or stationary).

Data Instance			Label
X1	X2	X3	Walk
X4	X5	X6	Walk
X7
...	...	Xi	Bus
...	Bus
...
...	...	Xn	...

Table 1. The Division of Data into Equal-Sized Instances
(Three in this case)

The data instances then are divided into two thirds and one third for training and testing respectively. These similar-mode instances enter the SVMs learning process using a stationary Gaussian kernel with a radial basis kernel function for training due to their flexibility. The SVM classification machine is also trained using a multiclass method.

As shown in figure 3, the classification gives better results for longer data instances. However, a longer sequence of mixed travel modes could introduce higher complexity. Therefore, we chose to use the small-sized instance that still contains a decent number of arcs to represent a realistic sequence; in this case three.

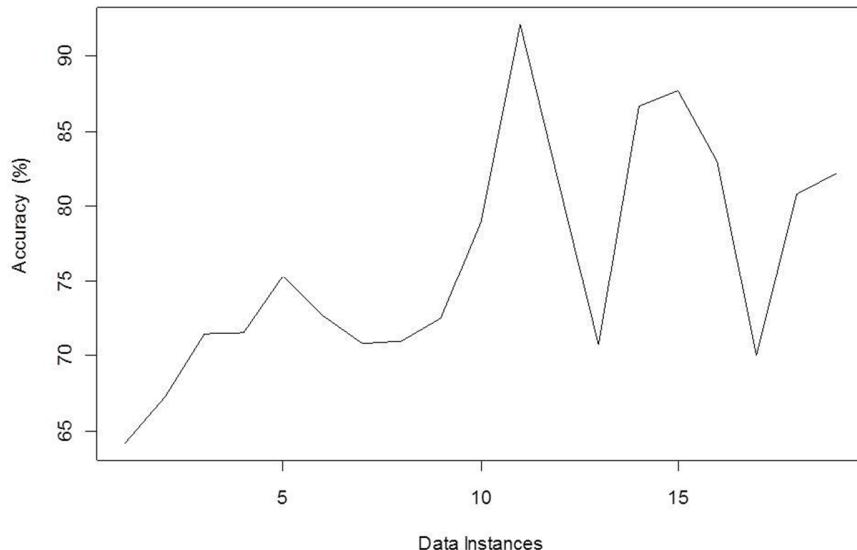


Figure 3. SVM Classification Accuracies due to the Usage
of Different Lengths of Data Instances

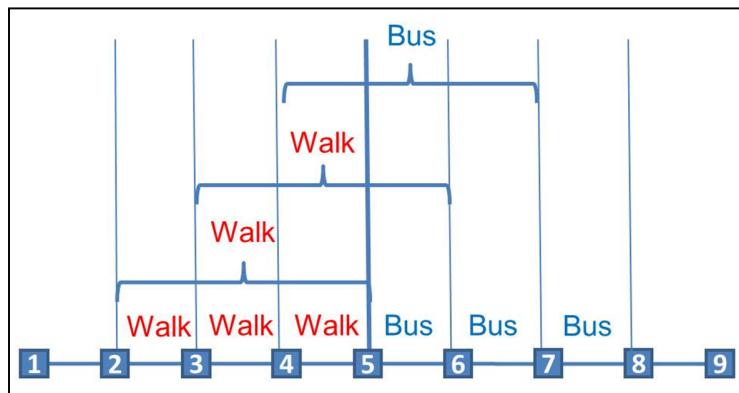
The accuracy achieved by this primary classification is around 71%, and the confusion matrix is illustrated in table 3. A good separation between the walk-stationary and the bus-cycle modes is achieved. However, the instances are non-realistic since they assume that the track is already segmented into similar-mode segments.

Inferred	Truth			
	Bus	Cycle	Stationary	Walk
Bus	20	16	0	2
Cycle	19	38	0	2
Stationary	0	0	41	4
Walk	2	8	34	110

Table 2. Confusion Matrix for 3-Arcs-Lengthed Instances

3.2 SVM moving window algorithm

To resolve the previously stated problem, we propose applying a fixed-length moving window on the whole track, and moving that window on an arc-by-arc basis along the track's velocity values. Every time the window slides, a classification of that instance of data is performed. Figure 4 (a & b) illustrates this method, where a moving window classifies each 3-sized instance moving arc-by-arc along the track.



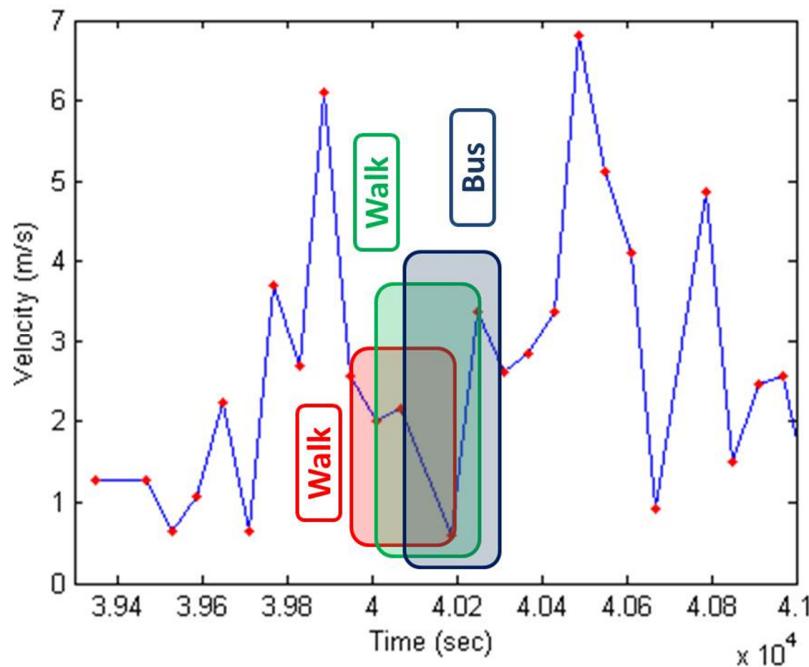


Figure 4. A Moving Window Classifying Each 3-Sized Instance Moving Arc-by-Arc along the Track (a) Illustrated Abstractly Above, and (b) with Velocity against Time Below

After each instance is classified, the algorithm runs through the new inferred labels and changes any improbable inferences such as a “bus” instance between a series of “cycle” instances. It also applies a transition matrix (Table 4) which amends the sequence according to the probability of switching between certain modes (such as cycle–bus–cycle), and changes the less probable mode. This matrix is based on (Zheng et al., 2008) and is compiled from this research. As we could note, almost all modes are followed by a “walk” mode. “Walk” and “stationary” modes are very interchangeable while a portion of “cycle” is followed by a stationary mode due to the chaining-the-bike activity.

Travel modes	Walk	Stationary	Bus	Cycle
Walk		59.8%	12.6%	27.6%
Stationary	50.0%		36.5%	17.5%
Bus	98.7%	0%		1.3%
Cycle	76.3%	23.7%	0%	

Table 3. Transition Matrix of Travel Modes

The amended classified instances are then switched back again into classified arcs by reverting the idea of the moving window. Once a change in mode occurs, this could indicate that the majority of the 3 arcs within that instance have changed into a new travel mode, and hence the first of these 3 arcs is classified as the previously ending mode. Figure 5 illustrates an abstract example of these classified arcs.

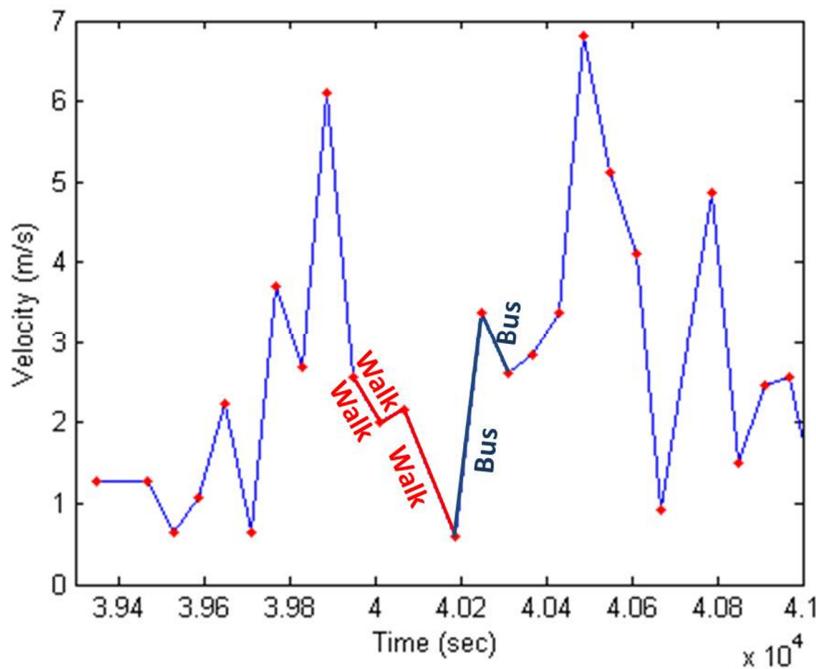


Figure 5. The Result of the Moving Window after Assigning the Classification to the Arc Level

4. Results

The results of the moving window algorithm reveal an accuracy of 70 %, without having to pre-segment the track. Figure 6 shows the classification results compared to the actual classes. Table 5 illustrates the confusion matrix of this classification. The high confusion between “stationary” and “walk” modes is not an error at all; in fact it is due to the existence of lots of actual stops within any walking pattern.

Inferred	Truth			
	Bus	Cycle	Stationary	Walk
Bus	115	98	13	21
Cycle	162	414	17	28
Stationary	8	6	407	160
Walk	86	39	235	848

Table 4. Confusion Matrix of the Moving Window Algorithm

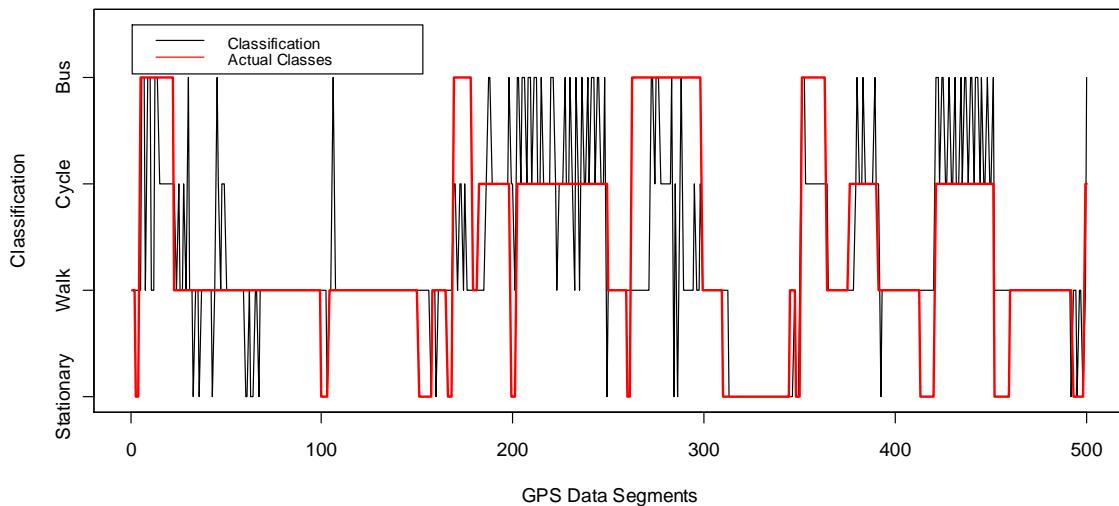


Figure 6. The Classification Results Compared to the Actual Classes (Sample)

Merging both classes into a single class called “walks” and re-running the algorithm again, an accuracy of 85% was achieved. This demonstrates a major improvement on the previous accuracy with no pre-segmentation. Table 6 illustrates the confusion matrix of this second run, while figure 7 shows the classification results compared to the actual classes. Some classification errors could be noted in the middle of long segments, such as the “walk” segment at the first 200 records of the dataset in figure 8.

Inferred	Truth		
	Bus	Cycle	Walks
Bus	125	105	32
Cycle	149	395	29
Walks	97	57	1668

Table 5. Confusion Matrix of the Moving Window Algorithm Merging “Walk” and “Stationary” Modes

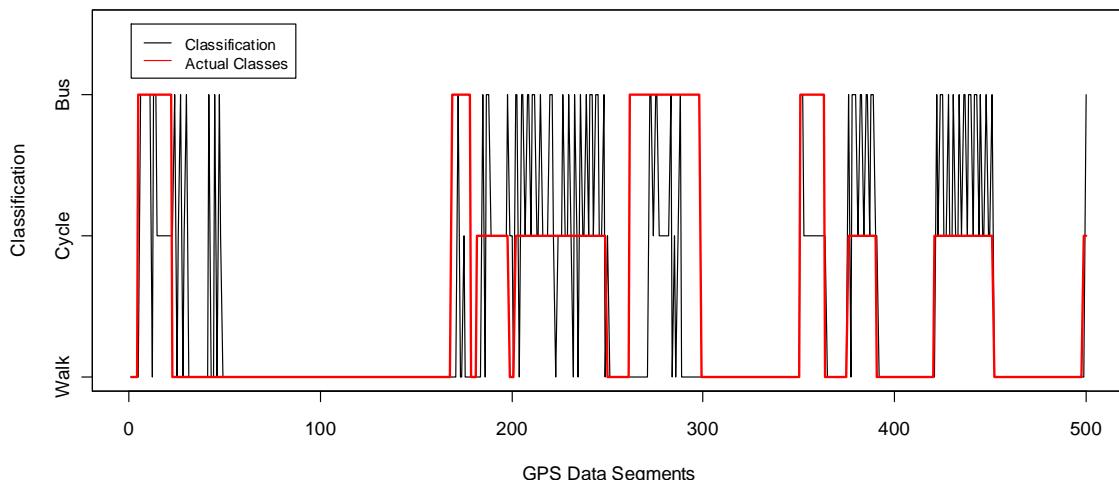


Figure 7. The Classification Results Compared to the Actual Classes – with Merged “Walk” and “Stationary” Classes (Sample)

5. Conclusions

We provided a novel approach for identifying the travel mode from GPS data. In contrast to existing techniques, our approach uses one consistent framework to classify each arc into its respective travel mode. This is done using SVMs to learn from data instances which each consist of a sequence of similar-mode data.

The moving window approach overcomes SVMs' shortcoming of requiring multiple attributes to give best results by learning from these instances. The power of SVMs here is taking motion patterns of each travel mode into consideration. The classification is carried out using a moving window to classify instances on an arc-by-arc basis along the track's velocity values. After applying a transitional matrix and merging intermediate modes, our model achieves 85% accuracy rate without the need of having pre-segmented data. The next task is to attempt to further separate "cycle" and "bus" modes which seems to be of high confusion because of their similar motion characteristics.

6. Acknowledgements

This research is supported by EPSRC and u-blox. We would like to express gratitude to Chris Marshall for his support and constructive comments on the work reported here.

7. References

Gonzalez P et al, 2008, Automating Mode Detection Using Neural Networks and Assisted GPS Data Collected Using GPS-enabled Mobile Phones, *Proceedings of the 15th World Congress on Intelligent Transportation Systems*, New York, New York, November 16-20.

Liao L, Fox D and Kautz H, 2007, Hierarchical Conditional Random Fields for GPS-based activity recognition. In Robotics Research: *The Eleventh International Symposium*, Springer Tracts in Advanced Robotics (STAR), Volume 28/2007, 487-506.

Zheng Y, Liu L, Wang L and Xie X, 2008, Learning transportation mode from raw GPS data for geographic applications on the Web. *In Proceedings of WWW 2008*, ACM Press, 247-256.

Putting the Geographical Analysis Machine on the Internet Revisited

Ian Turton¹ and Andy Turner²

¹GeoVISTA Center, Department of Geography, The Pennsylvania State University,
University Park, PA 16801 ijt1@psu.edu

²The Centre for Computational Geography, School of Geography, University of Leeds,
Leeds LS2 9JT A.G.D.Turner@leeds.ac.uk

Introduction

Openshaw et al. (1999) wrote about their experiences putting the Geographical Analysis Machine (GAM) on to the Internet to allow a wider use. Their motivation was that, at the time, proprietary GIS programs lacked sophisticated geographical analysis technology. The experimental system that was developed relied on a complex set of Unix scripts and FORTRAN code and proved to be unsustainable. The system also required a text file containing the locations of population and case points which was difficult to create for a naïve user. Lastly it required users to upload these potentially confidential information to a remote server trusting the server administrators who were unknown (but obviously trustworthy). In the last decade, since the original attempt to put GAM on the Internet, the situation appears not to have improved. There are still demand for more user friendly and secure ways of exploring geographical data for evidence of spatial clustering. Robertson and Nelson (2010) state that "...training and software availability were cited as the primary barriers to the uptake of space-time disease surveillance..." and provide a general assessment that for the programs they tested - handling the data formatting was difficult and the interpretation of outputs was challenging.

Motivation

This paper seeks to solve the same problem that Openshaw et al. (1999) attempted, making use of modern developments in cloud and grid computing, distributed spatial data management and improved computing power. From the literature (*e.g.* Olsen et al., 1996; Robertson and Nelson, 2010) there is a demand from epidemiologists for a simple system that will: import their case data; import population data (preferably from a Census site directly, or from files they download from one); and, exports a geographically referenced, easy to understand map of the potential clusters for them to investigate.

As with anyone handling confidential data, epidemiologists are concerned about data security. Any system that is to be used with confidential data needs some form of guarantee that the data will be secure and will not become available to others (at least not provided without clear usage restriction and only to other users of those confidential data). To guarantee the security of a software system running on a networked machine, the software source code needs to be inspected

and the authority for a guarantee trusted. A system that is completely open source and based on standards compliant open source service components allows anyone to inspect the source code and this helps to verify that data is secured in the system. Additionally, being open source allows for the academic rigour of the algorithm implementations to be assessed. In a closed source system, arguably there is too great a risk with respect to confidential data security.

Implementation

The system described is programmed in Java using the GeoTools library (Turton, 2008). The system makes use of the Web Processing Standard (WPS) (OGC, 2007) as implemented by GeoServer (an open source server which implements other important Open Geospatial Consortium (OGC) standards). Users are able to seamlessly import data from compliant Web Feature Servers (WFS) (OGC, 2005) and export the results via a Web Map Server (WMS) (OGC, 2002) as a layer or in a variety of georeferenced imagery formats. Thus a user need only install the latest version of GeoServer and add the required jar files to have a fully functioning system on a computer with Java installed. Ideally a user will be able to configure their system to pull data from a remote server which is serving up population data from a central WFS. WPS allows for the user to specify the input as coming from a sub-process, so a user can construct a model to calculate a more complex expectancy or Population At Risk (PAR) estimate. However, if needed, it is simple for the user to add required PAR data layers to their own GeoServer instance.

To allow for the generation and comparison of results from different spatial clustering methods, the system described has the following methods:

- The GAM/K system (Openshaw, 1996) which carries out an exhaustive search by applying a range of circles to the whole spatial area of the data set. While this method is sure to find a cluster (if one exists) it can be prohibitive to carry out this level of search on large data sets.
- The rare disease cluster detection method of Besag and Newell (1991) searches for clusters by examining circles centred on cases with a radius determined by the k neighbouring cases. This reduces the number of circles to be examined but the nearest neighbour calculation can be time consuming with large data sets.
- The SatScan algorithm (Kulldorff, 1997) makes use of a scan statistic calculated for circles centred on each population point and extended to include up to half the total population at risk. The paper is unclear on the preferred method to expand the circle so we opt to extend the circle point by point though this leads to issues with nearest neighbour calculations again.
- A random circle method formalized by Fotheringham and Zhan (1996) allows a quick but non exhaustive scan of the data set. For very large datasets a user might choose to search quickly using random circles across the whole map and then apply one of the other methods in a smaller rectangle constrained to interesting areas.

By providing a standardised interface to these contrasting systems this system will allow epidemiologists to investigate large data sets using a fast scan method (such as Random Circles) and then rerun the analysis restricted to areas of interest using an exhaustive method (such as GAM). As can be seen in Figure 1 both methods find the same clusters, but the random search is 15 times faster, but the GAM analysis is more detailed and finds more structure. Users are also able to analyse very large data sets by connecting the back end of the processing system to a cloud computing environment.

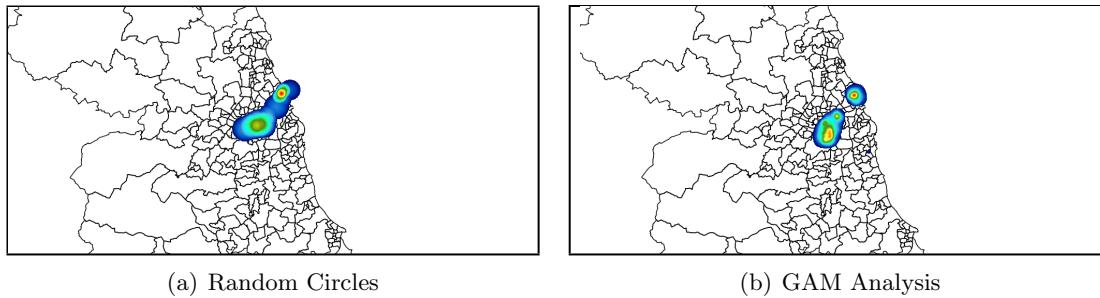


Figure 1: Analysis of Cancer in NE England (a)1000 Random Circles, (b)GAM (15,649 Circles)

Conclusions

This paper describes a system developed for epidemiologists to use to search large databases to find clusters of rare diseases (such as Childhood Leukaemia). The system is made available as open source software and is based on standards compliant OGC services. Providing the system as open source allows it to be verified as secure to work in networked environments with confidential data and it allows the academic rigour of the algorithmic implementations to be assessed. The system allows the user to pull in Census data from servers that serve it via a WFS. The system can be readily installed locally and on Grid and Cloud computing infrastructures. Results are made available to the user using the WMS standard which allows for them to be overlaid with other data which allows for further geographical exploration of the data which may help to explain spatial clustering in the incidence data.

References

- Besag, J. and Newell, J. (1991). The Detection of Clusters in Rare Diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 154(1):143–155.
- Fotheringham, A. S. and Zhan, F. B. (1996). A Comparison of Three Exploratory Methods for Cluster Detection in Spatial Point Patterns. *Geographical Analysis*, 28(3):200–218.
- Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and Methods*, 26(6):1481–1496.
- OGC (2002). *Web Mapping Service Standard 1.1.1*. Number 01-068r3. Open Geospatial Consortium.
- OGC (2005). *Web Feature Service Standard 1.1.0*. Number 04-094. Open Geospatial Consortium.
- OGC (2007). *Web Processing Service Standard 1.0.0*. Number 05-007r7. Open Geospatial Consortium.
- Olsen, S. F., Martuzzi, M., and Elliott, P. (1996). Cluster Analysis And Disease Mapping: Why, When, And How? A Step By Step Guide. *BMJ: British Medical Journal*, 313(7061).
- Openshaw, S. (1996). *Methods for Investigating Localised Clustering of Disease*, chapter Using a geographical analysis machine to detect the presence of spatial clusters and the location of clusters in synthetic data, pages 68–87. Number 135. IARC Scientific Publication, Lyon, France.

- Openshaw, S., Turton, I., Macgill, J., and Davy, J. (1999). Putting the Geographical Analysis Machine on the Internet. In Gittings, B., editor, *Innovations in GIS 6*, chapter 10, pages 121–132. Taylor and Francis, London.
- Robertson, C. and Nelson, T. (2010). Review of software for space-time disease surveillance. *International Journal of Health Geographics*, 9:16+.
- Turton, I. (2008). GeoTools. In Hall, B. G. and Leahy, M. G., editors, *Open Source Approaches in Spatial Data Handling (Advances in Geographic Information Science)*. Springer, 1st edition.

Inverse Estimation of the Point Position from an Image of Kernel Density Estimation

A. Takizawa¹

¹Department of Architecture and Architectural Engineering, Kyoto University Graduate School of Engineering
Kyoto University Katsura Campus 4, Nishikyo-ku, Kyoto, 615-8540, Japan
Telephone: +81-75-383-2941
Fax: +81-75-383-2941
Email: kukure@archi.kyoto-u.ac.jp

1. Introduction

Kernel density estimation (KDE) (Parzen 1962) which smoothes the distribution of data and is widely used in the field of statistical analysis. Let x_1, x_2, \dots, x_N denote the observed variables where N represents the number of variables. The estimate value at x by KDE is given as

$$\hat{f}_K(x) = \frac{1}{nh} \sum_{i \in N} K\left(\frac{x - x_i}{h}\right)$$

where h denotes the extent of the impact of the variable, and $K(x)$ denotes a kernel function. As a kernel function, a Gaussian kernel:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

is often used.

In the field of GIS, KDE has been used to visualize the density of event points that occur in a plane. If a certain number of event points exist, KDE can visualize the area where a lot of events tend to occur without showing actual points. This is useful for visualizing the distribution of the event points whose privacy needs to be protected. Therefore, if their position is estimated from an image of KDE, it might be useful for researchers who need spatially precise data. For example, points are not shown in the Japanese criminal maps of burglary unlike other street crimes. Although, if the criminal occurrence points of burglary are estimated, the relationship between space and crime can be analyzed from more directions.

In this study, the author formulates the problem of estimating the point position from an image of KDE as a nonlinear optimization problem under the condition that the number of event points is given. Since density distribution which is originally continuous will be discretized to create an image, the amount of information will lose. Moreover, since not only the point position but also another parameters required for defining KDE are unknown, this problem is more difficult than the problem estimating only the point position. For solving the problem, quasi-Newton methods (QN) and differential evolution (DE) (Strom and Price 1996) which is a relatively recent meta-heuristic are adopted as a

local optimizer and a global one, respectively. In the implementation, general-purpose computing on graphics processing units (GPGPU) that is recently showing drastic improvement of computational time in various fields is applied to for solving the problem in practical time and accuracy. Then, the proposed method is verified with sample images of KDE created artificially using GIS.

2. Estimation Method

2.1 Image preprocessing

Various information other than density distribution such as a line segment representing road are often contained in a map image. In such a case, they are removed, and an image which contains only the density distribution is created. Furthermore, if a color image is provided, they are transformed to a gray-scale one. Then, a matrix data of the image is constructed and the brightness level of each pixel is translated to an ordinal number in ascending order from 0 (i.e. the density is the lowest) to bl (i.e. the density is the highest).

2.2 Problem setting

Let $p \in P$ denote a pixel of an image, (x_p, y_p) denote the x-y coordinates of p , $q \in Q$ denote an event point, and (x_q, y_q) denote the estimated position of q . The total number of event points is given as $|Q| = N$. The estimation value of density at p derived from KDE with Gaussian kernel is defined as

$$mk(p) = \frac{1}{\sqrt{2\pi}|Q|h} \sum_{q \in Q} \exp\left(-\frac{(x_p - x_q)^2 + (y_p - y_q)^2}{2h^2}\right).$$

Let $mm(p)$ denote the brightness level of the image at p , and s denote the scale parameter. Parameter s adjusts linearly the estimate value $mk(p)$ to the brightness level whose interval is $[0, bl]$. The problem estimating the point position from an image of KDE is reduced to find the optimal parameter minimizing the error between each pixel's values of the actual image and of KDE created by the estimated position and parameters. This problem is defined as:

$$\begin{aligned} \text{minimize } & err(Q, s, h) = \frac{1}{|P|} \sum_{p \in P} (s \cdot mk(p) - mm(p))^2 \\ \text{s. t. } & 0 \leq x_p \leq num_x - 1 \\ & 0 \leq y_p \leq num_y - 1 \\ & 0 < s \\ & 0 < h, \end{aligned}$$

where num_y and num_x denote the number of pixels of the image in each direction.

2.3 Solving the problem

Above objective function is a nonlinear continuous one but has many local optimum solutions. Therefore we adopt and compare two solvers: QN and DE. While there are

several variations in DE, we adopt the one denoted as DE/rand/1/bin/simple (called “The Joker”) which can find highly accurate solution in many problems (Erik et al. 2008).

3. Implementation

Since it takes $\mathbf{O}(|P||Q|)$ time to calculate the objective function, the computational time increases in proportion to the increase of event points. Especially, since DE needs to calculate the objective function many times, reduction of the computational time is an important issue. Since the objective function is obtained by summing up each error of a pixel independently, it is easy to be parallelized. Therefore, we implement the problem in C++ with OpenMP for adopting a multi-core CPU, and with CUDA for adopting a GPGPU, and compare both computational times.

4. Experiment

4.1 Setting up

25 and 50 event points were artificially generated and their KDE images with Gaussian kernel was created using ESRI ArcGIS 9.3. Size of the image is 148 x 112 pixels of every direction. The tone level of its brightness is nine. The convergence threshold of QN is 0.001. The parameters of DE is as follows: CR (crossover rate) = 0.98, F (blending rate) = 0.30 (in case of 25 points) and 0.25 (in case of 50 points). The computational environment is as follows: CPU = Intel Core i7 960, memory = 12GB, GPGPU = NVIDIA TESLA C2050, OS = Windows 7 Professional (64bit), compiler = Microsoft Visual C++ 2008 SP1 Professional + NVIDIA Parallel NSight (CUDA 3.1).

4.2 Comparison of the computational time between two implementations

At first, we compare the computational time of two implementations by CPU and GPGPU. We set remained parameters of DE as population = 1,000 and max generation = 10. Table 1 lists the result. The computational time by GPGPU is much faster than that by CPU, and we can see that the result of GPGPU is 42.2 times faster than that of CPU in case of 50 points. Moreover, we can see that the computational time by GPGPU does not almost increase even when the number of points doubles. From above observation, it can be said that GPGPU is very effective for implementing our estimation method.

Implementation	25 points	50 points
CPU	62.5 sec.	113.9
GPGPU	2.5	2.7
CPU/GPGPU	25	42.2

Table 1. Comparison of the computational time by DE between two implementations.

4.3 Result of the point estimation

Finally, the point position is estimated. QN is performed 7,500 times by changing the initial point position, and the best solution among them is adopted. The remained parameters of DE are as follows: the number of population = 5,000 (25 points) and 7,500 (50 points); the number of generation = 50 (25 points) and 100 (50 points).

Table 2 lists the result. DE shows the best accuracy in both cases. Figure 1 shows the best estimated point position by DE. The estimated position of 25 points is close to the actual data. However, we can see the distinguished error of the estimated position of 50 points. Since the tone level of the image used in the experiment is only nine, the information of the slope of KDE was lost so much and the position might not be estimated well.

Solver	25 points	50 points
QN	0.416	0.346
DE	0.095	0.079

Table 2. Comparison of $err(Q, s, h)$ of the best solutions.

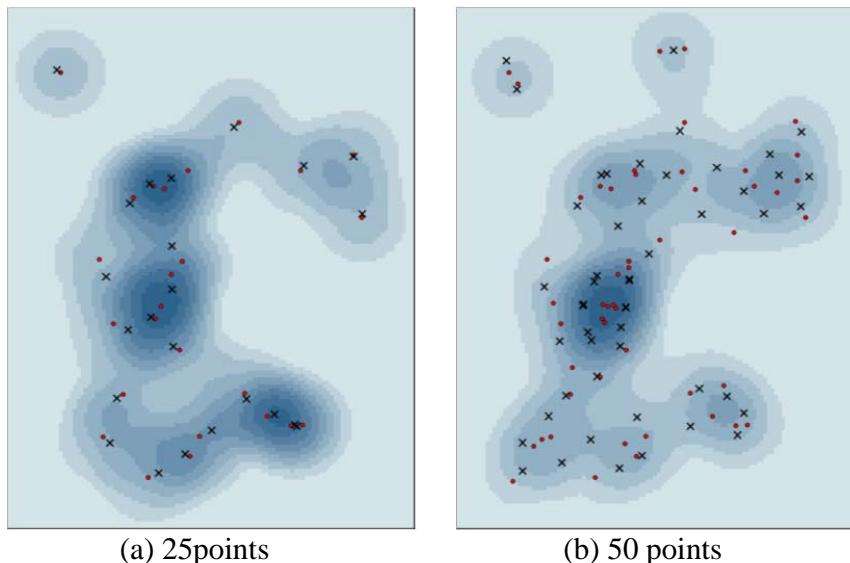


Figure 1. The best estimated position of event points by DE (\circ : actual, \times : estimated).

5. Conclusion

In this research, the author proposed the estimation method of the position of event points from a gray-scale image created by KDE under the condition that the number of event points is given. It turned out that GPGPU improves the computational time of the objective function remarkably. The result by DE was quite good compared with the result by quasi-Newton methods. If the number of event point is 25, the estimated position was similar to the actual points. If the number of points increases to 50, the error became larger. The author wishes to try further improvement of this method to be applicable with even 50 or more points. Although, since it seems to be difficult to estimate the point position completely, the estimated position should be evaluated to contain error value. As a future work, the confidence area of each estimated point position should be considered.

6. Acknowledgements

This study was supported by a Grant-in-Aid for Young Scientists (B) (No. 22760457) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

7. References

- Erik M, Pedersen H and Chipperfield A J, 2008, Parameter tuning versus adaptation: proof of principle study on differential evolution, *Hvass Laboratories Technical Report*, no. HL0802.
- Parzen E, 1962, On Estimation of a Probability Density Function and Mode. *Ann. Math. Statist.*, 33(3):1065-1076.
- Storn R and Price K, 1996, Minimizing the real functions of the ICEC'96 contest by differential evolution. In: *Proc. of the International Conference on Evolutionary Computation*, 842-844.

Kernel Regression for space-time traffic prediction under missing data

J. Haworth¹, T. Cheng², J. Shawe-Taylor³

¹CEGE, University College London, Gower Street, London, WC1E 6BT
Email: j.haworth@ucl.ac.uk

² CEGE, University College London, Gower Street, London, WC1E 6BT
Telephone: +44 (0)20 7679 2738
Fax: (international codes)
Email: tao.cheng@ucl.ac.uk

³ Dept. of Computer Science, University College London, Gower Street, London, WC1E 6BT
Telephone: +44 (0)20 7679 7680
Fax: +44 (0)20 7387 1397
Email: j.shawe-taylor@cs.ucl.ac.uk

1. Introduction

Prediction of traffic variables such as flows, speeds and travel times is of vital importance in Intelligent Transportation Systems (ITS). To date, various parametric and non-parametric techniques have been used, of which Vlahogianni et al (2004) provide a good review.

Real time sensor networks such as those installed to collect traffic data are prone to missing data caused by equipment failure and other factors. In this situation, most prediction algorithms break down as they no longer have access to the current traffic patterns. Missing data are usually replaced using imputation techniques which make use of historic data in one of two ways. The first way involves harnessing the seasonal temporal autocorrelation in traffic data by using the historic data from the same time, day and location in previous weeks/months/years. These methods ignore the day to day stochastic nature of traffic. The second way involves searching the historic data for the most similar patterns to the recently observed data, either temporally or spatially. These methods rely on the assumption that all possible traffic patterns have been observed and that recent traffic patterns are available (Qu et al, 2009). Furthermore, if the spatial neighbourhood is used, they assume the presence of spatio-temporal autocorrelation (Yue and Yeh, 2008). Although more sophisticated techniques have been developed, in practice, simple algorithms such as historical average methods and exponential smoothing are widespread (Zhong et al, 2004).

In this study, we make use of both the inherent seasonality in traffic data and the assumed spatio-temporal autocorrelation to predict unit journey times (UJT, inverse of speed) under missing data on a section of road in central London. Kernel regression is used as the prediction algorithm. The results are validated using real UJT observations collected using automatic number plate recognition (ANPR) cameras on a section of the London Congestion Analysis Project (LCAP) road network.

2. Methodology

Kernel regression is a non-parametric technique simultaneously developed by Nadaraya (1964) and Watson (1964) that is used to estimate the conditional expectation of a random variable. Given a set of n pairs of variables $(X_1, Y_1), \dots, (X_n, Y_n)$, the goal of kernel regression is to estimate a regression function of Y on X as in equation 1:

$$m(x) = E(Y|X = x) \quad (1)$$

Which reads as $m(x)$ is equal to the expectation of Y given that X equals x . Using the Nadaraya-Watson estimator, an estimate of $\hat{m}_h(x)$ can be obtained using equation 2:

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n K_h(\|x - X_i\|)Y_i}{\sum_{j=1}^n K_h(\|x - X_j\|)} \quad (2)$$

Where K_h is a kernel with parameter h . The kernel regression estimator gives a weighted average of the observed independent variables Y_i and the denominator ensures that the weights sum to 1. The kernel K_h is usually chosen to be a radial basis function (RBF) kernel with bandwidth h .

In this study, we assume that the sensor at the current link c is not functioning and no data is being collected. Therefore, we must make use of the current traffic conditions on a neighbouring link n to make the predictions. This is achieved by computing the similarity between the current and previously observed patterns on link n and computing a weighted sum of the corresponding previously observed points on link c . Due to the seasonality in the data it is not necessary to examine all previous patterns, however; only searching the patterns that occurred at the same time on previous days may omit important information as traffic may develop differently from day to day. Therefore, a window w is defined centred on the current time of day t within which patterns are compared. Based on this, the UJT on the current link c at time of day $t + 1$ is predicted according to equation 3:

$$\hat{m}_h(Y^c)_{t+1} = \frac{\sum_{l=1}^w \sum_{j=1}^d K_h(\|X^n_t - X^n_{ij}\|)Y^c_{ij}}{\sum_{k=1}^w \sum_{l=1}^d K_h(\|X^n_t - X^n_{kl}\|)} \quad (3)$$

Where $X^n_t = x_t, x_{t-1}, \dots, x_{t-m}$ is the current traffic pattern on neighbouring link n with temporal embedding dimension m ; X^n_{ij} is the corresponding traffic pattern at window point i on day j ; Y^c_{ij} is the observation recorded at current link c corresponding to window point i on day j ; w is the size of the window centred at time t and d is the number of days in the training dataset. Therefore, the estimator is computed using $w * d$ pairs of training examples at each time point.

3. Data and Experimental Procedure

The data obtained are from the London Congestion Analysis Project (LCAP) network. LCAP is a system of automatic number plate recognition (ANPR) cameras maintained by Transport for London (TfL) that collect travel time information, aggregated at the 5

minute level, on London's road network (see fig. 1). The LCAP network has very high levels of missing data, in some locations up to 85%, and often cameras are not operational for days at a time. In these circumstances, the missing data are replaced with historical profiles for the whole day, which cannot properly reflect the true network conditions. For real time applications, and also for operational reasons, it is desirable to obtain a more accurate estimate of the true conditions.

The road link selected for prediction is link R425, which has a length of 925.6m. Its upstream and downstream neighbours are used separately to predict its future values. They are link R1592 (1644.4m) and R2140 (3854.7m) respectively. The data used are 33 Tuesdays between January 6th and August 18th 2009. The data are split into training (25 days), testing (7 days) and prediction (1 day) sets. The training and testing datasets are used to determine the best values for the kernel parameter h , the embedding dimension m and the window size w , and the parameters are then used for prediction on the prediction set. One step ahead predictions are made.

4. Results

Two models are constructed for comparison purposes; an historical average predictor as currently used by TfL, and exponential weighted moving average, which has also been used in practice (Zhong et al, 2004). The results are shown in table 1. The root mean squared error (RMSE) index is used to measure performance.

Model	RMSE	$h(\theta)$	m	w
Hist. Avg.	0.0207	-	-	-
Exp. MA.	0.0203	(0.6)	-	-
Upstream	0.0183	0.1	2	5
Downstream	0.0194	0.1	2	5

Table 1. Comparison of models.

From the results it can be seen that the exponential moving average model outperforms the historical average model as it takes into account the seasonal temporal autocorrelation in the data. However, the improvement is small at 1.93% because it ignores the day to day stochastic variation. By making use of the current neighbourhood conditions, the kernel regression technique is able to produce results that are a further 9.85% and 4.43% better than the exponential moving average using upstream and downstream links respectively. Examining figure 1, it can be seen that the kernel regression method is able to follow the general traffic pattern more closely, although it performs less well in the PM peak period.

5. Conclusions and future directions

This study has demonstrated the potential of kernel regression for predicting future traffic conditions on an urban road link under the hypothesis of missing data. Further validation is needed to test the sensitivity of the model parameters and the performance of the model under different scenarios such as recurrent and non-recurrent congestion. Additionally, an

optimal combination of the predictions made at upstream and downstream locations needs to be found and adaptive weighting schemes will be considered.

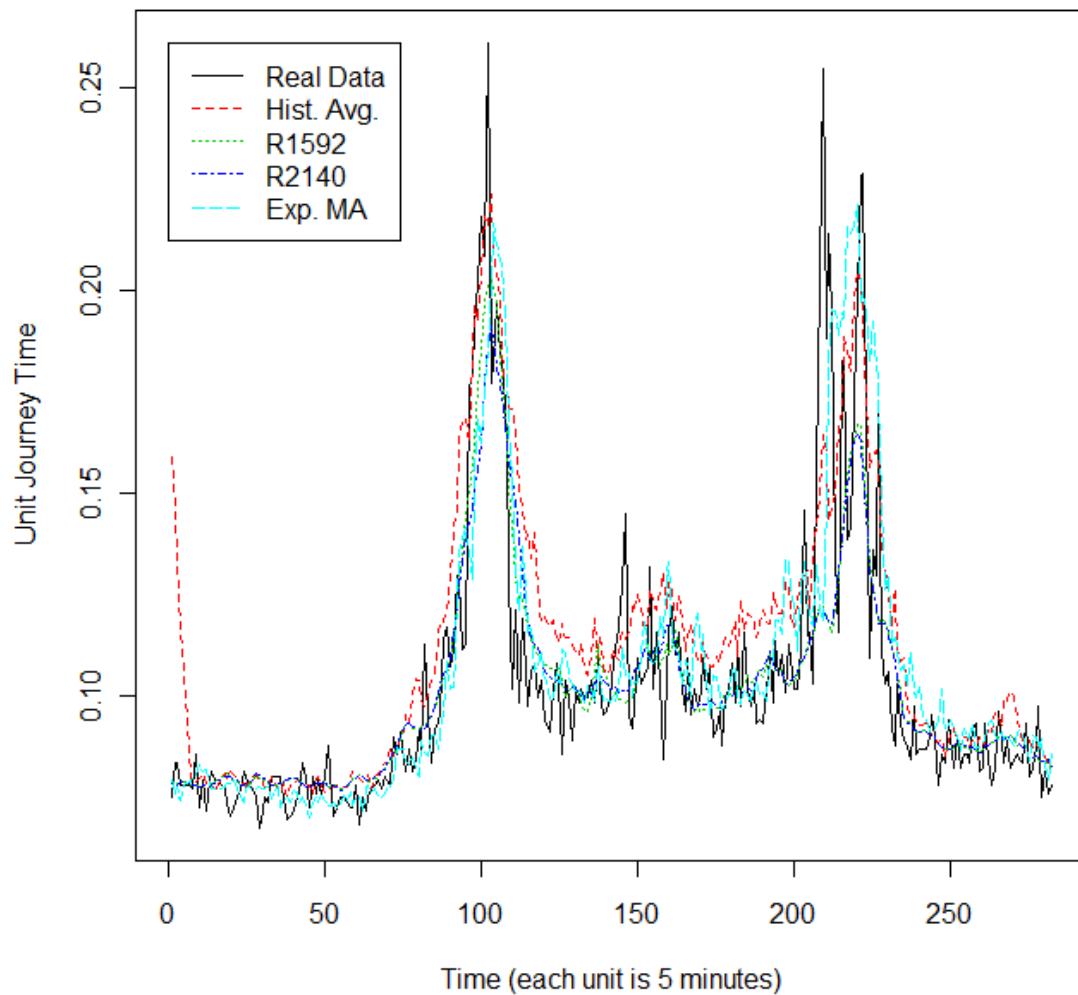


Figure 1. Comparison of predictors.

6. Acknowledgements

The authors would like to thank Transport for London for their kind support and for providing the data used in this study.

7. References

- Huang, R., Sun, S. & Liu, Y., 2011. Sparse Kernel Regression for Traffic Flow Forecasting. In D. Liu et al., eds. *Advances in Neural Networks – ISNN 2011*. Springer, Berlin, Germany: 76-84.
- Lei Han et al., 2010. Locally kernel regression adapting with data distribution in prediction of traffic flow. In *Proceedings of the 18th International Conference on Geoinformatics*: 1-6.
- Nadaraya, E.A., 1964. On Estimating Regression. *Theory of Probability and its Applications*. 9(1):141.
- Qu, L. et al., 2009. PPCA-Based Missing Data Imputation for Traffic Flow Volume: A Systematical Approach. *IEEE Transactions on Intelligent Transportation Systems*, 10(3):512-522.
- Sun, S. & Chen, Q., 2008. Kernel Regression with a Mahalanobis Metric for Short-Term Traffic Flow Forecasting. In C. Fyfe et al., eds. *Intelligent Data Engineering and Automated Learning – IDEAL 2008*. Springer, Berlin, Germany: 9-16

- Vlahogianni, E.I., Golias, J.C. & Karlaftis, M.G., 2004. Short-term traffic forecasting: Overview of objectives and methods. *Transport Reviews: A Transnational Transdisciplinary Journal*, 24(5):533.
- Watson, G.S., 1964. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*. 26(4): 359–372.
- Yue, Y. & Gar-On Yeh, A., 2008. Spatiotemporal traffic-flow dependency and short-term traffic forecasting. *Environment and Planning B: Planning and Design*, 35:762-771.
- Zhong, M., Lingras, P. & Sharma, S., 2004. Estimation of missing traffic counts using factor, genetic, neural, and regression techniques. *Transportation Research Part C: Emerging Technologies*, 12(2): 139-166.

Selective progressive transmission of vector data

Fangli Ying, Peter Mooney, Padraig Corcoran, Adam Winstanley

Department of Computer Science, National University of Ireland Maynooth,
Maynooth, Co. Kildare, Ireland.
email {fyng}@cs.nuim.ie

1. Introduction

Delivering raw geospatial data to mobile devices is an interesting and challenging computational and user-interface problem. Geospatial data can be rendered in real-time on the mobile device using appropriate visualisation software running on the mobile device operating system. Currently the majority of approaches in delivering geospatial data to mobile devices provide pre-rendered maps (tiles, images). While tile-based approaches have evolved into a defacto standard we feel there are a number of advantages in delivering geospatial data in raw vector formats (XML, GML, Shapefile, etc) to mobile devices including: *User personalisation*: User can choose which geographical features are displayed, change map themes, set visualisation preferences, etc. *Timeliness*: The user is always provided with the most up-to-date and recent versions of the spatial data. A number of constraints imposed by the mobile device environment provide major challenges including: screen resolution, available network bandwidth, and usability issues arising from providing map visualisation on small screens (Raper et al.; 2007). In this extended abstract we describe an implementation of a selective progressive transmission scheme for vector data. We use OpenStreetMap (OSM) as the case-study vector dataset. OSM data has a number of attractive features which make it a useful case study, these including: in many areas, OSM data often changes very quickly; OSM attempts to map a very wide range of geographical features; and is freely and openly available. In our implementation a user requests an area of OSM data they wish to view on their mobile device. This OSM data is downloaded immediately on the server where it is generalised. This OSM data package is then progressively transmitted beginning with a low level of detail version of the dataset. In an iterative process additional spatial detail is transmitted to the mobile device until the full resolution dataset is delivered. Our paper provides a brief overview of the implementation of our progressive transmission scheme. We describe an example of selective progressive transmission for a sample OSM dataset.

2. Implementation and Discussion

In previous papers (Ying et al.; 2010b,a) we proposed a model for progressive transmission. This model has been implemented in the Android platform. Figure 1 provides a flowchart of the implementation of this model. The user selects an area from an OpenStreetMap (OSM) slippy map on their mobile device. The Android client application sends a request to our server system. The OSM-XML data corresponding to the area requested is downloaded immediately from OpenStreetMap.org using the OSM API. For improved performance the OSM-XML is processed using data streaming. A Java-based implementation was written

for the OSM-XML processing and subsequent generalization of the spatial data. We use two generalization approaches. The well-known Douglas Peucker algorithm is used for polyline simplification. For polygons it is very important to preserve shape/contour attributes for rendering on the small screen of a mobile device (Setlur et al.; 2010). We employ a very well known method from the domain of computer vision which preserves the shape of a contour across levels of detail. The method by Latecki and Lakamper (2000) is a contour preserving approach to generalization of polygons. Some OSM polygons and polylines are greatly under-represented while others are very well represented with many hundreds of nodes (Mooney et al.; 2010). Consequently some of the features in the input dataset are more heavily generalized than others. Figure 2 outlines the data structure used to maintain the ordering of nodes from the geographic features which undergo generalization. For a given node n_i a number of characteristics are maintained in the data structure including: the nodes n_j and n_k which are connected to n_i in the polygon P or polyline L ; the significance KS_i to the overall polygon which is calculated from the angle at n_i and length of this node's incoming and outgoing edges (from (Latecki and Lakamper; 2000)); the order or position where n_i was removed during the generalization process is used by the progressive transmission to progressively rebuild the polygon or polyline.

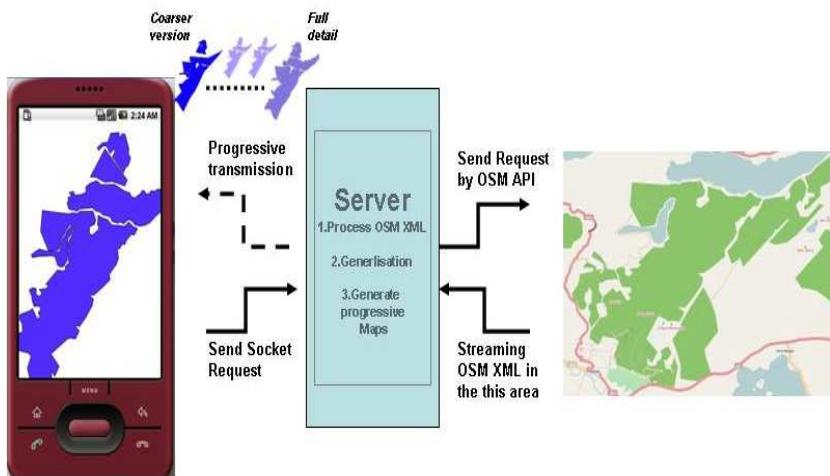


Figure 1: A schematic diagram illustrating the software implementation of our progressive transmission model

Figure 3 shows screenshots from the Android device of an example of progressive transmission (top row of images) and selective progressive transmission (bottom row of images) applied to a sample OSM dataset. The left-most column shows the progressive transmission process when only 20% of the original nodes are present in the input dataset. Subsequent columns show 40%, 60%, 80%, and finally the right-most column shows the full resolution (100%) dataset. Two polygons are coloured in blue. The large polygon is NP_a and the smaller polygon is NP_b . In the progressive transmission example nodes are added in the reverse to how they were removed during generalization. The most significant nodes are added to the transmitted dataset first. Only close to the end of the progressive transmission are the nodes with very low overall significance transmitted. The problem with this approach is that shapes with small area (relative to other shapes in the map) containing a large number of nodes are only provided with additional spatial detail close to the end of the progressive transmission. In a selective transmission scheme (bottom row of images in Figure 3) the area

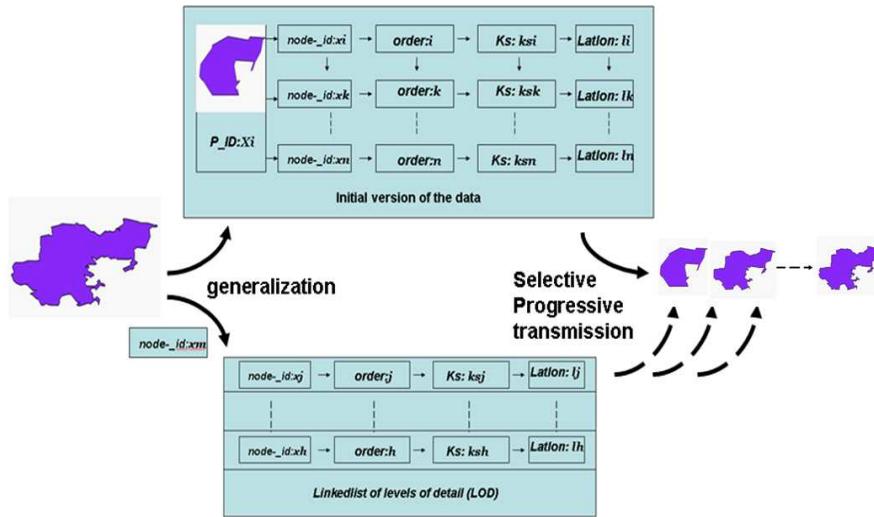


Figure 2: A schematic diagram of the indexed data structured used to maintain the ordering of nodes from the geographic features after generalization

of each polygon shape in the map is used as the selection criteria. Regardless of the significance of nodes the larger shapes in the map receive additional spatial detail before smaller shapes. This could help reduce the cognitive load on users as the larger shapes are more visible on small screen displays (Burigat and Chittaro; 2008). A number of different selection criteria could be used to drive the selective progressive transmission process. These include: measures of circularity or rectangularity of the shapes, area ratio (normalised ratio of difference between area of the polygon and its convex hull), shape complexity based on convexity of shapes (Brinkhoff et al.; 1993), map clutter indicators (Harrie and Stigmar; 2010), etc.

3. Conclusions and Future Work

With the gradual move of cartography from paper maps to web and mobile maps the requirement for real-time cartography has come into play (Yang and Weibel; 2009). We have described the implementation of a model for selective progressive transmission of vector data over the Internet to mobile devices. In this phase of our research we have used the area of the polygon shapes in the map as a shape metric to guide the selective transmission after the generalisation of the data on the server side. Using different shape metrics will affect how the spatial data is transmitted to the client device. To quantify which shape metrics work best for delivery of raw spatial data, such as OpenStreetMap, to mobile devices we are carrying out extensive user trials. During these trials with the Android-based mobile device we are collecting large quantities of additional information including zooming and panning behaviour of users as the map display progressively becomes more detailed and click/point interaction from the user with the map display. The long-term goal of this research is to develop a robust model for the smooth and seamless delivery of large quantities of raw vector data (in our case OSM data) to mobile devices. Progressive transmission strategies will become more important going forward resulting from the increased requirement of spatial content and the ubiquitous nature of mobile devices. The commercial aspect of this research is summarised by Khurri and Luukkainen (2009) who comment that to continue innovation in mapping services and user-generated content for Location-based Services map vendors will only gain competitive advantage by providing “up-to-date maps as a primary precondi-

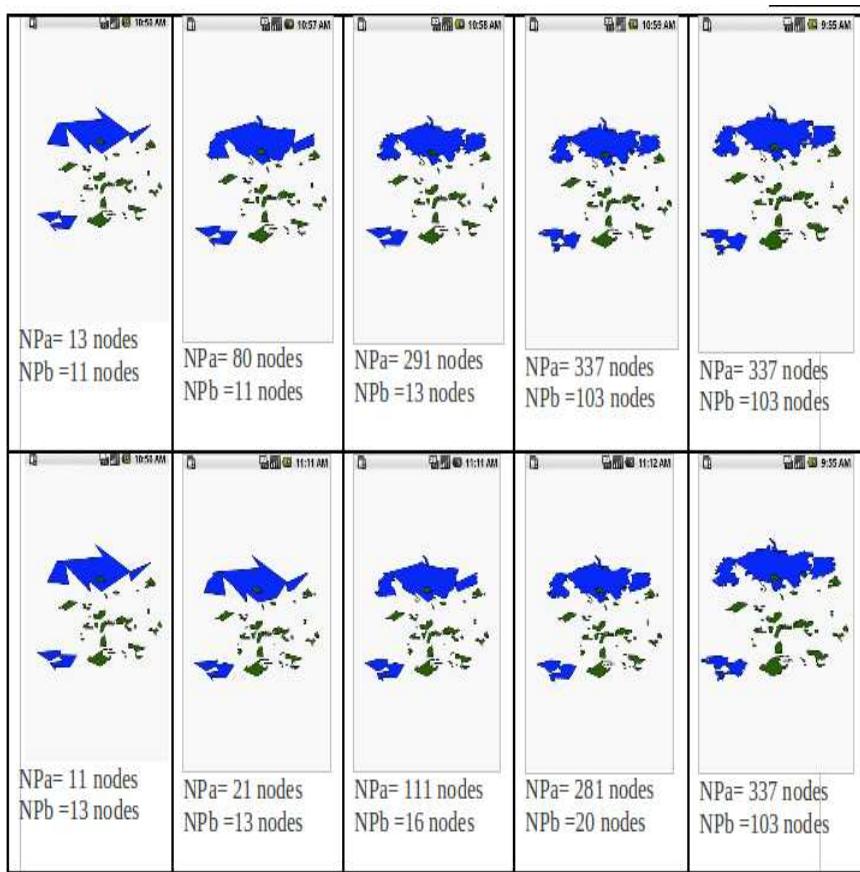


Figure 3: An example of progressive transmission (top row) and selective progressive transmission (bottom row). In the progressive transmission example detail (nodes) are added in reverse to the order they were removed while in the selective case larger shapes receive detail early in the transmission. NP_a is the large blue polygon while NP_b is the small blue polygon

tion for supplying accurate, timely and relevant content to LBS consumers”.

Acknowledgements

Fangli Ying is funded by Irish Universities Association and China Scholarship Council. Research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. Peter Mooney is funded by the Irish Environmental Protection Agency STRIVE programme (grant 2008-FS-DM-14-S4).

References

- Brinkhoff, T., Kriegel, H.-P. and Schneider, R. (1993). Comparison of approximations of complex objects used for approximation-based query processing in spatial database systems, *Proceedings of the Ninth International Conference on Data Engineering*, IEEE Computer Society, Washington, DC, USA, pp. 40–49.
- Burigat, S. and Chittaro, L. (2008). Interactive visual analysis of geographic data on mobile devices

- based on dynamic queries, *Journal of Visual Languages & Computing* **19**(1): 99–122. Spatial and Image-based Information Systems.
- Harrie, L. and Stigmar, H. (2010). An evaluation of measures for quantifying map information, *ISPRS Journal of Photogrammetry and Remote Sensing* **65**(3): 266 – 274.
- Khurri, A. and Luukkainen, S. (2009). Identification of preconditions for an emerging mobile LBS market, *Journal of Location Based Services* **3**(3): 188–209.
- Latecki, L. and Lakamper, R. (2000). Shape similarity measure based on correspondence of visual parts, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **22**(10): 1185–1190.
- Mooney, P., Corcoran, P. and Winstanley, A. (2010). A study of data representation of natural features in openstreetmap, in R. Purves and R. Weibel (eds), *Proceedings of GIScience 2010: the 6th International Conference on Geographic Information Science*, Published online at <http://www.giscience2010.org>, p. p150.
- Raper, J., Gartner, G., Karimi, H. and Rizos, C. (2007). A critical evaluation of location based services and their potential, *Journal of Location Based Services* **1**(1): 5–45.
- Setlur, V., Kuo, C. and Mikelsons, P. (2010). Towards designing better map interfaces for the mobile: experiences from example, *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application*, COM.Geo '10, ACM, New York, NY, USA, pp. 31:1–31:4.
- Yang, B. and Weibel, R. (2009). Editorial: Some thoughts on progressive transmission of spatial datasets in the web environment, *Computers & Geosciences* **35**(11): 2175–2176. Progressive Transmission of Spatial Datasets in the Web Environment.
- Ying, F., Mooney, P., Corcoran, P. and Winstanley, A. (2010a). A shape complexity approach to simplification of geospatial data for use in location-based services, in G. Gartner and Y. LI (eds), *Proceedings of the 7th International Symposium on LBS & TeleCartography 2010*, South China Normal University (SCNU), Guangzhou, China, pp. 149–154.
- Ying, F., Mooney, P., Corcoran, P. and Winstanley, A. C. (2010b). A model for progressive transmission of spatial data based on shape complexity, *SIGSPATIAL Special* **2**: 25–31.

Software Prototyping of a Heuristic and Visualized Modelling Environment for Digital Terrain Analysis

C.-Z. Qin¹, Y.-J. Lu², A-X. Zhu^{1,3}, W.-L. Qiu²

¹State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, 11A Datun Road, Anwai, Beijing 100101, China

Telephone: 86-10-64889777

Fax: 86-10-64889630

Email: qincz@lreis.ac.cn; axing@lreis.ac.cn

²School of Geography, Beijing Normal University, Beijing 100875, China

Telephone: 86-10-64889461

Fax: 86-10-64889630

Email: luyj@lreis.ac.cn; qiuweili@bnu.edu.cn

³Department of Geography, University of Wisconsin Madison, 550N, Park Street, Madison, WI 53706-1491, USA

Telephone: 1-608-2620272;

Fax: 1-608-2653991

Email: azhu@wisc.edu

1. Introduction

Digital terrain analysis (DTA) is widely used in hydrological, pedological, and geomorphological applications (Wilson and Gallant 2000). DTA in practical application is typically a modelling process of organizing different DTA tasks into a workflow with specific structure (e.g., chain, network). Construction of a proper DTA workflow is related to many aspects of knowledge in DTA domain, such as those of assigning the DTA tasks, selecting the specific algorithm for every task, setting the data flow between DTA tasks, setting parameter(s) of a given algorithm, and ensuring the match between algorithms and the specific application. This is a non-trivial process for users, especially for those not being familiar with DTA.

Current DTA-assisted software provides very limited support on the process of modelling DTA workflow. The software includes both software/toolboxes focusing mainly on DTA (e.g. TauDEM (Tarboton 1997), TAS (Lindsay 2005)) and common GIS (e.g. SAGA¹, GRASS²) which have functions of DTA. The functions of DTA in the software are often provided with a traditional menu-based style. So user must know the details on the workflow in advance and must manually set and run every

¹ www.saga-gis.uni-goettingen.de

² grass.itc.it

task in the workflow step-by-step. Recently, visualized modelling environment has been used to facilitate the geoscientific modelling process by a codeless visual means (Takatuska and Gahegan 2002; Watson and Rahman 2004; Gregersen et al. 2007). An example related to DTA is the ModelBuilder tool in the recent version of ArcGIS³. However, users still have to manually locate every task in the workflow step-by-step based only on user-owning knowledge in DTA domain. This limits the efficiency and even quality of modelling DTA workflow.

In fact, many of the knowledge in DTA domain (such as the organization among tasks, the data flow among them, selection of algorithm for a specific task, etc.) could be formalized in advance and then enable the non-expert users to model DTA workflow in a much easier way. The similar idea has been implemented in the automatic modelling tools recently emerged in semantic web and geospatial web services (Lutz et al. 2007; Yue et al. 2007) so as to facilitate the automatic discovery, access, and chaining of geospatial web services. However, existing automatic modelling tools often lack a visualized environment and rely on both web services and huge knowledge base (e.g. SWEET Ontologies⁴).

By combining the formalized DTA knowledge with visualized modelling environment, a heuristic interaction modelling environment could be implemented to support effectively on the modelling process of user-specific DTA workflow. Currently there is few DTA-focused implementation of this idea. This abstract presents a preliminary work on a light-weighted, offline software prototyping of a heuristic and visualized modelling environment for DTA (named as SimDTA VisModeler).

2. Software Prototyping of SimDTA VisModeler

2.1 Design

SimDTA VisModeler is conceptually designed to consist of three bases and four modules (fig. 1). Three bases (i.e. *knowledge base*, *model base*, and *database*) storage the formalized DTA knowledge, tasks/algorithms, and input/output data, respectively. The knowledge base supports the ability of heuristic modelling.

Four modules designed in SimDTA VisModeler are described below.

Visualized Modelling Module. This module has a graphical user interface (GUI) which has a catalog of available DTA tasks and a canvas. The canvas shows the view picture of the user-specific DTA workflow formalized, including the selected DTA tasks, the data-flow between tasks, and the organization of them (fig. 2a). This module supports the visualized modelling at three levels: task-level, algorithm-level, and execution-level. The task-level modelling means the interactive, visualized formalization of the user-specific DTA workflow. When a task is requested by user in a way of dragging the task icon into the canvas, objects drawn immediately on canvas are not only the task icon but also the related input/output data icons and the

³ www.esri.com

⁴ <http://sweet.jpl.nasa.gov/>

unidirectional links (from each input to task, also from task to each output) (fig. 2a). The algorithm-level modelling permits user to change the default settings about specific algorithm and its parameter(s) for each given task in a dialog-box activated from the corresponding task icon on canvas. The execution-level modelling is to specify input/output data for executing the workflow.

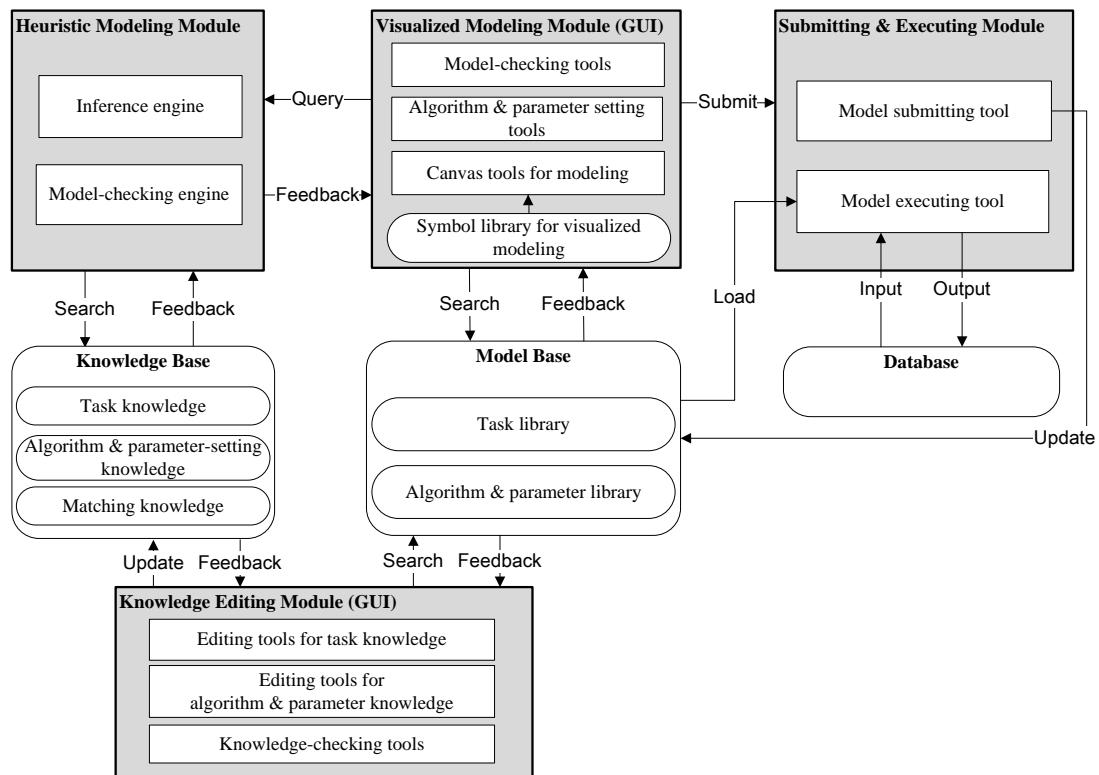


Figure 1. Framework of SimDTA VisModeler

Heuristic Modelling Module. With a inference engine inside this module it can be automatically determined to add a specific task into the current workflow in order to prepare a necessary, but not ready-made input data (fig. 2b, 2c). Thus, the workflow shown in the visualized modelling module can be interactively expanded from a user's initial target task to a complete DTA workflow in which all input data are ready for execution. Such modelling process of automated tracing from the last step to the first step is more natural for the non-expert user than the traditional modelling process of user-defining from the first step to the last step in existing DTA-assisted software. We call this feature of SimDTA VisModeler “heuristic modelling”.

Submitting & Executing Module. This module can save the user-specific DTA workflow into the model base and execute it, after the workflow is built and passed the model-checking in the visualized modelling module.

Knowledge-Editing Module. This module has a GUI by which the knowledge base can be interactively updated. This module is independent to other modules.

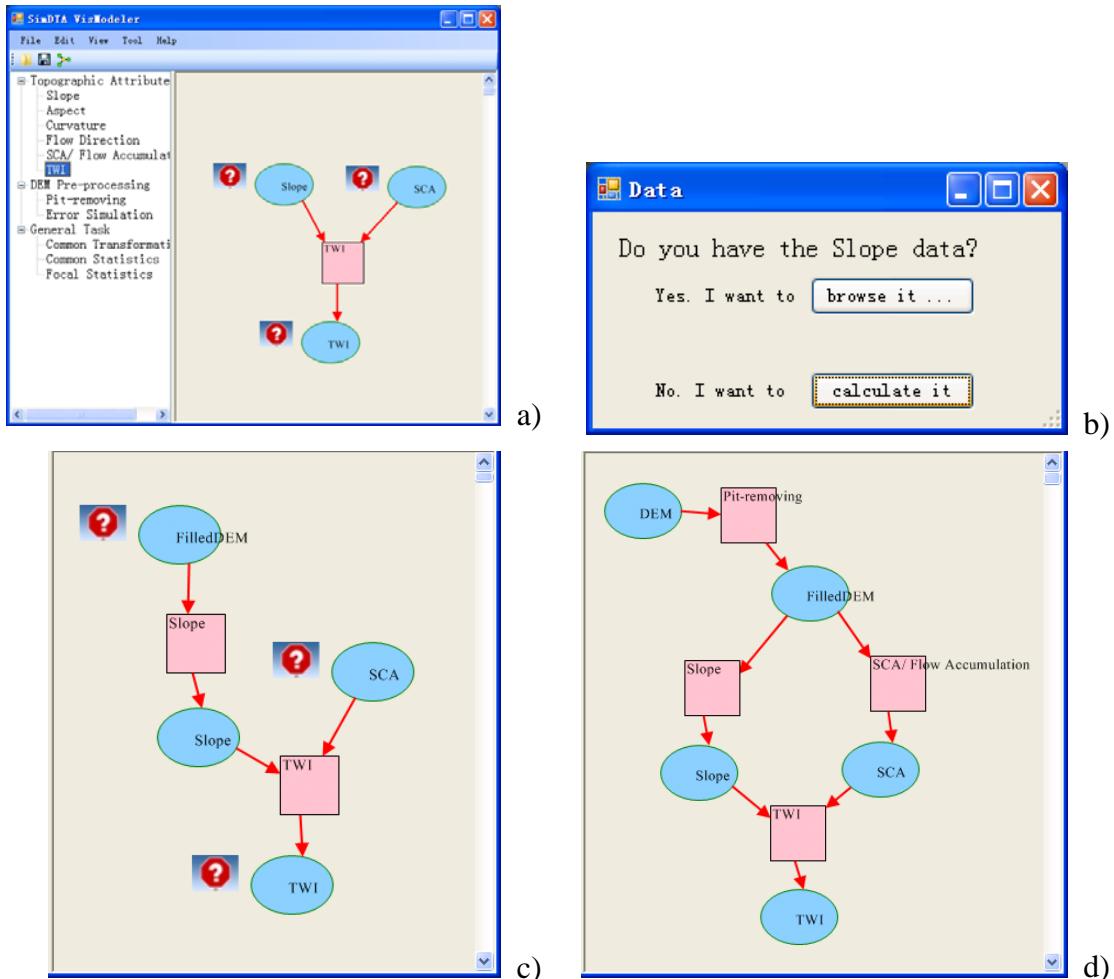


Figure 2. Modelling with SimDTA VisModeler. a) Drag a task to canvas (Question mark attached to the data-icon means that the data source is not specified yet); b) Interaction on whether the current workflow should be extended; c) A task is automatically added; d) A complete workflow for calculating TWI.

2.2 Preliminary implementation

SimDTA VisModeler is developed with Microsoft .NET 4.0 libraries. The visualization is developed based on .NET GUI library. The model executing tool is implemented by .NET reflection mechanism.

Currently the DTA knowledge is stored in XML⁵ files. The independent knowledge-editing module is under development.

3. Application

A simple case of modelling and executing of the calculation of topographic wetness index (TWI), one of the most important topographic attributes, is used to illustrate the availability of SimDTA VisModeler. If a user need calculation of TWI, not only a TWI algorithm but also the input data (two topographic attributes, i.e. slope gradient

⁵ <http://www.w3.org/XML/>

and specific contributing area) for the TWI algorithm are necessary for this TWI-calculating task (Hengl and Reuter 2008). As soon as the user drags the TWI task into the canvas, these requirements will be visualized (fig. 2a) and a specific TWI algorithm with default parameter settings will be chosen as default algorithm based on existing research on this algorithm. If the necessary input data is available, the modelling process could be ended with assigning input/output file names by user. Or else, additional tasks of calculating slope gradient and specific contributing area would be added in the user's DTA workflow (fig. 2b, 2c). The remaining process of modelling user's TWI-calculating workflow will be the analogy of the above process until all input data have been ready for execution (fig. 2d).

4. Conclusions

By the heuristic and visualized modelling environment proposed in this abstract, a non-expert user can handily construct a complete DTA workflow from his initial target task, even if the user has little knowledge on DTA tasks/algorithms or the relationship between them.

5. Acknowledgements

This study was supported by the Knowledge Innovation Programs of the Chinese Academy of Sciences (KZCX2-YW-Q10-1-5), and the National Natural Science Foundation of China (40501056).

5. References

- Gregersen JB, Gijsbers PJA and Westen SJP, 2007, OpenMI: Open modelling interface. *Journal of Hydroinformatics*, 9(3):175-191.
- Hengl T and Reuter HI (eds), 2008, *Geomorphometry: Concepts, Software, Application. Developments in Soil Science*, vol. 33. Elsevier.
- Lindsay JB, 2005, The Terrain Analysis System: a tool for hydro-geomorphic application. *Hydrological Processes*, 19:1123-1130.
- Lutz M, 2007, Ontology-based descriptions for semantic discovery and composition of geoprocessing services. *Geoinformatica*, 11(1):1-36.
- Takatsuka M and Gahegan M, 2002, GeoVISTA Studio: a codeless visual programming environment for geoscientific data analysis and visualization. *Computers & Geosciences*, 28(10):1131-1144.
- Tarboton DG, 1997, A new method for the determination of flow directions and upslope areas in grid digital elevation models. *Water Resources Research*, 33:309-319.
- Watson FGR and Rahman JM, 2004, Tarsier: a practical software framework for model development, testing and deployment. *Environmental Modelling & Software*, 19(3):245-260.
- Wilson JP and Gallant JC (eds), 2000, *Terrain analysis: principles and applications*. John Wiley & Sons, New York.
- Yue P, Di L, Yang W, Yu G and Zhao P, 2007, Semantics-based automatic composition of geospatial Web services chains. *Computers & Geosciences*, 33(5):649-665.

Interactive visualisation of spatial turnover

Shawn W. Laffan

¹School of Biological, Earth and Environmental Sciences,
UNSW, Sydney, Australia, 2052
Telephone: +61 2 9385 8093
Fax: +61 2 9385 1558
Email: Shawn.Laffan@unsw.edu.au

1. Extended abstract

Spatial turnover, the rate of change of a phenomenon or process with distance, is an important component of many research fields. Examples include species turnover (of which beta diversity is a specific type; Tuomisto, 2010), environmental gradients such as rainfall, income classes by census districts, language groups (Jones & Laffan, 2008), and pedodiversity. It forms the basis of simple correlogram plots and analytical methods such as Generalised Dissimilarity Modelling (Rosauer *et al.*, 2009a; Ferrier *et al.*, 2007) where one relates the turnover of species composition to the changes in environmental phenomena.

Turnover for continuous fields such as elevation or climatic variables are merely the gradients between pairs of locations, and can be easily inferred from displays of geographic layers in a GIS (e.g. Figure 1). However, the visualisation of turnover becomes more difficult for cases where one is investigating the turnover of collections of objects that overlap in geographic or other spaces. An important example of this is species diversity, where one is interested in the rate of change of species composition with increasing geographic distance and direction, as opposed to the rate of change of the absolute number of species. In such cases the composition cannot be directly displayed or stored as a single surface, making such an approach appropriate.

For compositional turnover one must use a matrix of turnover values, where the n rows and columns represent the set of locations used, and each of the values v_{ij} represents the turnover between the pair of locations represented by row i and column j . The visualisation of turnover matrices can be extremely complex because, for each location in an $n \times n$ matrix, there will be $n-1$ possible values that could be plotted and each of these relates to a different location in the data set. One approach is to plot lines radiating from each location i to each neighbour j , symbolised by their relative values. However, such a diagram rapidly becomes cluttered as n increases. Subsetting by dominant directions is possible, but could exclude multi-modal relationships, and subsetting to distance classes can result in interpreters missing insights where relationships span distance classes.

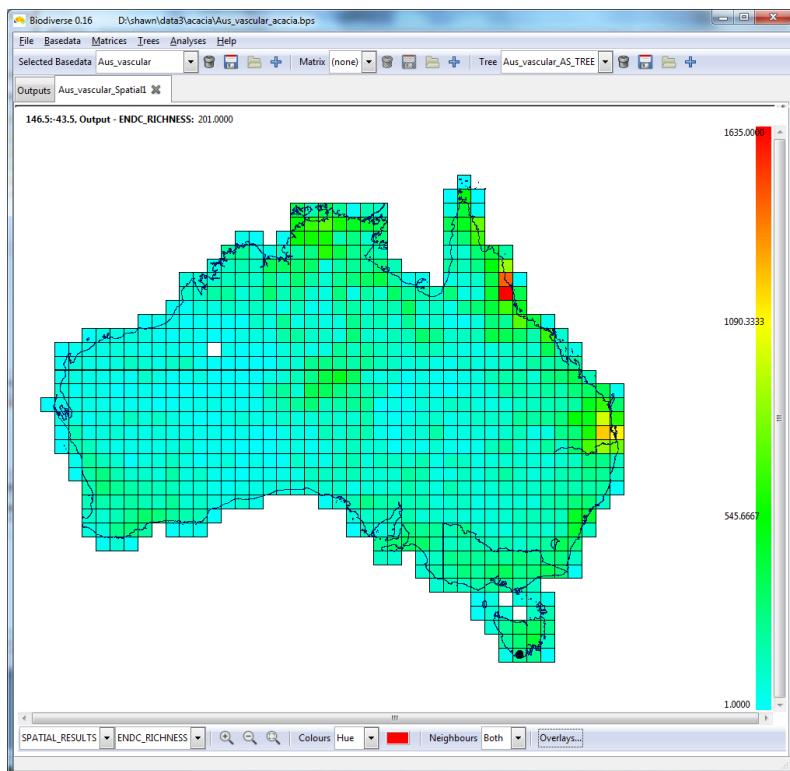


Figure 1. Species richness of Acacia species across Australia at a 1 degree resolution. The rate of change of richness is easy to infer, but not the relative changes in species composition. Data are derived from Laffan and Crisp (2003).

The simplest solution to the visualisation problem in this case is to subdivide the matrix into n geographic layers and plot each one separately. This process is comparatively simple to implement as a post-hoc process where each layer is extracted and plotted in sequence. However, an interactive approach is far more user-friendly, and working directly from the matrix is more computationally efficient. Such an approach has been implemented as an extension to the Biodiverse software (Laffan *et al.*, 2010; <http://www.purl.org/biodiverse>). In this system a matrix is calculated and then displayed for one index location, with the remaining locations plotted using their turnover values relative to the index location. When the user clicks on a new location it is set as the index location, and the display is updated to show the turnover from that location to all other locations using values extracted dynamically from the matrix. An example plot of four locations this is given in Figure 2, where one can observe the rate of change of Acacia species composition at a one degree resolution across Australia for a west to east transect through central Australia. Of particular note is that one can easily visualise for each location the spatial scale, anisotropy and non-stationarity that underlie analyses that use aggregate measures of turnover.

A further advantage of using Biodiverse is that one can analyse any one of the more than 150 currently supported scalar indices, so one can assess more than simply species turnover, for example species endemism (Laffan & Crisp, 2003), phylogenetic endemism (Rosauer *et al.*, 2009b), phylogenetic turnover, trait data, continuous fields, and the like. Biodiverse has been developed for complex spatial analyses, so one can also construct spatially constrained matrices such that the sets of neighbours considered for each

location can be restricted to a set radius around each location, or within a biome, or some other arbitrarily complex spatial condition.

2. References

- Ferrier, S., Manion, G., Elith, J. & Richardson, K. (2007) Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions*, **13**, 252-264.
- Jones, C. & Laffan, S. W. (2008) Lexical similarity and endemism in historical wordlists of Australian Aboriginal languages of the greater Sydney region. *Transactions of the Philological Society*, **106**, 456-486.
- Laffan, S. W. & Crisp, M. D. (2003) Assessing endemism at multiple spatial scales, with an example from the Australian vascular flora. *Journal of Biogeography*, **30**, 511-520.
- Laffan, S. W., Lubarsky, E. & Rosauer, D. F. (2010) Biodiverse: a tool for the spatial analysis of biological and other diversity. *Ecography*, **33**, 643-647.
- Rosauer, D., Ferrier, S., Manion, G., Laffan, S. W. & Williams, K. (2009a) Nice weather for frogs - using environmental data to model phylogenetic turnover. *10th International Conference on GeoComputation* (ed. by B.G. Lees & S.W. Laffan), Sydney, Australia.
- Rosauer, D. F., Laffan, S. W., Crisp, M. D., Donnellan, S. C. & Cook, L. G. (2009b) Phylogenetic endemism: a new approach to identifying geographical concentrations of evolutionary history. *Molecular Ecology*, **18**, 4061-4072.
- Tuomisto, H. (2010) A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography*, **33**, 2-22.

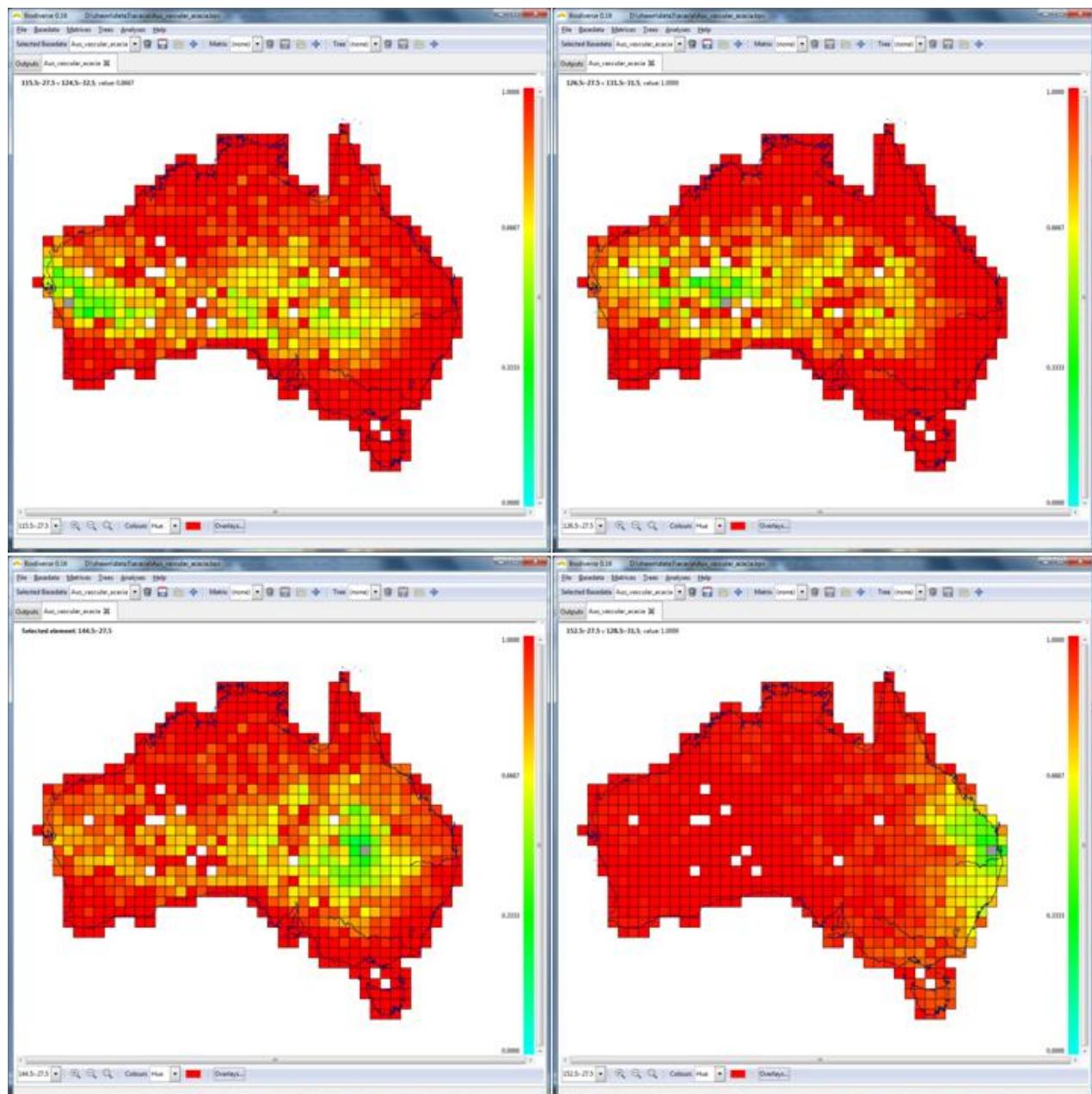


Figure 2. Spatial turnover plots for four cells along a west-east transect for Acacia species in Australia. Turnover is measured using the Sorenson dissimilarity metric. Index cells are denoted in grey. The colour scale progresses from cyan (most similar) to red (most dissimilar). Data are derived from Laffan and Crisp (2003).

Automatic Terrain Analysis in Railway Transportation Corridors with Regard to Asset Line-of-Sight from Monocular Video

Thomas Warsop

Research School of Informatics, Computer Science Department

Loughborough University, Loughborough, LE11 3TU, UK

Email: T.E.Warsop@lboro.ac.uk and S.Singh@lboro.ac.uk

1 Abstract

For the safe and continued operation of a railway environment, it is important railway assets such as signals, signs and level crossings are clearly visible to train drivers. Safety guidelines stipulate drivers' view of assets must be unobstructed for eight seconds upon approach (Railway Group Standards [2003]). Even if assets are clearly visible from prescribed distances at the point of initial construction, the environment in which they exist is dynamic. Thus the environment around an asset can change (e.g. vegetation can grow) which can cause unwanted asset line-of-sight (LoS) obstructions. Therefore, asset LoS must be regularly checked. Traditional methods for performing these checks either involve manual trackside labour (US Army Corps of Engineers [2007]) which present safety concerns with regard to personnel on the track or expensive laser scanning equipment (FLI-MAP [2010]). In this work, we present a system which creates a three-dimensional model of the environment surrounding an asset using monocular video data captured from a train mounted video camera.

Once this terrain model has been created, it must be analysed with regard to LoS between asset position and possible driver positions. Traditional methods of LoS analysis provide information regarding the inter-visibility of points within a terrain map and focus on reducing the computational load of such tasks (Salomon et al. [2004], Washtell et al. [2009], Duvenhage [2009]

and Y. Xia and Shi [2010]). Whilst this information is useful, it is important in the presented application to provide analysis of terrain, indicating which parts of the terrain cause dangerous LoS obstructions now or in the future. Hence, we present a novel metric to provide simultaneous LoS and terrain analysis for segmented terrain elements from the generated model. This metric combines information regarding the closest driver position which can be obstructed, using the smallest amount of modelled terrain *growth*.

The *minimum obstructing distance* of a terrain element is defined as the closest distance of all the observation points obscured by the terrain element:

$$MOD(T) = \min(D(asset, obs_i), \forall i \in OBS) \quad (1)$$

where $OBSTRUCT(T, obs_i) = true$

where, T is the terrain element under consideration, OBS is the set of all observation points, $D(asset, obs_i)$ returns the distance between the asset and observation points ($asset$ and obs_i respectively) and $OBSTRUCT(T, obs_i)$ returns true if the terrain element obstructs obs_i , otherwise it returns false. This concept is demonstrated in Figure 1.

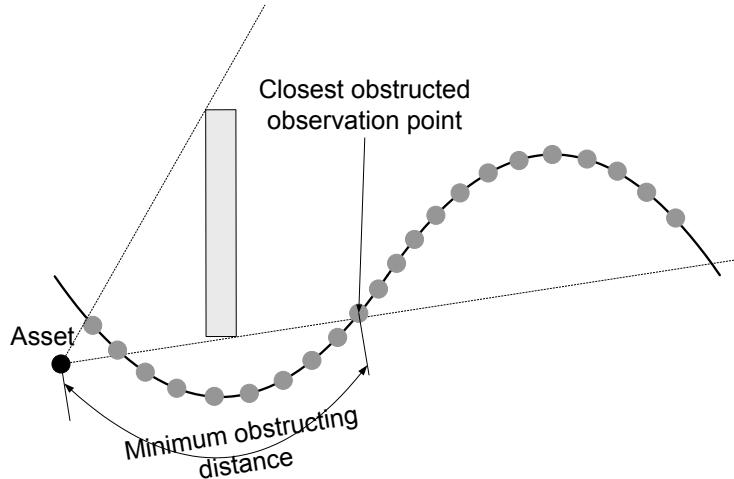


Figure 1: Minimum obstructing distance of a terrain element.

For any set of terrain elements, the lower this minimum obstructing distance is, the more important the associated terrain element is with regard to

asset LoS. Consider the example shown in Figure 2. Clearly, terrain element A should be removed before terrain element B, as the minimum obstructing distance is much closer for A. This concept also extends to groups of terrain elements which appear together. This is shown in the (albeit) simple example of Figure 3. The terrain element marked 1 has the lowest minimum obstructing distance and has the greatest impact on asset LoS. For example, even if all other terrain elements are removed, the asset LoS will still be blocked by terrain element 1.

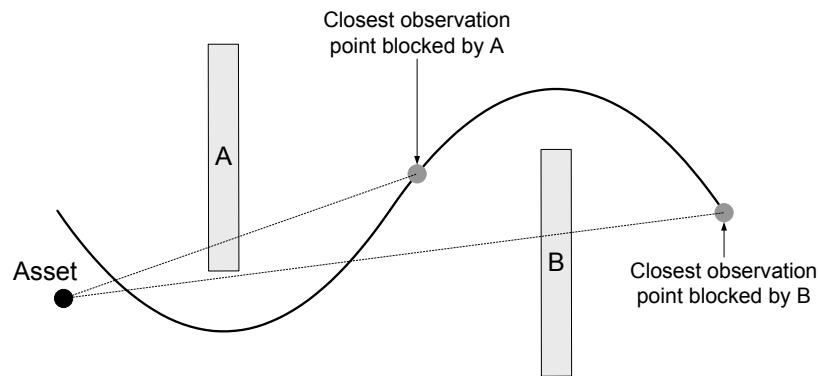


Figure 2: Comparing minimum obstructing distance.

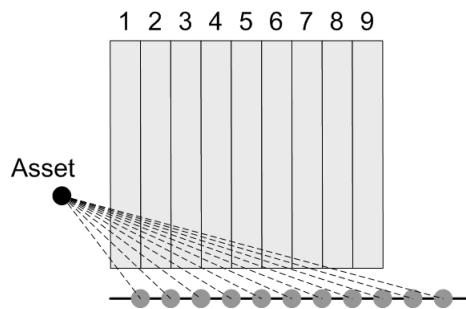


Figure 3: Minimum obstructing distance in terrain groups.

Not all terrain elements will obscure observation points and hence will have no associated minimum obstruction distance. However, it is possible to compute what the minimum obstruction distance would be in the future.

This growth is computed as the minimum distance between the terrain element and the set of lines created between the asset point and observation points:

$$GROWTH(T) = \min(D_{LINE}(T, \overrightarrow{AssetObs_i}), \forall i \in OBS) \quad (2)$$

where, $D_{LINE}(T, \overrightarrow{AssetObs_i})$ returns the distance between terrain element T and the line created by the asset point ($Asset$) and the current observation point under consideration (Obs_i) and OBS is the set of all observation points. This concept is shown in Figure 4, where the terrain element (white rectangle) would require to grow horizontally the distance demonstrated by *Growth* to block LoS from the observation point (grey circle). The minimum obstructing distance is then the distance from the asset associated with this blocked observation point.

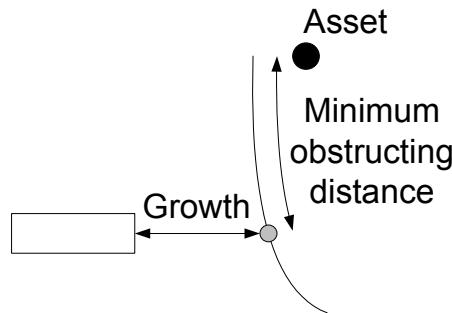


Figure 4: Terrain element growth and minimum obstructing distance.

The lower this growth value, the more important the corresponding terrain element with respect to asset LoS. This is highlighted in Figure 5 where terrain element 3 is the most important as it requires the least amount of growth to intersect with the LoS between the asset and an observation point.

Computing the minimum obstruction distance and terrain growth for each terrain element, it is possible to produce a ranking. This ranking sorts the elements so that those which cause the closest obstruction, the soonest are ranked higher. This is achieved by sorting terrain elements into ascending order in terms of computed values of growth and the minimum obstruction distance. Such an example is shown in Figure 6 and Table 1, the numbers represent the ranking of the terrain element - lower means the terrain element is more important.

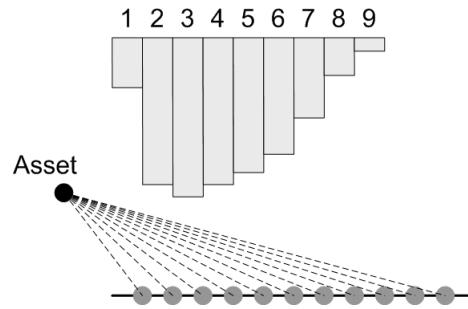


Figure 5: Terrain element growth example.

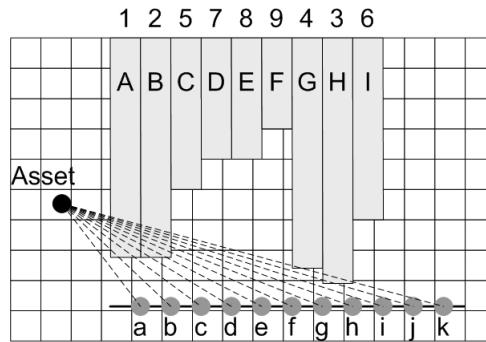


Figure 6: Terrain ranking example

However, the calculation of this metric for a given terrain model can be computationally expensive. Following the lead from previous literature, we therefore present a further novelty for decreasing computation associated with our proposed metric. The main idea behind this reduction divides single (expensive) three-dimensional based computations into a series of two-dimensional problems, each of which can be solved using binary search techniques. Several different such techniques are compared in this work. The best performing of these utilises an Adelson-Velskii and Landis (AVL) balanced binary tree and persistent AVL tree pair, decreasing computation to 18% of an exhaustive method whilst producing results which are 99.8% the same.

The core output of this work is generated reports presenting the terrain and line-of-sight analysis for use by railway engineers to perform maintenance

Terrain	MOD	Growth	Rank
A	b	-	1
B	d	-	2
C	k	1.5	5
D	k	2.5	7
E	k	3.0	8
F	k	4.5	9
G	k	-	4
H	j	-	3
I	k	2.0	6

Table 1: Terrain ranking data for Figure 6.

(or preventative maintenance) with regard to the terrain surrounding an asset, to help improve asset line-of-sight. The following presents example results used in the construction of such a report. Figure 7 shows an asset and some of the image frames from the video sequence preceding this location. Figure 8 shows line-of-sight and terrain analysis results for this sequence, in a top-down view of the corresponding terrain model.

References

- B. Duvenhage. Using an implicit min/max kd-tree for doing efficient terrain line of sight calculations. In *Proceedings of the 6th International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa*, pages 81–90, 2009.
- FLI-MAP. 2010. URL <http://www.flimap.com/site4.php>.
- Railway Group Standards. Signal positioning and visibility. In *Railway Group Standard*, December 2003. P. Woolford and A. Blakeney.
- B. Salomon, N. Govindaraju, A. Sud, R. Gayle, M. Lin, and D. Manocha. Accelerating line of sight computation using graphics processing units. In *In Proceedings of the 24th Army Science Conference*, pages 1–5, 2004.
- US Army Corps of Engineers. Chapter 8 - total station topographic survey procedures. In *Engineering and Design*

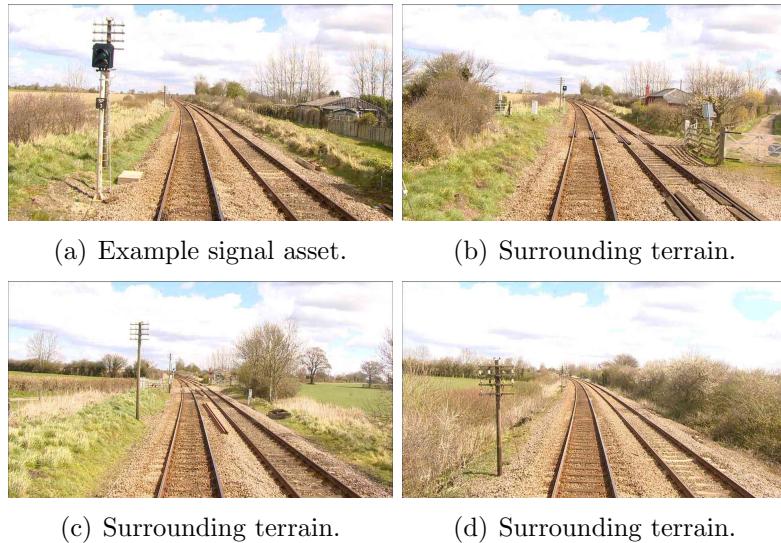
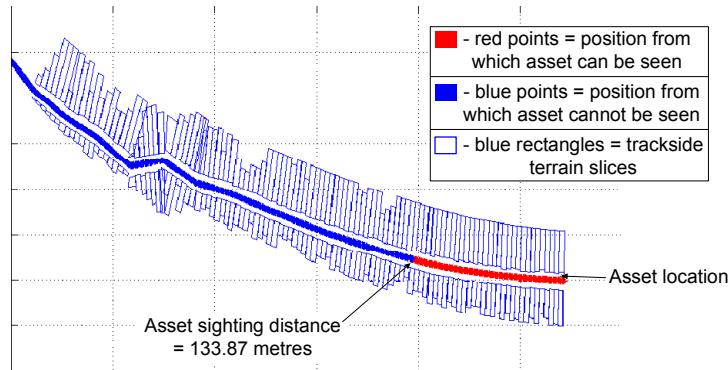


Figure 7: Example asset (the signal to the left-hand side of the track) and image frames containing the surrounding terrain.

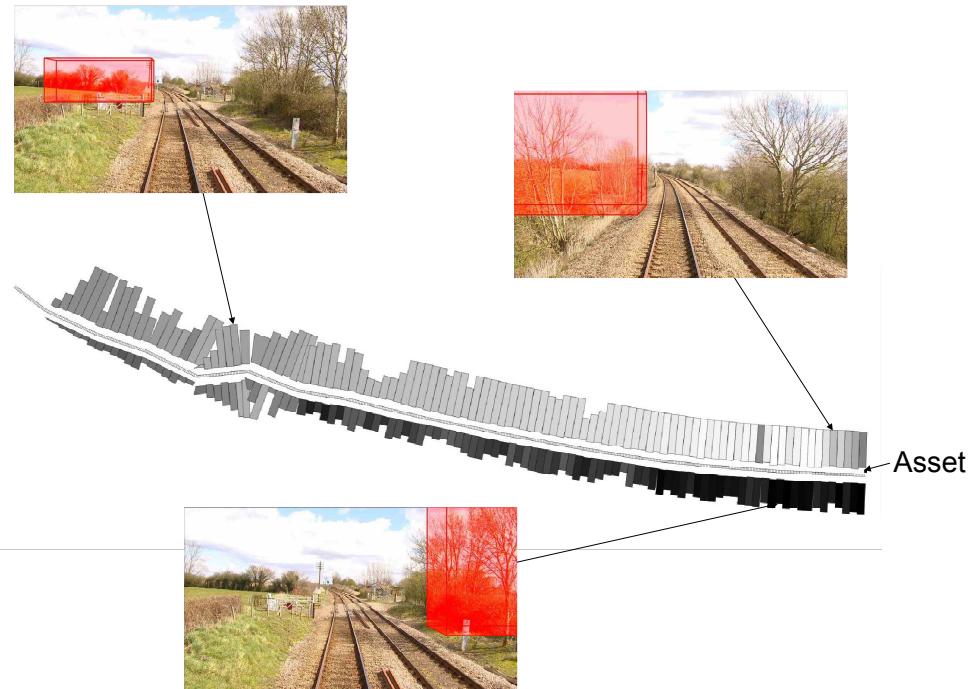
- *Control and Topographic Surveying*, January 2007. URL <http://140.194.76.129/publications/eng-manuals/em1110-1-1005/c-8.pdf>. Last accessed: 24th December.

J. Washtell, S. Carver, and K. Arrell. A viewshed based classification of landscapes using geomorphometrics. In *Proceedings of Geomorphometry*, pages 44–49, 2009.

Y. Li Y. Xia and X. Shi. Parallel viewshed analysis on gpu using cuda. In *Third International Joint Conference on Computational Science and Optimization*, pages 373–374, 2010.



(a) Example asset line-of-sight profile presented in a *top-down* view of the track. Asset sighting distance refers to the distance from which an unobstructed view of the asset is provided upto the location of the asset.



(b) Terrain analysis with respects to asset line-of-sight, the trackside terrain elements are coloured according to effect on asset LoS - white means the terrain element has more effect on asset line-of-sight, black means less.

Figure 8: Example output providing analysis of asset and terrain line-of-sight.

Model Selection in GWR: the Development of a Flexible Bandwidth GWR

Wenbai Yang, A. Stewart Fotheringham, and Paul Harris

National Centre for Geocomputation, National University of Ireland Maynooth
Telephone: +353 (0) 1 708 6731
Fax: +353 (0) 1 708 6456

1. Introduction

Model selection is a key issue in geographically weighted regression (GWR; Fotheringham et al. 2002). This not only includes the selection of a variable subset, the bandwidth size, the type of kernel, but also the form of the GWR model itself. In this study, we investigate this issue with respect to basic GWR and semi-parametric or mixed GWR (MGWR; Brunsdon et al. 1999; Fotheringham et al. 2002; Mei et al. 2004), where for the latter, some relationships are modelled as stationary across space, whilst others are not. However, this mixed model can be seen as a special case of a more general model, where different bandwidths are defined for different independent variables. This flexible bandwidth GWR (FBGWR) model should not only allow an extremely useful exploratory investigation of data relationships that may vary at different spatial scales, but it is hypothesised that its results may guide model selection with more parsimonious GWR and MGWR fits.

2. Model selection

A multiple linear regression (MLR) model describes stationary relationships between dependent and independent variables throughout the study area. When relationships vary over space, a GWR model is preferred, which allows parameters to be estimated locally. In this study, a stepwise Akaike Information Criterion (AIC) method is used for variable subset selection with MLR, whereas for GWR, an approximate stepwise AIC procedure is employed. Model selection with MGWR involves not only whether or not to include a variable, but also whether it should vary spatially. There are two practical methods to help this decision; one uses AIC, the other uses a Monte Carlo test with the basic GWR model. The Monte Carlo test evaluates the variability of a local parameter estimate of a given variable; if it varies significantly across space, then this variable should be a geographically varying term, otherwise, it should be a fixed term in the ensuing MGWR fit.

3. Case study data

We highlight model selection issues using a dataset of the Irish Famine, where relationships between population decline and thirteen demographic, locational and land use characteristics (see table 1) are investigated (Fotheringham et al. 2010). The dataset consists of 3,250 Electoral Divisions (EDs) and were derived from the 1841 and 1851 Population and Agricultural Censuses. To reduce computational complexity with MGWR (and FBGWR), whilst at the same time preserving the pattern of relationships across the country, a stratified sample containing 446 EDs is drawn to act as our study dataset. Datasets are shown in fig. 1.

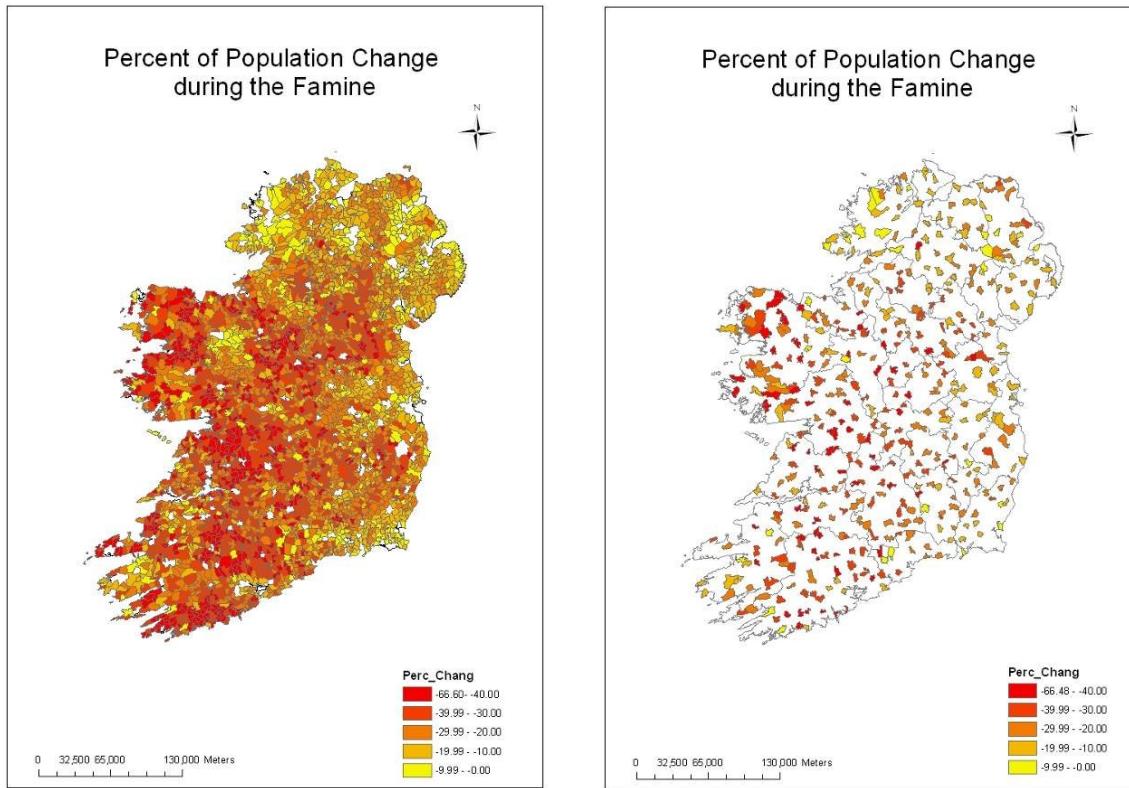


Figure 1. Full (left) and sample (right) datasets.

4. Analysis

The many paths for model selection with GWR and MGWR are depicted in fig. 2. Starting from a full set of independent variables (leftmost path), we can fit a basic GWR model using all variables and then find two competing MGWR models via an AIC or Monte Carlo approach. Alternatively (the middle path), we use AIC with GWR to find a reduced set of independent variables and then find a further two MGWR models via AIC or Monte Carlo. This provides four possible MGWR fits, which are then doubled to eight, if we start from a reduced set of variables, resulting from a stepwise MLR fit (rightmost path).

There is no one best path: for example if we follow the path on the far right, we may exclude variables that only have locally significant coefficients. If, as an example, we focus on the results from the four MGWR paths shown in fig. 2, we are presented with clear inconsistencies (see table 1).

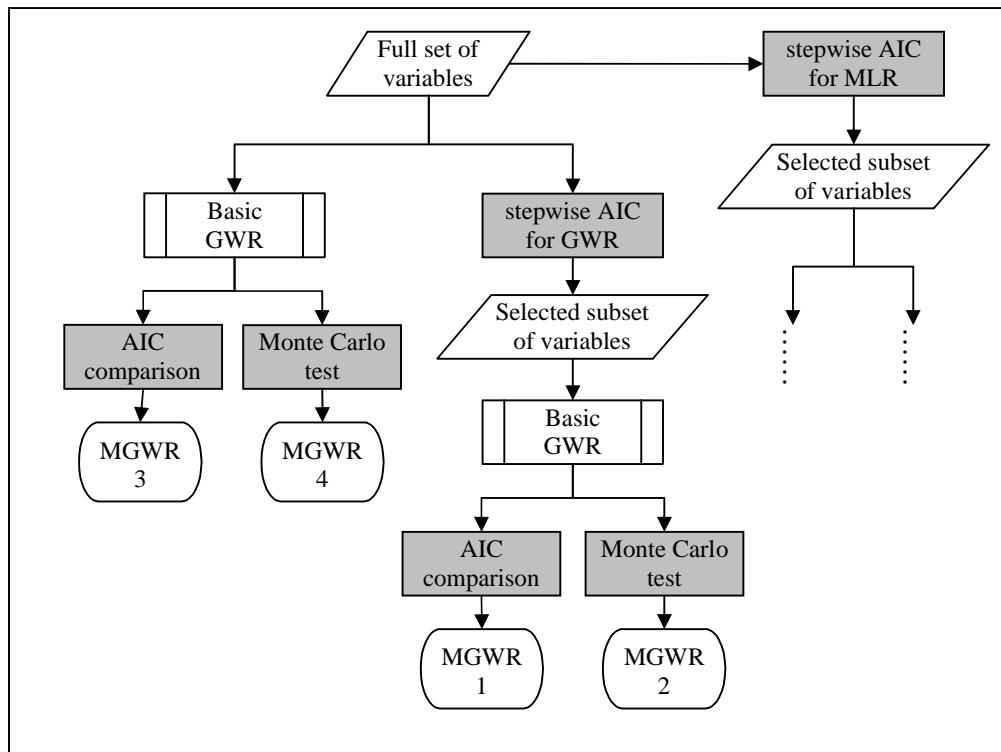


Figure 2. Paths for model selection.

Furthermore, if we focus on MGWR1 and MGWR2, more mixed models can result. For example, using the AIC approach MGWR1 results in a first pass suggesting that the parameters for PopCrop_41 and Potat_Cult have no significant spatial variability (call this MGWR1-1). Then if we conduct a further AIC-based selection with MGWR1-1, a new mixed model (MGWR1-2) can be found which suggests that the parameter of Perc_Town also has no significant spatial variability (i.e. three variables should be fixed). Alternatively, the Monte Carlo route via basic GWR indicates that Perc_Town and InWHOUSE should be the fixed terms (i.e. the MGWR2 model). This model can also be specified with different bandwidths: MGWR2-1 selects its bandwidth through AIC minimisation; MGWR2-2 employs the same bandwidth as GWR1; while MGWR2-3 employs the bandwidth of MGWR1-1.

Performances of the five MGWR models are compared in table 2, along with the corresponding basic GWR (GWR1) and MLR models. MLR is the poorest performer, whilst according to AIC, MGWR1-2 performs the best. The three MGWR models built on the Monte Carlo test, perform no better than basic GWR.

As an example, the local parameter estimate surfaces of VALUATION are compared in fig. 3. Here, only the parameter estimates that are significantly different from zero (indicated by a t-value in excess of ± 1.96) are displayed. As a comparison, the global parameter estimate is 0.0015, with a t-value of 5.52, suggesting a positive relationship between population decline and land value. Here, a positive relationship means population decline was more severe in areas where the corresponding variables had lower values. A negative relationship has the opposite meaning. GWR1 detects some areas with significant negative effects in the middle of the country which does not happen with the mixed models. Among the mixed models, MGWR1-2 reveals the most variation of parameter estimates, while MGWR1-1 masks the areas with significant effects in the north-west of the country.

Variable	Model	MGWR1	MGWR2	MGWR3	MGWR4
(Intercept)	Varying	Varying	Varying	Varying	Varying
VALUATION (land value per hectare)	Varying	Varying	Varying	Varying	Fixed
Coast_Dist (distance to coast)	Varying	Varying	Varying	Varying	Varying
Perc_Towns (percentage of population in towns)	Varying	Fixed	Fixed	Fixed	Fixed
ACC41_20 (accessibility to urban areas)	Varying	Varying	Varying	Varying	Varying
lnWHOUSE (ln proximity to workhouses)	Varying	Fixed	Varying	Fixed	Fixed
PopCrop_41 (population per acre of cropped land)	Fixed	Varying	Varying	Fixed	Fixed
Potat_Cult (percentage of cropped land under potatoes)	Fixed	Varying	Fixed	Varying	Varying
UNINHABPCT (percentage of uninhabited dwellings)	-	-	Fixed	Fixed	Fixed
lnPPB (ln persons per building)	-	-	Varying	Fixed	Fixed
PCorn_Cult (percentage of cropped land under grain)	-	-	Varying	Fixed	Fixed
Ratio1841 (male/female population ratio)	-	-	Varying	Fixed	Fixed
Crops_Hold (average holding size)	-	-	Varying	Varying	Varying
MEAN_ELEV (mean elevation)	-	-	Fixed	Fixed	Fixed
Optimal adaptive bandwidth	72	107	172	118	

Table 1. Model selection results from different path.

	MLR	GWR-1	MGWR1-1	MGWR1-2	MGWR2-1	MGWR2-2	MGWR2-3
AIC	3383.44	3240.38	3233.64	3229.29	3281.13	3283.78	3301.50
BIC	3419.93	3547.68	3528.66	3483.85	3492.80	3530.59	3595.54
R square	0.18	0.61	0.61	0.58	0.49	0.52	0.54
Adjusted R square	0.17	0.49	0.49	0.48	0.40	0.41	0.41
Bandwidth		89	72	72	107	89	72
			<i>Based on AIC comparison</i>		<i>Based on Monte Carlo test</i>		
	MLR	GWR			Mixed GWR		

Table 2. Performances of different models (AIC is actually AIC corrected)

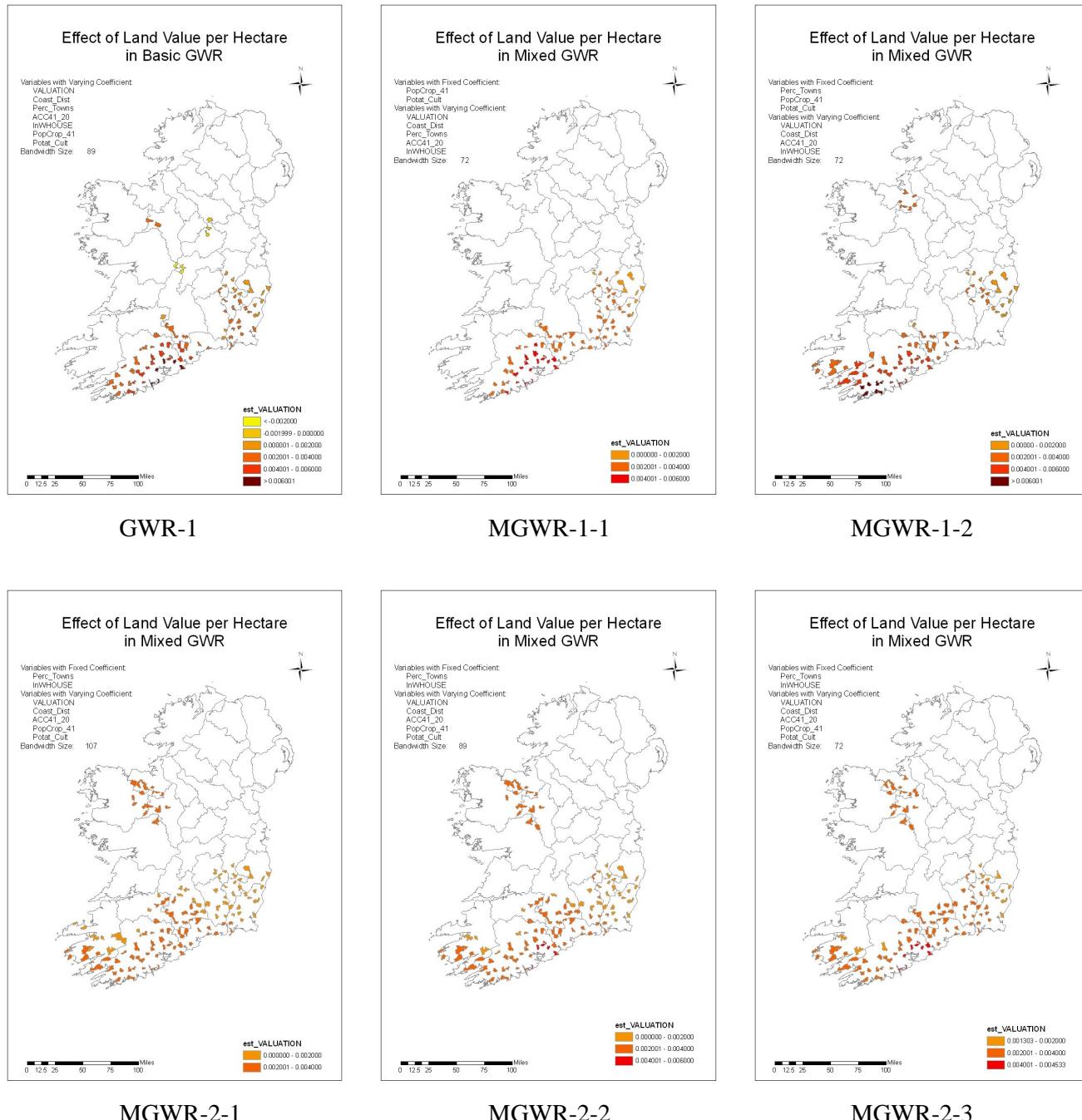


Figure 3. Parameter estimates for VALUATION from GWR and MGWR models.

5. FBGWR

Our results indicate clear problems in selecting a GWR model, whether in basic or mixed form. It is proposed that the results from a FBGWR model may aid model selection in these simpler models, of which FBGWR is a generalisation. On calibrating a FBGWR model, variables that have relatively large bandwidths would suggest that these should be fixed in a mixed model, whilst possibly, variables that have very small bandwidths should be omitted altogether.

Ideally, a fully developed and proven FBGWR model should enable a full investigation of data relationships that vary at different spatial scales. However such a model is inherently complex and its calibration is likely to present a significant computational burden. In this respect and as a first step in the development of FBGWR, we present an approximate form of this model using back-fitting and we use this model as a diagnostic tool for the simpler GWR models described. The details and validity (including its accuracy) of this model will be expanded upon in due course.

6. Acknowledgements

Research presented in this article was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support.

7. References

- Brunsdon, C., Fotheringham, A.S. & Charlton, M., 1999. Some Notes on Parametric Significance Tests for Geographically Weighted Regression. *Journal of Regional Science*, 39(3), pp.497-524.
- Fotheringham, A.S., Brunsdon, C. & Charlton, M., 2002. *Geographically weighted regression: the analysis of spatially varying relationships*, John Wiley and Sons.
- Fotheringham, A.S., Kelly, M. & Charlton, M., 2010. *The Demographic Impacts of the Famine: Towards a Greater Geographical Understanding*. Available at: <http://ncg.nuim.ie/content/projects/famine/Index.htm>.
- Mei, C.-L., He, S.-Y. & Fang, K.-T., 2004. A Note on the Mixed Geographically Weighted Regression Model. *Journal of Regional Science*, 44(1), pp.143-157.

Spatial planning – an inverse problem?

Ricardo Crespo¹, Adrienne Grêt-Regamey²

¹Swiss Federal Institute of Technology,
Zurich
Email: crespo@nsl.ethz.ch

²Swiss Federal Institute of Technology,
Zurich
Email: gret@nsl.ethz.ch

1. Introduction

This study aims at introducing a new approach for sustainable spatial planning, where the point of departure is not current data, but a future desired by stakeholders. To this end, we propose an inverse modeling approach where the result is a set of values for parameters identified as being key to reach a desired future. The goal of the preliminary assessment is to show, what kind of planning questions can be answered using the approach, e.g. (i) given the current state-of-the-art of modeling future spatial states, are the spatial developments desired by the stakeholders reachable?, (ii) where should the desired development take place?, and (iii) what are the trade-offs between the different solutions? We discuss advantages and shortcomings of the approach for planners and conclude about the effectiveness of the approach as a means of encouraging lay people and stakeholders to get involved efficiently in sustainable spatial development issues.

2. The inverse approach

Solving an inverse problem is not solving a mathematical exercise, but solving a problem coming from a scientific model. The phenomenon that relates the parameters to the observations can be represented by an operation G linking the space of observations d with the space of parameters m . While the set of parameters is usually called the model space (m), the set of observed data is called the data space (d). Aster et al. (2005) describe an inverse problem as shown in Equation 1:

$$d = G(m_{\text{true}}) + \varepsilon \quad (1)$$

where d may be a function of time and/or space or a collection of discrete observations. $G(m_{\text{true}})$ a “perfect” experiment and ε a noise component. d_{true} exactly satisfies for m equal to the true model, m_{true} , if we assume that the forward model is exact, ε can often be neglected as it often has little or no correspondence to m_{true} . The goal of inverse modeling is to search a solution such that $G(m_{\text{true}})$ becomes close to d . Thus, we make inferences about the parameters from the set of observed data thus fitting the model to the data.

3. The case study: Planning for new dwellings in metropolitan area

3.1 The model and the study area

The metropolitan area of Zurich (Switzerland) is one of the Europe’s economically strongest area and Switzerland’s economic centre. As a hot spot of living, working, and commuting, the metropolitan area of Zurich is characterized by high density and high development dynamic. We illustrate how inverse modeling can support decision-makers for identifying urban development options in the canton Zurich. Given a desired house price level, we show how one can determine what are the relevant trade-offs between locational, structural, and socioeconomic characteristics of new dwellings given a certain price.

Our forward model is based on a hedonic house price model employed by Loechl and Axhausen (2009). The underlying data for the model were taken from a webpage including rent offers from various Swiss real estate online platforms between December 2004 and October 2005. The addresses for all dwelling units in the dataset were geocoded at building level and matched with a wide set of spatial variables. Overall, rent prices are regressed against a set of structural, locational and socioeconomic explanatory variables. The forward problem can be solved by Ordinary Least Square (OLS) by which beta parameters are estimated and the expected value of rents can be estimated for a given set of explanatory variables. Conversely, in the inverse problem, a set of explanatory variables are to be found for a desired house price level and a given set of beta parameter estimates. For example, a desired house price level may be selected so as to compensate the economic loss caused by variables which are generally negatively related to house prices, such as: air noise level, population density, and proportion of foreigners. Figure 1 illustrates the forward and the inverse model framework for the hedonic house price model.

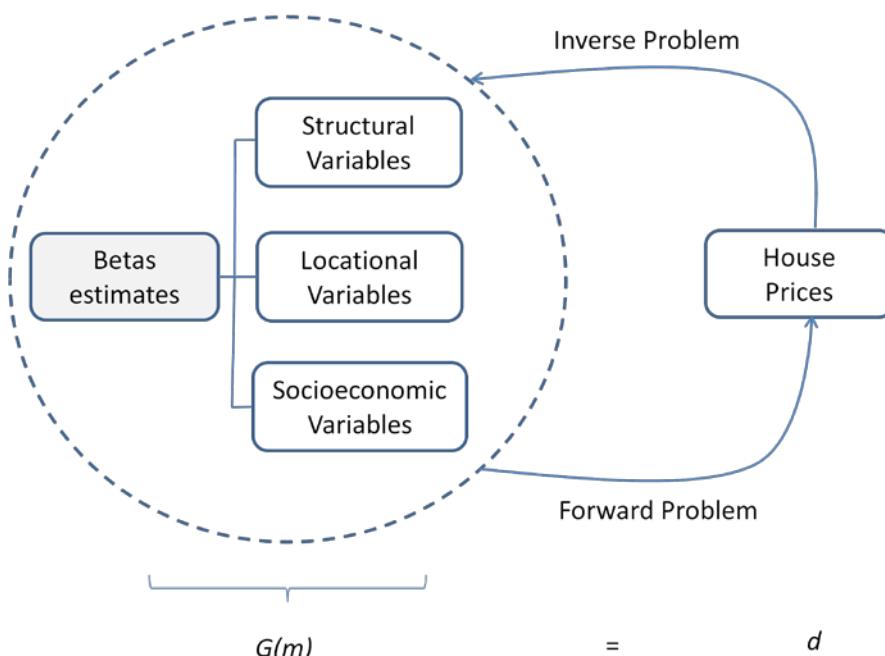


Figure 1. Framework illustrating the link between the forward and the inverse modeling approach.

3.2. Methodology: a mixed-GWR approach

As we intend to explore how compensation scheme between the different variables differ across space, we used a local regression analysis as the forward model. However, it may occur that some explanatory variables do not exhibit sufficient local variability, yielding misleading local parameter estimates associated with such variables. For this reason, we perform a mixed geographically weighted regression (GWR) model (Fotheringham et al. 2002) which allows some parameters to vary over space while others remain fixed. The selection of which variable will be globally or locally considered in the model can be done simply by analyzing the variability of the explanatory variables over space as well as the sign and t-values of the local estimates from GWR.

The functional form of the model is represented in Equation 2:

$$P = X_a \beta_a + X_b \beta_b + \epsilon \quad (2)$$

where P is a vector of the monthly rents prices X_a and X_b are the matrices of explanatory variables associated with global and local coefficients respectively, β_a is a vector of global coefficients (without including the intercept term), β_b is a matrix of location-specific coefficients (including the intercept term), and ϵ is a vector of residuals assumed to be random and spatially uncorrelated.

The next step is to **invert** our model in order to determine the values of chosen explanatory variables for a given rent value. As stated above, our intention is to explore the compensation scheme over space, thus we invert Equation 2 so as to solve the model for the X_b matrix as shown in Equation 3, where P^d denotes a “desired price” for which X_b is to be solved. Following the notation used in Equation 1, we express our inverse problem as:

$$G^{-1}(P^d, X_a, \hat{\beta}_a, \hat{\beta}_b) = X_b \quad (3)$$

where $\hat{\beta}_a$ and $\hat{\beta}_b$ are the mixed-GWR estimates associated with global al local coefficients respectively.

3.3 The air noise problem

It is generally argued that the locational variable air noise is negatively correlated to house prices. In our model, we use a global dummy variable denoted by AIRNOISE which is equal to 1 if the air noise level is above 52 dB and 0 otherwise. The value of the AIRNOISE parameter estimate is -85.02 [CFH] which represents the extent to which monthly rent prices drop as a result of high level of air noise in the area. Figure 2 shows the locations where the air noise level exceeds the 52 dB (referred to as *noisy locations*) coupled with the spatial distribution of the sampled data.

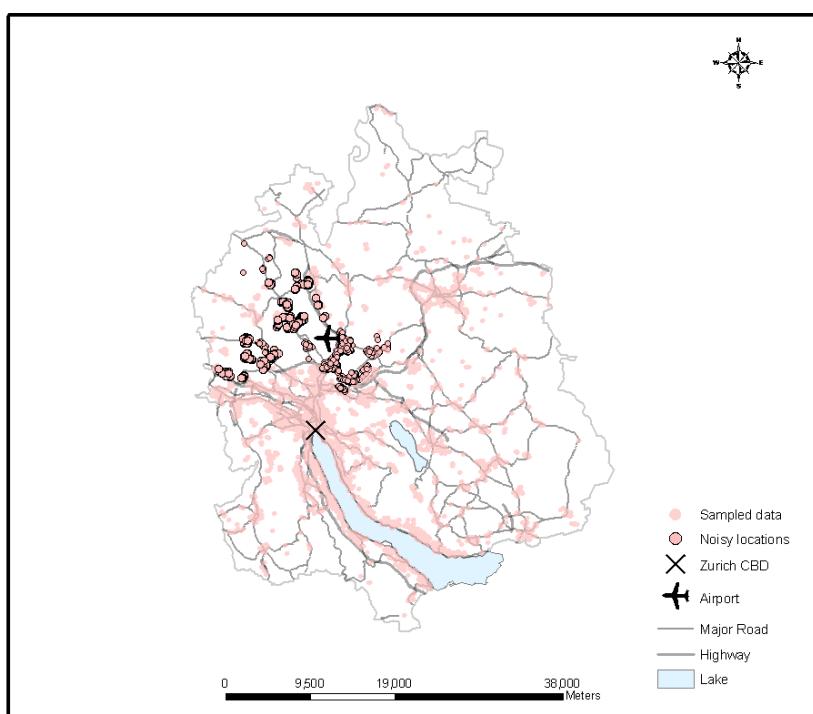


Figure 2. Noisy locations and sampled data.

As a way to demonstrate how our inverse approach operates, we solve Equation 3 for a local variable used in the hedonic house price model which we denote by CARTT_CBD (employed in a logarithmic form in the model) and which measures the average travel time to the Zurich CBD by car in minutes. In other words, our motivation is to find out the extent to which the CART_CBD has to be reduced to compensate the economic loss caused by the air noise (85 CHF monthly). To explore how the compensation scheme varies across the existing noisy locations, we define three clusters from the existing noisy locations, as shown in Figure 3.

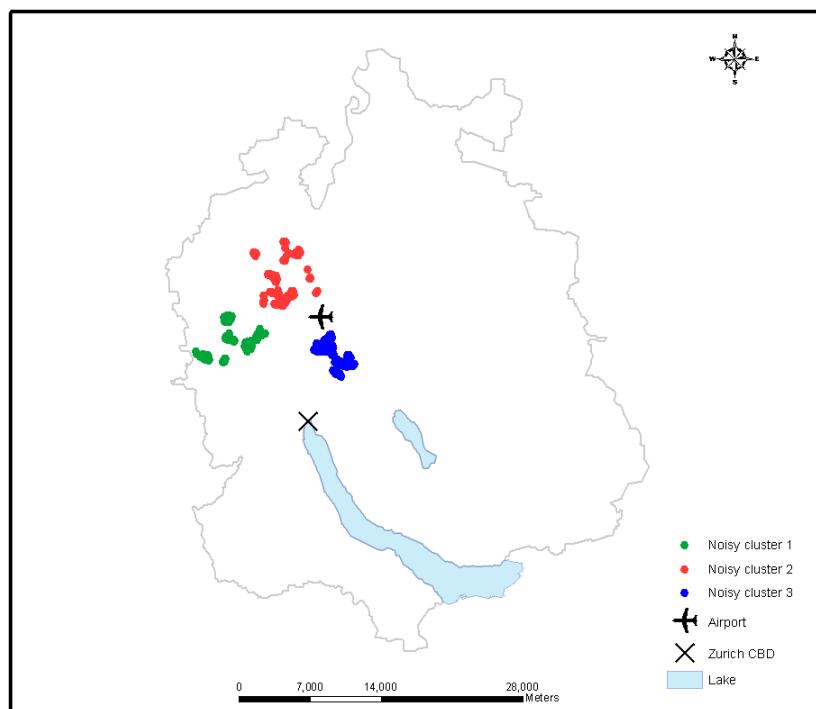


Figure 3. Clusters of noisy locations.

Figure 4 exhibits the results of the inverse modeling approach for the CARTT_CBD variable. Colored circles were drawn proportionally to show the extent to which the accessibility to the Zurich CBD must be improved in order to compensate for the 85 CFH loss caused by the air noise. As can be seen, a significant difference can be observed between cluster 1 and cluster 3, which can be explained because cluster 3 is closer to the Zurich CBD than cluster 1. In addition, it might occur that the workplace of most of inhabitants from clusters 1 is not located by the Zurich CBD, as it may be in case of cluster 1, so that the rent of properties in cluster 1 would respond more slowly to reductions in the CARTT_CBD variable than the rent of properties in cluster 3.

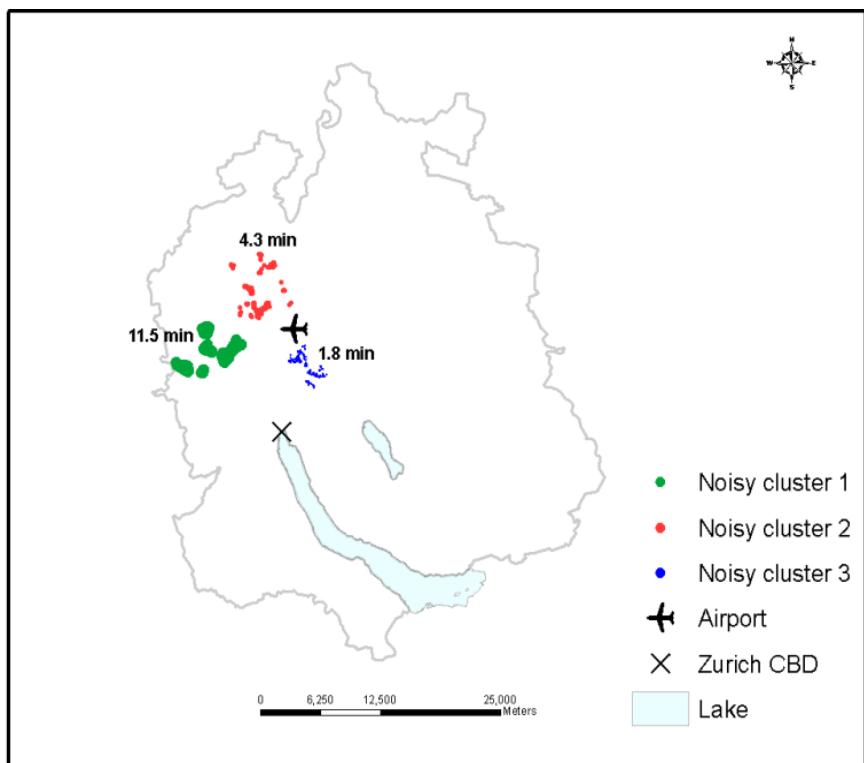


Figure 4. Representation of the optimal value of the CART_CBD variable for each cluster.

4. Final remarks

The preliminary results illustrate in a simply manner how the inverse modeling can be employed in planning decision-making processes. It is however worth pointing out that more complex planning decisions can be made by including more variables in the inverse model analysis. By doing this, we can investigate the trade-offs between different solutions and alternatives for sustainable spatial planning.

5. References

- Aster R, Borchers B and Thurber C (2005). *Parameter Estimation and Inverse Problems*. Elsevier Academic Press.
- Fotheringham S, Brunsdon C and Charlton M (2002). *Geographically Weighted Regression: the analysis of spatial varying relationships*. John Wiley & Sons Ltd.
- Loech M and Axhausen K (2010). Modeling hedonic rent prices for land use and transport simulation while considering spatial effects. *Journal of Land Use and Transport*. (In Press)

A spatial analysis of perceptions of health services accessibility, health status and geographic access using GWR

Alexis COMBER^{1*}, Chris BRUNSDON¹ and Robert RADBURN²

¹Department of Geography, University of Leicester, Leicester, UK;
* ajc36@le.ac.uk

²Leicestershire County Council, Leicester, UK

Introduction

This research uses a GWR analyses to compare stated attitudes to health service accessibility with measures of geographical access to those services. There is much concern over the effects of reductions public services, including health, in the UK and elsewhere, as result of central government cuts in funding. This research seeks to unpick the some of dimensions associated with access and service accessibility, including public perceptions of access, geographical access, and how they vary spatially as well as within and between different socio-economic groups and with health status.

Background and Method

The Place Survey was launched in 2008 by the Department of Communities and Local Government in the UK. It reports National Indicators and was collected by local authorities who used random sampling to select potential respondents. The DCLG commissioned the survey in order to capture public opinions (satisfaction) about local authorities and local services. The DCLG specified most of the the questions although local authorities were able to add their own questions in their area. The survey by Leicestershire County Council conducted on behalf of the DCLG in 2009 included a set of questions related to service accessibility, by capturing opinion over access to range of different public services.

Respondents were asked to indicate the ease / difficulty of access to a range of different services: "From your home, how easy is it for you to get to the following using your usual form of transport?" Services over which attitudes over accessibility were sought included general practitioners, dentist, pharmacies and local hospitals. Additionally the Leicestershire implementation of the survey captured information related to respondents' current health status, whether they have any long-standing illness, disability or infirmity. In Leicestershire there were 8530 responses to the Place Survey, with 415 / 8530 indicating dissatisfaction (replying either 'dissatisfied' or 'very dissatisfied') over access to GPs, 393 / 8530 stating that they had bad or very bad health (henceforth 'bad health') and 2824 / 8350 indicating that they had long term illness.

The Output Area Classification was developed by Vickers and Rees (2007). It allocates each output area to one of 7 groups (further subdivisions exists which were not considered in this analysis). Based on the Output Area they fell in, Place Survey responses were allocated to an OAC class. Geographic distances were calculated to the nearest GP surgery from the post-code of each respondent.

This analysis used a logistic GLM to explore the relationship between dissatisfaction over access to GP surgeries, health status and geographic distance. It then used a GWR analysis to explore the spatial non-stationarity and variability in respondent dissatisfaction over access to GPs with these variables.

Results

The GLM showed that Long Term Illness, Bad Health and Distance to the nearest GP surgery were significant predictors of dissatisfaction over access, whilst OAC classes were not (Figure 1). Analysis of the exponential of the coefficient estimates (in Figure 1) calculates the odds ratios associated with different factors. The odds ratios are shown in Table 1 and suggest the following statements:

- For respondents with Long Term Illness the relative proportion of them being dissatisfied is around 2 times that for those who not have Long Term Illness;
- For respondents with bad health the relative proportion of them being dissatisfied are around 2 times that for those who not have bad health;
- The relative proportion of being dissatisfied over access to doctors increases by 30% (1.307) per extra km distance to the nearest doctors.

```

Call:
glm(formula = Docy ~ Illness + BadHealth + OAC + DocDist, family = binomial,
  data = data.pt2)

Deviance Residuals:
    Min      1Q   Median      3Q     Max 
-1.0652 -0.3400 -0.2633 -0.2330  2.7837 

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -12.99392 324.74370 -0.040  0.968    
Illness       0.72225  0.10869  6.645 3.03e-11 ***  
BadHealth     0.69931  0.17229  4.059 4.93e-05 ***  
OACBlue Collar Communities 9.38625 324.74373  0.029  0.977    
OACCity Living 10.15737 324.74400  0.031  0.975    
OACConstrained by Circumstances 9.36875 324.74377  0.029  0.977    
OACCountryside 9.15778 324.74373  0.028  0.978    
OACMulticultural 10.09480 324.74399  0.031  0.975    
OACProlspering Suburbs 9.17904 324.74371  0.028  0.977    
OACTypical Traits 9.11494 324.74373  0.028  0.978    
DocDist        0.26801  0.03751  7.146 8.95e-13 ***  
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 1. Results of the GLM analysis of dissatisfaction over access to GP surgeries as run in R.

Variable	Illness	BadHealth	DocDist
Odds ratio	2.059	2.012	1.307

Table 1. The odds ratios generated from the exponential of the coefficient estimates associated with different factors

A GWR analysis was then used to analyse for any the spatial variation or non-stationarity in these relationships. The spatial distribution of effects of the different predictor variables on the proportion of being dissatisfied are shown in Figure 2. It is clear that, whilst there is little spatial variation in the effect of Long Term Illness and Distance, there is considerable variation in the effect of Bad Health, with a broad trend increasing from NorthEast to SouthWest.

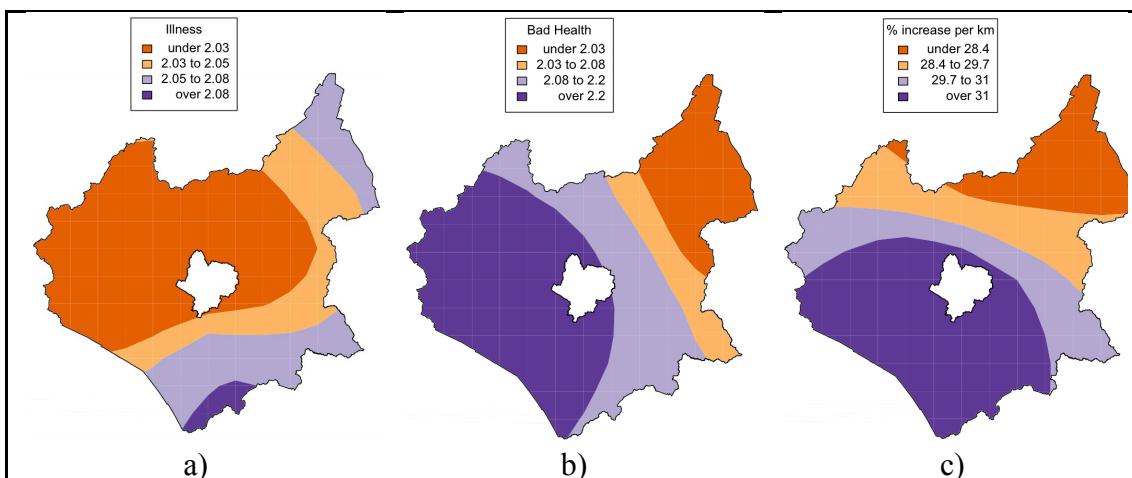


Figure 2. The spatial variation in the relative proportion of being dissatisfied over access to GP surgeries for those a) with Long Term Illness compared to without, b) with Bad Health compared to those in good health, and c) the percentage increase in dissatisfaction with distance. The class divisions in each case are derived from 25th, 50th and 75th percentiles of the distribution.

Discussion

Much work has used geographical information systems (GIS) to analyse accessibility to services in rural areas, often focusing on one or two specific services and on considering access from only a spatial distance-based perspective. For example, White et al (1997) and Langford and Higgs (2010) applied GIS to analyse changes in post office provision and associated subsequent impacts on accessibility, whilst Comber et (2009) developed a model to optimise closures patterns against stated accessibility objectives. Martin et al (2002) and Lovett et al (2000, 2002) used GIS analyses to examine access to various health services within rural parts of the UK and these techniques have been used in rural studies in other countries (Bamford 1999; Brabyn and Barrett 2004). Kaufman (1999), Morton and Blanchard, (2007) and McEntee and Agyeman (2010) have also used GIS in relation to food accessibility, while Morrison and OBrien (2001) have looked at banking provisions.

Perceptions of service access will be influenced by a number of factors, some of which may be quite local: the number of available services (choice), the ability to obtain those services (opening hours, public transport links), perceptions of service quality. Linking measures of perceived access and geographic access and exploring the spatial variation in significant relationship using GWR, provides a richer analysis of the issues relating to service provision that are locally important a number of ways. First, it identifies where perceptions and actual service provision and access are consistent, Second, where perceived access is not strongly related to actual access, GWR identifies areas that require further investigation.

The use of GWR to analyse attitude survey data relating to access in conjunction with physical measures of access, allows the relationship between different dimensions of access and accessibility to be examined. One might expect that as distance from services increased so might dissatisfaction over access to that service. Whilst the concept of accessibility is more complex than stated attitudes in postal survey and GIS-based distance measures, this type of analysis can be used to identify the locations where pockets of variation in the attitudes / distance relationship exist, for

example where dissatisfaction is high and access is high, where dissatisfaction is low and access is low and locations where either dissatisfaction or physical access is low and the other is high. Thus, by considering how such relationships vary in space and across different social groups, this method identifies subgroups that are potentially vulnerable to reductions in service provision. For example, communities where dissatisfaction over service access is high relative to distance are those with potentially low levels of social capital and which may be more vulnerable to such than others. The ability of communities to plug the service gaps resulting from reductions in public service provision is a crucial tenet of the Big Society agenda. Identifying vulnerable communities – those who may not have the social capital to bid for and run facilities at risk of closure or to take over local state-run services as envisioned in the Big Society (DLGC, 2010) – is important if those groups are not be socially excluded by the changes in service delivery. The use of GWR in this way demonstrates that it is possible to generate a richer analysis of accessibility by considering both the qualitative and quantitative dimensions of access.

References

- Bamford, E. J., Dunne, L., Taylor, D. S., Symon, B. G., Hugo, G. J., and Wilkinson, D. (1999) Accessibility to general practitioners in rural South Australia: a case study using geographic information system technology, *Medical Journal of Australia*, 171, 614–616,
- Brabyn, L. and Barnett, R. (2004). Population need and geographical access to general practitioners in rural New Zealand. *Journal of the New Zealand Medical Association*, 117 (1199), 1-13
- Comber, A.J., Brunsdon, C., Hardy, J. and Radburn, R. (2009). Using a GIS-based network analysis and optimisation routines to evaluate service provision: a case study of the UK Post Office. *Applied Spatial Analysis and Policy*, 2(1), 47-64
- Department for Communities and Local Government (2010) Draft Structural Reform Plan, Department for Communities and Local Government, London
www.communities.gov.uk/publications/corporate/structuralreformplan
- Kaufman, P. Rural poor have less access to supermarkets, large grocery stores, *Rural Development Perspectives*, 13(3), 19-26
- Langford, M. and Higgs, G. (2010) Accessibility and public service provision: evaluating the impacts of the Post Office Network Change Programme in the UK, *Transactions, Institute of British Geographers*, 35(4), 585–601
- Lovett, A., Haynes, R., Sunnenberg, G., and Gale, S. (2002) Car travel time and accessibility by bus to general practitioner services:a study using patient registers and GIS, *Social Science and Medicine* 55, 97–111
- Lovett, A., Haynes, R., Sunnenberg, G., Gale, S., (2000) Accessibility of primary health care services in East Anglia HPP Research Report Series 9, Norwich: School of Health Policy and Practice, University of East Anglia.
- Martin, D., Wrigley, H., Barnett, S. and Roderick, P., (2002) Increasing the sophistication of access measurement in a rural healthcare study, *Health and Place*, 8 (1), 3-13.
- McEntee, J. and Agyeman, J. (2010) Towards the development of a GIS method for identifying rural food deserts: geographic access in Vermont, USA, *Applied Geography*, 30, 165–176.
- Morrison, P. S. and OBrien, R. (2001) Bank branch closures in New Zealand: the application of a spatial interaction model, *Applied Geography* 21, 301–330

- Morton, L. W. and Blanchard, T. C. (2007). Starved for access: life in rural Americas food deserts. *Rural Realities*, 1(4), 1–10.
- Vickers, D.W. and Rees, P.H. (2007). Creating the National Statistics 2001 Output Area Classification. *Journal of the Royal Statistical Society, Series A*, 170(2), 379–403
- White, S.D., Guy, C. and Higgs, G. (1997) Changes in service provision in rural areas. 2. Changes in post office provision in mid-Wales: a GIS-based evaluation, *Journal of Rural Studies* 13 (4), 451-465.

Distance metric selection for calibrating a geographically weighted regression model

Binbin Lu, Martin Charlton

National Centre for Geocomputation, Ionias Building, North Campus, NUI Maynooth, Maynooth, Co. Kildare, Ireland
 Telephone: (353) 1 7086455
 Fax: (353) 1 7086456

1. Introduction

In reality, spatial processes tend to vary over space due to changing geographical contexts, and then spatial non-stationarity emerges (Jones & Hanham 1995). As more and more scholars are realizing this, they start to pay more attention to developing the local forms of spatial analysis methods, in which Geographically Weighted Regression (GWR) (Brunsdon et al. 1996; Fotheringham et al. 1998) is proposed as a kind of local technique to estimate regression models with spatially varying relationships. Coinciding with the first law of geography (Tobler 1970), “Everything is related to everything else, but near things are more related than distant things”, GWR makes a point-wise calibration around each regression point with especially concerning a ‘bump of influence’: around each regression point nearer observations have more influence in estimating the local set of parameters than observations farther away (Fotheringham et al. 1998).

How “far” is far? For this question, distance is a clear indicator by giving a quantitative metric. In mathematics, it is defined as a metric between elements in a metric space, and follows four conditions: non-negativity, identity, symmetry and triangle inequality. In related domains of GIS, Euclidean distance (straight-line distance) is the most commonly used metric. GWR is not an exception either that almost all the models have been calibrated by regarding it as the default metric. However, Euclidean distance might be not always the best choice for measuring the proximity between modeled objects; it could be an unreasonable measure due to surface distortion or partition from natural/man-made features. Theoretically, a GWR model can be calibrated perfectly with the potentially best distance measured, but the underlying rule of this is a black box due to the diversity of data and complexity of geographical context. One feasible way is to find an approximately optimal metric. Unfortunately there is no generic method for the distance metric selection, and in previous papers they have adopted descriptive statistics or prior information to evaluate performances of different distance metrics (Mitra et al. 2002; Kamarainen et al. 2003; Shahid et al. 2009). In the case of GWR diagnostic statistics is concerned to be applied in the distance metric selection. For this purpose, a generalized form of distance metric in Euclidean space, Minkowski distance (MD) function, is introduced, and parameters are clarified according to the scores of Akaike Information Criterion (AIC).

2. Minkowski distance function

As a generalization of all the commonly used metrics in Euclidean space, MD is defined as:

$$d = \left(\sum_{i=1}^n |x_i - x'_i|^p \right)^{\frac{1}{p}} \quad (1)$$

where (x_1, x_2, \dots, x_n) and $(x'_1, x'_2, \dots, x'_n)$ are two vectors in n-dimension Euclidean space, and p is a positive real number. When p is 1, 2 and infinity, the distance is known as Manhattan distance, Euclidean distance and Chebyshev distance respectively. Different values of p for this function mean different metrics for the space, which could be validated straightforwardly

by the distance iso-surfaces with sampled values of p (as shown in figure 1). As noticed from figure 1, the rotation of coordinate system will lead to changes in distance measurement when the value of p is not 2. Then the rotated angle θ could be the other factor for this selection. The selection comes down to choosing an optimum value of p , together with considering the rotation angle θ of the coordinate axes. From a practical aspect, the 2-dimension Euclidean space is focused on the here, and in this case the formula(1) could be rewritten according to the coordinate transformation rule:

$$d_{p,\theta} = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2} \left(|\sin(\theta + \alpha)|^p + |\cos(\theta + \alpha)|^p \right)^{\frac{1}{p}}, \text{ where } \alpha = \arctan\left(\frac{x_1 - x'_1}{x_2 - x'_2}\right) \quad (2)$$

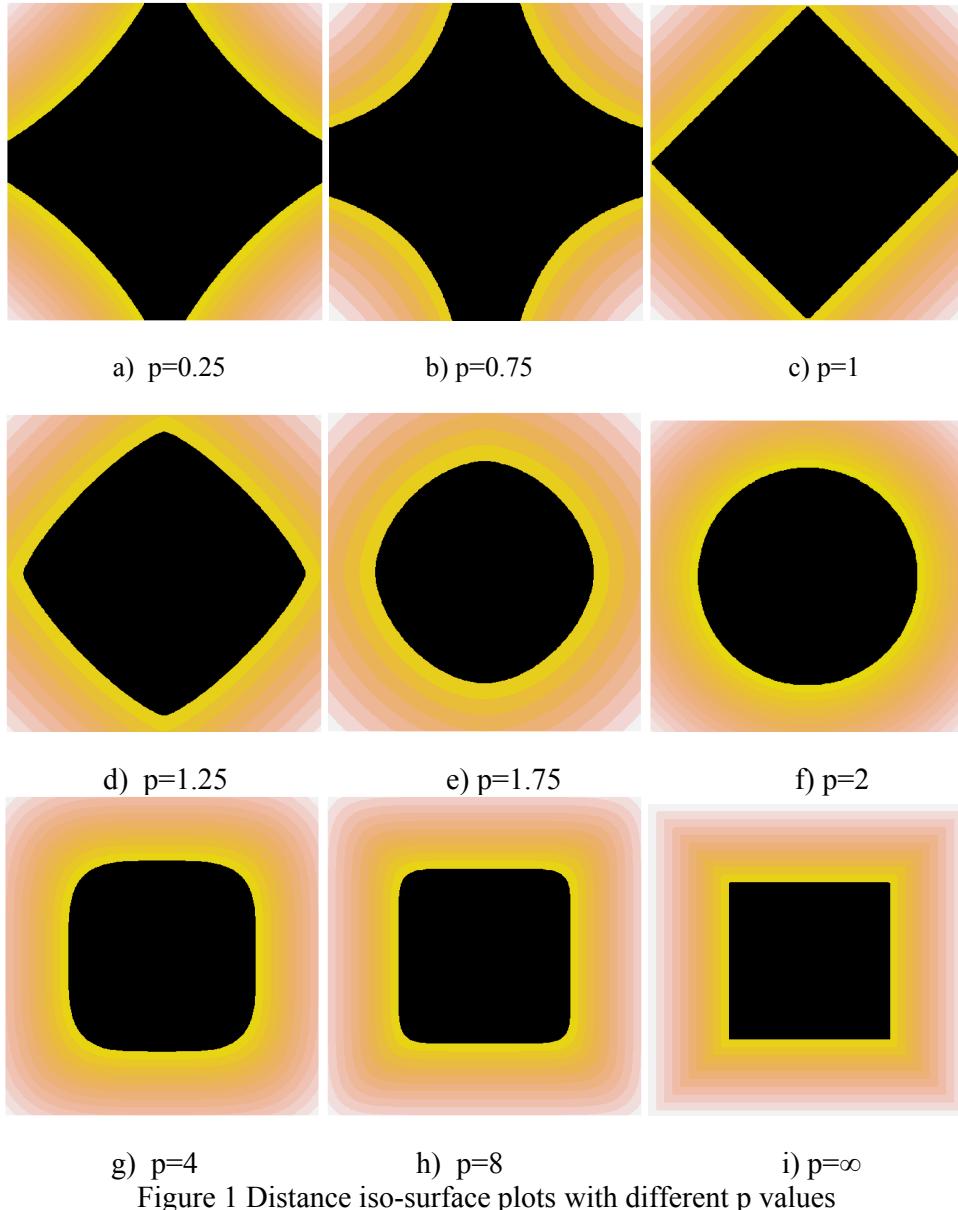


Figure 1 Distance iso-surface plots with different p values

3. Experiment

In this paper, AIC is used to evaluate this selection. It measures the goodness of fit of a GWR model calibration, and the smaller AIC value means the better model calibration (Fotheringham et al. 2002).

In GWR methodology, two kinds of kernel functions, fixed and adaptive, are available. As a matter of fact, this could be regarded as a good example of using different distance metrics: the former is the usage of absolute Euclidean distance while the latter is the consideration on relative Euclidean distance. Both kinds of kernel functions are also suitable for any specific MD metric. Following the procedures of the GWR technique, an optimum bandwidth should be determined for each calibration with different specific metric. On the consideration of the computational complexity, the cross-validation approach is used for the bandwidth selection instead of using AIC as the indicator. Here a rough procedure of distance metric selection is proposed in line with the above discussion:

1. For each specified pair (p, θ) , select the optimum bandwidth based on the scores of the CV approach;
2. Compute AIC values with specified MD functions and selected bandwidth, and determine the values of (p, θ) with the smallest AIC value for specifying the optimum MD function and corresponding rotation angle.

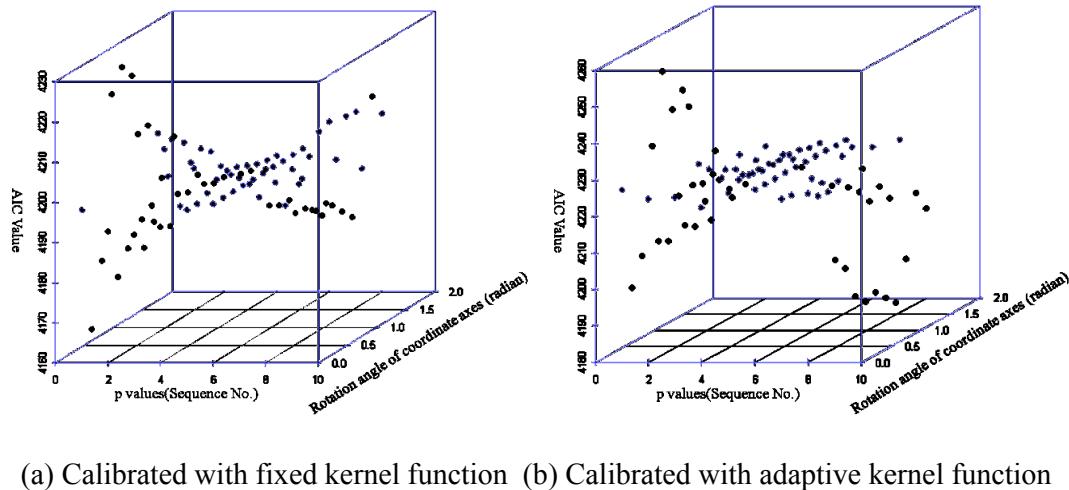


Figure 2 AIC values of the example GWR model calibrated with MD metrics specified by different sets of (p, θ)

In this paper, a sampled house price data-set in London is modeled to explore a spatially varying relationship between house price and floor area. For calibrating this GWR model, 90 sets of (p, θ) , which are permutations of the groups $P\{0.25, 0.75, 1, 1.25, 1.75, 2, 4, 8, \infty\}$ and $\Theta\{\text{which is an arithmetic sequence from } 0 \text{ to } 90^\circ \text{ with 10 elements}\}$, are tested. In figure 2, the scatter plot shows us the computed AIC value for each specific (p, θ) . Evidently the AIC values vary a lot with different values of p and rotated angles. As shown in table 1, the minimum AIC values for both kernels have significant reduction compared with results from using Euclidean distance, by 32.8 and 41.145 respectively.

Table 1 Comparison of results using Euclidean distance and the optimum (p, θ)

	Fixed bandwidth		Adaptive bandwidth	
p	2	0.25	2	Inf
θ	---	0.1710423	---	1.0262536
AIC	4199.701	4166.880	4227.42	4187.275

4. Conclusion

The results show us that the calibration of a GWR model could be improved significantly by choosing an optimum distance metric with corresponding coordinate system rotation. However, this selection is concerned as a whole for the model calibration, while actually the distance metric is somehow depended on a particular feature attribute. If this selection could be fulfilled for each independent variable in a GWR model, then the better estimations might be made. Moreover the mixed GWR model (Brunsdon et al. 1999) could be regarded as a special case of calibrating a regression model using different distance metrics individually for different attributes, and it forms an interesting future challenge.

Acknowledgements

Research presented in this paper was jointly funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan and China Scholarship Council (CSC). The authors gratefully acknowledge their support.

Reference

- Brunsdon, C., Fotheringham, A.S. & Charlton, M., 1996. Geographically weighted regression: a method for exploring spatial non-stationarity. *Geographical Analysis*, 28, p.281–298.
- Brunsdon, C., Fotheringham, A.S. & Charlton, M., 1999. Some notes on parametric significance tests for Geographically Weighted Regression. *Journal of Regional Science*, 39(3), pp.497-524.
- Fotheringham, A.S., Brunsdon, C. & Charlton, M., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships 1st ed., London: John Wiley & Sons Ltd.
- Fotheringham, A.S., Charlton, M. & Brunsdon, C., 1998. *Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis*. Environment and Planning A, 30(11), pp.1905-1927.
- Jones, J.P., Hanham R.Q., 1995. Contingency, realism, and the expansion method. *Geographical Analysis*, 27, p.185–207.
- Kamarainen, J.-K., Kyrki V., Jarmo I. & Kalviainen H., 2003. Improving similarity measures of histograms using smoothing projections. *Pattern Recognition Letters*, 24(12), pp.2009-2019.
- Mitra, P., Murthy, C.A. & Pal, S.K., 2002. Unsupervised Feature Selection Using Feature Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3), pp.301-312.
- Shahid, R., Bertazzon, S., Knudtson, M. & Ghali W.A., 2009. Comparison of distance measures in spatial analytical modeling for health service planning. *BMC health services research*, 9, p.200.
- Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2), pp.234-240.

Eight Challenges for Social Flows in a GIS

C. Andris¹, J. Ferreira¹

¹Massachusetts Institute of Technology, Department of Urban Studies and Planning
 77 Massachusetts Avenue, Room 9-532
 Cambridge, MA 02139
 Telephone: +1 301 908 4855
 Fax: +1 617 253 2654
 {clio, jf}@mit.edu

1. Introduction The built environment and its inhabitants can be thought of as a set of multiple systems, where (1) humans interact with one another, (2) humans interact with the landscape, and (3) the landscape interacts with itself. These systems are not self-contained or closed: they interact with one another, and across system lines. Thus, our set of multiple systems can be thought of not as a group of systems, but as a system of systems. When considering social and spatial as standalone systems, there is much evidence for the successful modelling of social processes with behavioural, agent-based, game-driven or discrete choice models, and geographic processes with a proven toolbox of spatial statistics and agent-based processes such as the work of Batty (2007) in cellular automata and other modelling capabilities. But to model social and spatial as a system of systems, where social/cognitive choices are represented spatially, current approaches seem to fall short. For example, cutting-edge research in social network analysis (see Onnela et al 2010) uses a “geography” that only accounts for Euclidean distance, and moreover, represents geography as an oversimplified node-link network, that neglects the adjacency, areal boundaries, cost distance and natural geographic processes. This paper will outline obstacles to the development of social/spatial research under the assumption that flow (also called interaction, connectivity, or network) data connects two places either by the transfer of humans or information. This paper develops a framework for theoretically approaching, implementing, integrating, and learning from these “systems of systems”, or more specifically, the intersection of social networks and geographic space, following the seminal work of Torrens (2010), but in the context of flow data. We assume that social flows fall into three major groups: transportation, communications, and social network representative flows. In this framework, we address eight pragmatic issues for the use of these place-to-place connectivity measures in a spatial context.

2. Eight Challenges for Social Flows within a GIS Framework

(1) **Difficulty characterizing system nodes:** A system node can be characterized by its local features, but also by its relationship to other nodes. Each flow radiating from a node has a destination, properties of that destination, flow direction, distance and magnitude. Since nodes participate in many flows, it is hard to summarize these variables and group by node without major fidelity loss due to summarization. We suggest that unsupervised classification methods can be used to classify system nodes by their specific geometric radial configurations, as represented by a series of numeric vectors. Some examples include Self-Organizing Maps (Guo 2006), Eigenvector decomposition (Calabrese et al 2006) and K-means (Andris 2011).

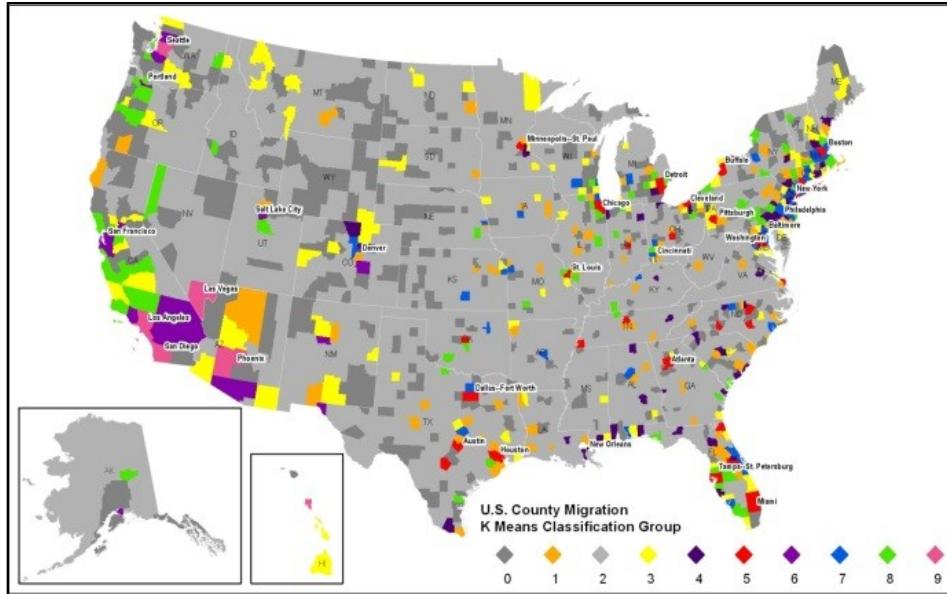


Figure 1: A clustering approach is used to classify counties by their incoming migration flow geometries.

(2) **Unconstructed theories of edge assignment:** Edges that represent telecommunications have little interaction with the space between calling agents, Adams (2010) while in comparison, edges between a pedestrian's origin and destination represent an embeddedness in the area between start and stop points. We find difficulty expressing edges that overlay on maps, as they do not inform the user as to the relevance of the trace. A framework for representing these connections is presented as a spectrum that ranges from fine-grained traced paths to a more tabular origin/destination representation. (Figure 2)

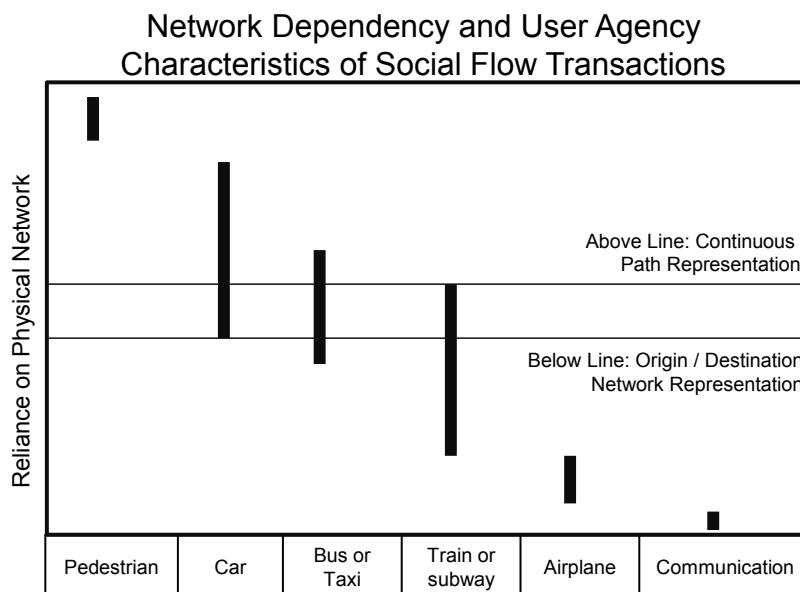


Figure 2: A system for classifying different geodesic traces informs a user of the appropriate representation of a trace by its reliance on the physical network.

(2) Inability to employ spatial analytic methods: These data are mostly unable to be analysed with spatial statistics, like point-pattern analysis, and spatial operations (like a clip or spatial join) as these statistics are fit for single points, polygons, and less frequently, lines, based on each entity's proximity or adjacency to one another. For example, the widely-used Geographically-Weighted Regression (Brunsdon et al 1998) has uncovered relationships between variables using continuous and point space, but is not currently fit for flow data. (see, for example, Figure 3) Here, we suggest amended versions of traditional spatial analytic methods by considering how edges should be assigned to geographic space (see (2)) in order to use the most appropriate representation. Thus, we can apply proximal clustering and statistical methods to origin/destination nodes, or assigning statistical properties to the edges themselves. Without sensitivity to how the flow interacts with the ground below it, the entity's 'existence' in geographic space can have various meanings.

(3) Muddled visualization: Since flows connect two places in absolute, discrete space, the edges that connect these places are poorly suited for large-scale visualization because the number of edges in a typical dataset is too dense for the constraints of 2D space. For this, we suggest querying, filtering and automating visual and tabular ordinal hierarchies, in addition to the classification methods mentioned in (1).

(4) Lack of visual-analytic systems: There are currently a lack of ESDA and software platform systems for exploring (a) the relationship between a social network and geographic space and (b) spatial relationships between non-adjacent entities. Following the work of Takatsuka and Gahegen (2002) and Anselin et al (2006) with the GeoVISTA Studio and GeoDA, respectively, we suggest that interactive dynamic environments be created for users to explore social and spatial configurations, and statistical properties, in a single view. (Figure 4)

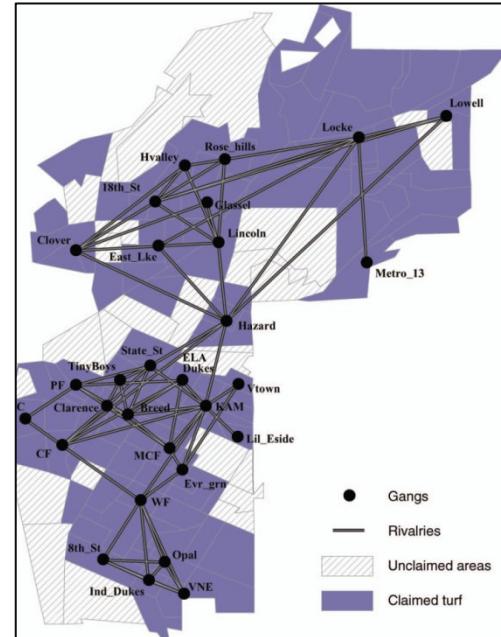


Figure 3: A social network of reported gang fraternity overlayed in geographic space shows that closeness does not always correspond with adjacency. (Figure from Radil et al 2010)

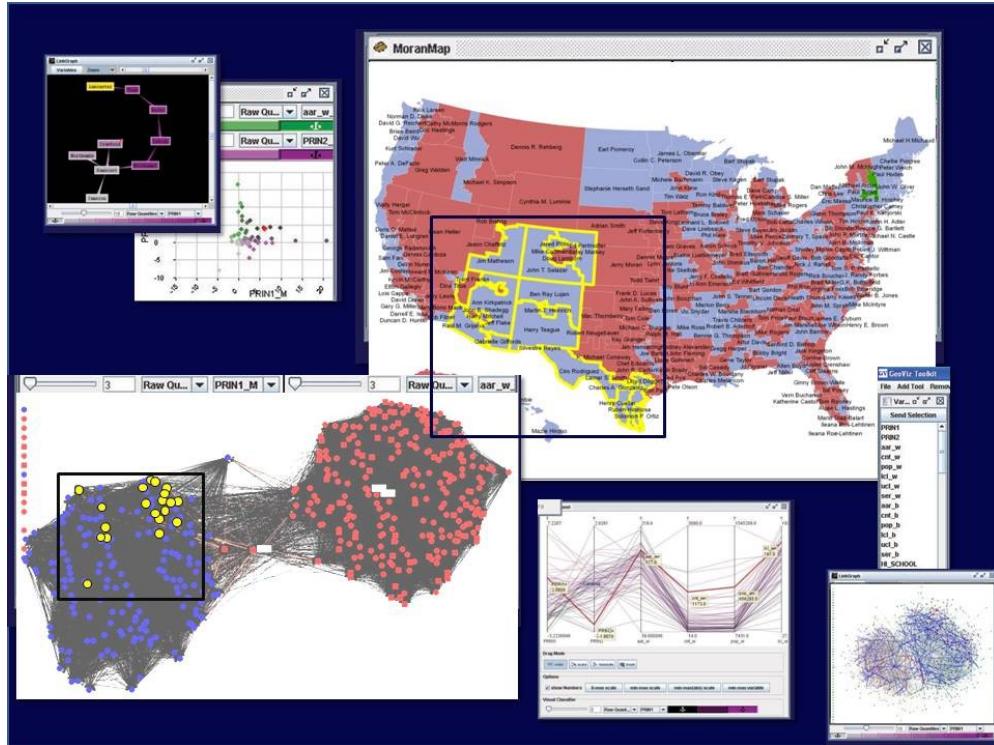


Figure 4: A prototype for an interactive social/spatial geo-visualization environment allows the user to manipulate entities in a force-directed social network in tandem with a linked map. In this example, yellow entities in the network are political figures, and correspond to their representative districts, as also highlighted on the map.

(6) Few socially-driven flow prediction methods: Geographically-dependent decisions to meet, migrate, travel or communicate are comprised of more than distance factors, but social factors. There is a higher “cost” of traveling somewhere unknown and without friend recommendations, than to a familiar place. This familiarity is evidenced by previous travel or communications flow data, in migration, this phenomena is known as ‘chaining’. (Castells 2000) Given that decisions to relocate or choose a job are often driven by social reasons, we present new prediction methods that account for chaining by using components of Bayes’ Law, and focusing specifically on the disaggregate, unique relationship between place-pairs, in addition to traditional gravity models.

(7) Lack of GIS infrastructure for flow manipulation: The phenomenon of a spatial flow incorporates the features at its origin and its destination, as well as features of the flow. In the field of Database Management Systems (DBMS), the theory and implementation of Spatial Data Infrastructure (SDI) is not yet developed to treat two unique points and a connecting edge as a unit of analysis, making selection, operations and manipulation difficult. (Figure) Perhaps it is for these reasons that flow data has not been as prominent of a fixture in Geographic Information Systems, in terms of software, computation, statistical endeavours, academic and professional use. Solutions to these challenges are manifold, but early advancements suggest combining graphs for easier querying. (Doytsher et al 2009) We suggest creating a new object (called not point, line or polygon, but ‘flow’ etc.) that packages nodes and edges into a single entity. This entity

can then be queried and interactively selected by its named components (e.g. the origin, destination and trace).

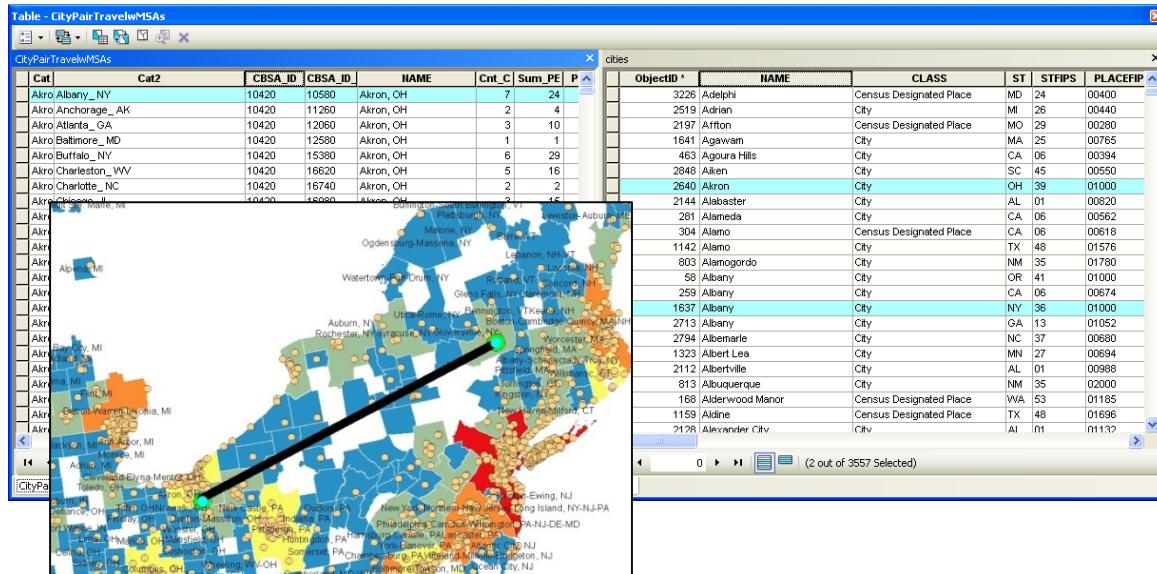


Figure 5: An example flow in the ArcGIS software environment, and its corresponding tables illustrates the geographic entity selection framework.

(8) Difficulty developing a lingua franca and taxonomies for the field: There seems to be little research on ontologies, nomenclatures, typologies and frameworks for flow data, although these linguae seem to be successfully implemented for the separate fields of Complex Network Analysis and Geographic Information Systems. Refining these typologies may enable progress in the synergistic field, and its ability to be communicated, improved, and understood. Initiative for this streamlining should come from multiple GIS entities: (1) software corporations, (2) GIS textbooks, (3) Higher Education terminology in labs and lectures, and (4) peer-reviewed literature.

3. Conclusion Given that geographers can benefit from better manipulation of social flows and connectivity over space, we consider eight pragmatic challenges and possible solutions for use of social flows and social distance in the digital era. We hope these issues can aid in analysis for operations, logistics, and planning, in tasks such as city-to-city airline passenger magnitudes, future road capacity, inventories of infrastructure needs for urban growth, city shrinkage, migration forecasts, site suitability for cooperative meeting places, epidemiological spreading or containment models, and suggestions for business franchise or advertisement expansion.

References

- Adams, P, 2011, A Taxonomy for Communication Geography. *Progress in Human Geography*, 35(1): 37-57.
- Andris, C, 2011, Weighted Radial Variation for Node Feature Classification. *Arxiv preprint arXiv:1102.4873*.

- Anselin, L, Syabri, I, and Kho, Y, 2006, GeoDa: An Introduction to Spatial Data Analysis. *Geographical Analysis*, 38(1): 5-22.
- Batty M, 2007, *Cities and Complexity*. MIT Press, Cambridge, MA, USA.
- Brunsdon, C, Fotheringham, S and Charlton, M, 1998, Geographically Weighted Regression. *The Statistician*, 47(3): 431-443.
- Calabrese, F, Reades, J, and Ratti, C, 2009, Eigenplaces: Segmenting Space through Digital Signatures. *IEEE Pervasive Computing*, 9: 78-84.
- Castells, M, 2000, *The Rise of the Network Society*. Wiley-Blackwell, Cambridge, MA, USA.
- Doytsher, Y, Galon, B, and Kanza, Y, 2010, Querying Geo-social Data by Bridging Spatial Networks and Social Networks. *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, ACM, 39-46.
- Guo, D, Chen, J, MacEachren, A, and Liao, K, 2006, A Visualization System for Space-Time and Multivariate Patterns (vis-stamp). *IEEE Transactions on Visualization and Computer Graphics*, 12(6): 1461-1474.
- Onnela, J-P, Arbesman, S, Barabasi, A-L, Christakis, N, 2010, Geographic Constraints on Social Network Groups, arXiv:1011.4859v1 [physics.soc-ph].
- Radil, S, Flint, C, Tita, G, 2010, 'Spatializing social networks: Using Social Network Analysis to Investigate Geographies of Gang Rivalry, Territoriality, and Violence in Los Angeles. *Annals of the Association of American Geographers*, 100(2): 307 – 326.
- Takatsuka, M, and Gahegan, M, 2002, GeoVISTA Studio: A Codeless Visual Programming Environment for Geoscientific Data Analysis and Visualization. *Computers & Geosciences*, 28(10): 1131-1144.
- Torrens P, 2010, Geography and Computational Social Science. *GeoJournal*, 75(2): 133-148.

A flexible model for haptic-assisted pedestrian navigation mobile applications

Ricky Jacob, Peter Mooney, Padraig Corcoran, Adam Winstanley

Department of Computer Science, National University of Ireland Maynooth,
Maynooth, Co. Kildare, Ireland.
email {ricky.jacob}@nuim.ie

1. Introduction

Pedestrian navigation applications, specifically for mobile devices, have received much research and development attention over the past decade or so with many different types of solutions developed (Kenteris et al.; 2011). The most common interface is a map-based interface with written and/or verbal turn-by-turn directions (possibly including landmark information). Haptic technology, or haptics, is a tactile feedback technology that takes advantage of our sense of touch by applying forces, vibrations, and/or motions to the user (Nakao et al.; 2010). The potential of haptic technology has only recently started to receive the attention of the research community (Jacob et al.; 2010). In mobile devices haptic-feedback is delivered in the form of vibrations which can be programmatically controlled using the phone software API. In this paper we describe a simple, flexible, model for the integration of haptic feedback into pedestrian navigation applications on mobile devices. A constraint is that the mobile device must have an onboard GPS and compass. The vibration motor on the mobile device must also be capable of being controlled from software running on the device itself. Our model allows a “heads up” approach to pedestrian navigation with the mobile device where the user is not required to keep looking down to check the screen of the mobile device. For testing purposes text-based navigation assistance is provided in conjunction with the haptic-feedback on the device screen in our prototype implementation. Three distinct modes of vibration of the device are used to provide haptic feedback to the user.

2. Description of Model

Our model is presented in Algorithm 1 and described in Section 3.. The user starts our application on their mobile device. The first step involves choosing both the start location (default is their current location taken from the device GPS) and destination. A simple slippy map interface is provided for this purpose. When the user has selected the route start and end points the Cloudmade routing service (Cloudmade; 2011) is automatically invoked with the parameters describing the requested route. A GPX file (illustrated in Figure 1) is returned to our application on the device. This file is immediately parsed and stored in the spatial database (PostGIS). The application then indicates that the user must scan (see the *Scan()* function in Algorithm 1 below) for the direction they should proceed in. *Scan()* requires the user to hold the device in front of them and slowly move it to find the correct forward direction. The circular buffer size around each route point is also set (see Figure 2). If the user is inside the buffer of a route point the application causes the device to vibrate (pattern

1) to indicate the user must scan for the correct direction. When the correct direction is found the device vibrates again (pattern 2) to indicate the correct direction. If the user strays off in the wrong direction the application vibrates (pattern 3) and they must scan again or physically return to the route themselves. This process is repeated until the user has reached their destination.

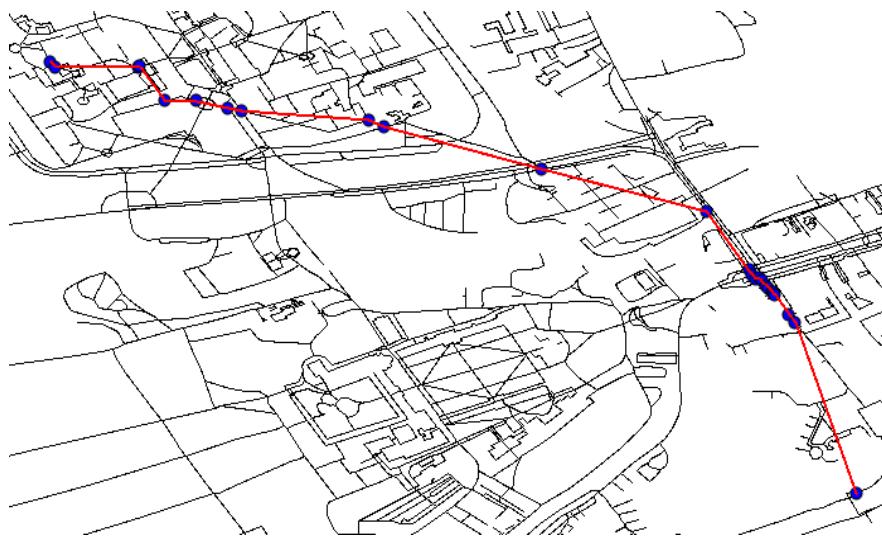


Figure 1: A GPX file output from the Cloudmade Routing Service overlayed on an OpenStreetMap *roads* shape file. Circular icons represent route points in the computed shortest route.

3. Implementation of our Haptic-Feedback Model

A HTC Magic, running the Android Mobile Operating system, was used for development and user testing. All software development on the software device was carried out in Java. The Android Software Development Kit provides the tools and APIs necessary to begin developing applications on the Android platform using the Java programming language. The Cloudmade Routing service (Cloudmade; 2011) was used as the web-service for generating the shortest pedestrian paths. Cloudmade use the global OpenStreetMap database for computation of shortest paths. With the Cloudmade routing service we have more “human orientated” walking routes. If the OSM data has paths across open areas properly tagged then these are considered in the computation of the shortest walking route. The Cloudmade routing service returns computed routes in JSON or GPX formats. Figure 1 shows the resulting GPX output file generated by the Cloudmade Routing Service for the shortest pedestrian route from the Computer Science Department at NUI Galway to the Maynooth Business park. The circular icons represent route or turning points in the computed shortest pedestrian route. A schematic diagram of our implementation is shown in Figure 2. A PHP script runs on the database webserver and this script acts as a broker service between our local spatial database and the mobile device. The script receives the user location data every t seconds. As described in Algorithm 1 the computed route is stored in this database. The PHP script matches the user location to this route and returns a response to indicate if the user is: going in the correct direction (no vibration), within the buffer of a route point (vibration 1), going in the wrong direction or is off route (vibration 2), or has reached their destination (no vibration). Initial user trials with this application have focused on observing and eliciting feedback on how test participants interact with the haptic application, use the application at the route

points, respond to feedback, and rate the usability or usefulness of the application. We will present quantitative results of a series of user trials at the GeoComputation conference and in future papers.

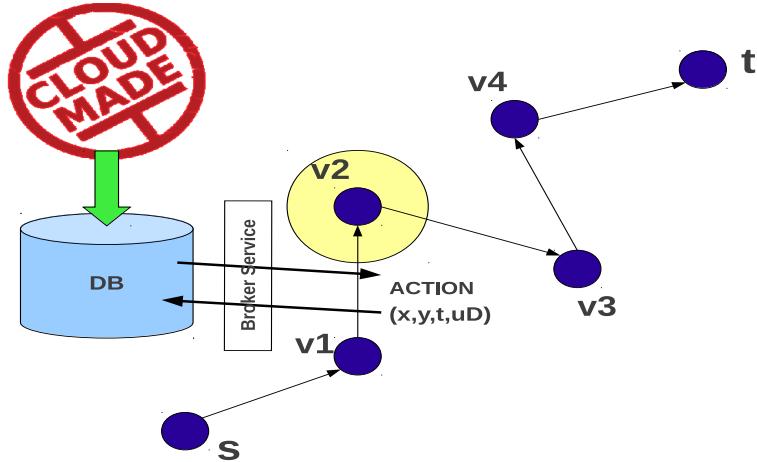


Figure 2: An example route. The user position and bearing is taken from the mobile device every x seconds. Upon delivery of this information to the broker software at the DB an action is initiated on the mobile device

4. Conclusions and Future Work

This paper has given a brief overview of a flexible model for integrating haptic-feedback into pedestrian navigation applications on mobile devices. To properly quantify the advantages and disadvantages of the model we will need to carry out additional user tests which will involve a wider range of participants. Initial feedback from our trials of the application with a small group of test participants was positive. The users found the application novel and quickly learned the vibration patterns and how to respond to this feedback. However, users, unfamiliar with the route, who used the haptic-enabled device, paused for longer times at route points compared to users familiar with the test routes. In an extensive study Ishikawa et al. (2008) compared the wayfinding behavior and acquired knowledge by participants who received information about routes from a GPS-based navigation system, from maps, and from direct experience of the routes. With respect to wayfinding behavior, participants who used the GPS-based navigation system traveled longer distances, made more stops during the walk than participants who viewed maps, and walked slower overall. We shall be carrying out similar user testing to investigate if walking performance of pedestrians, using the haptic application on their mobile device, improves as less “head down phone viewing” is required. Finally, using the vibrate function on the mobile device can drain power quickly from the battery. Most mobile device users would find the possibility of a dead battery a serious drawback to haptic-assisted navigation. Battery lifetime is a major usability concern to mobile phone users (Rahmati and Zhong; 2009). Rahmati and Zhong (2009), in a study of mobile phone users, “found that most recharges are driven by pressures of time and location instead of low battery”.

Acknowledgements

Research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. Peter Mooney is funded by the Irish Environmental Protection Agency STRIVE programme (grant 2008-FS-DM-14-S4).

References

- Cloudmade (2011). Cloudmade routing web services, Online API available at <http://developers.cloudmade.com/projects/show/routing-http-api> - Checked January 2011.
- Ishikawa, T., Fujiwara, H., Imai, O. and Okabe, A. (2008). Wayfinding with a gps-based mobile navigation system: A comparison with maps and direct experience, *Journal of Environmental Psychology* **28**(1): 74 – 82.
- Jacob, R., Mooney, P., Corcoran, P. and Winstanley, A. C. (2010). Haptic-gis: Exploring the possibilities, *SIGSPATIAL Special* **2**: 13–19.
- Kenteris, M., Gavalas, D. and Economou, D. (2011). Electronic mobile guides: a survey, *Personal and Ubiquitous Computing* **15**: 97–111.
- Nakao, M., Kitamura, R., Sato, T. and Minato, K. (2010). A model for sharing haptic interaction, *Haptics, IEEE Transactions on* **3**(4): 292 –296.
- Rahmati, A. and Zhong, L. (2009). Human-battery interaction on mobile phones, *Pervasive and Mobile Computing* **5**(5): 465 – 477.

A Description of Algorithm for a Haptic-assisted Pedestrian Navigation Application

Data: The input start location s and the destination location t (both longitude,latitude) of the route required

Result: In-route Haptic-assistance for pedestrian using their mobile device on route between s and t

Call Cloudmade Routing Service;
Download XML-encoded result from Cloudmade;
Parse and store route in Database;

```

begin
     $d \leftarrow setRoutePointBufferSize(10m);$ 
    while ( $U.location \neq buffer(t, d)$ ) do
         $U \leftarrow getCurrentUserLocation()$  This includes user direction;
         $D \leftarrow getLocationOfNextRoutePoint();$ 
        if ( $U.location = buffer(v, d)$ ) then
            repeat
                Until user points their mobile device in the correct direction;
                Vibrate when correct direction is found;
            until ( $Scan() = true$ );
            User can now proceed in the correct direction;
        end
        if ( $U.location \neq buffer(v, d)$ ) then
            if ( $U.direction = D$ ) then
                Everything is OK;
                Put the green light logo on the phone display;
                No feedback necessary;
            else
                Not going in the correct direction;
                Vibrate;
                repeat Until user corrects their direction until ( $Scan() = true$ );
                Update direction variables;
                OK to proceed;
            end
        end
    end

```

Algorithm 1: The algorithm describing the integration of haptic feedback and a pedestrian route

Trajectory Data Mining: Classification and Spatio-Temporal Visualization of Mobile Objects

A. Nara & Paul, M. Torrens

Arizona State University, 975 S. Myrtle Ave., Coor Hall, Tempe, AZ, 85287
Telephone: +1-480-965-7533
Email: {anara; paul.torrens}@asu.edu

1. Introduction

Trajectory-based data mining is a very active research topic in the field of Knowledge Discovery in Databases (KDD) in response to the influx of mobile object data. Using a set of spatio-temporal sequences of mobile object data collected from various types of Location Aware Technologies (LATs) or generated by simulation models, trajectory data mining discovers spatio-temporal knowledge through exercises including pattern detection, clustering, classification, generalization, outlier detection, and visualization. Potential applications across various fields include, for example, vehicle and pedestrian traffic control (e.g., transportation management and facilities design); Location-Based Services (LBS) (e.g., navigation assistance and mobile advertising); weather forecasting (e.g., hurricane trajectory prediction and risk analysis); law enforcement (e.g., video surveillance for criminal activities); animal conservation (e.g., tracking at-risk animal populations); and logistics for goods and human.

In recent years, many approaches have been proposed and applied to various fields to investigate patterns and trends from massive datasets of mobile objects (e.g., Gaffney et al. 2007; Lee, et al. 2007; Andrienko et al. 2009; Guo et al. 2010). Research challenges identified in previous works include characterization, generalization, and visualisation of massive and complex trajectories to discover interesting patterns, trends, and useful knowledge across scales.

In this paper, we propose a trajectory data mining framework that employs trajectory partitioning and clustering algorithms to extract behavioural patterns of mobile objects, as well as visual analysis to display extracted patterns and trends in space and time. As a case study, we developed an Agent-Based Model of pedestrian evacuation based on the social force model and generated crowd evacuation dynamics on a street corridor. The proposed framework successfully differentiated and visualized spatio-temporal clusters of local movement behaviours including smooth evacuation and bottleneck.

2. Methodology

To investigate movement behaviours in trajectory datasets, our proposed trajectory data mining framework includes three methodological steps, trajectory partitioning, trajectory clustering, and spatio-temporal visualization of trajectory clusters.

- Step1: Trajectory partitioning
 - Distance-Threshold approach

- Step 2: Trajectory clustering
 - Quantification of sub-trajectory
 - Principal Component Analysis (PCA)
 - K-means cluster analysis
- Step3: 3D visualization of trajectory clusters
 - Spatio-Temporal Kernel Density Estimate (STKDE) and volume rendering technique

A set of trajectory dataset is described as {Trajectory Set: $TR_{set} = TR_1, TR_2, TR_3, \dots, TR_i$, where i denotes the number of mobile objects}. Each trajectory is composed of a sequence of three-dimensional points $\{TR_i = p_1, p_2, p_3, \dots, p_j\}$, where j denotes the number of points in the trajectory $i\}$, $\{p_j = x, y, t\}$. The trajectory partitioning process partitioned an entire trajectory into trajectory partitions (sub-trajectories), the process of which is a key to extract local movement behaviours. In this study, a Distance-Threshold approach was employed. It uses a distance threshold value to partition a trajectory into sub-trajectories. This is based on the assumption that in many situations human movements involve stopping/staying when a person changes its behaviour. Such behaviours can be seen at multiple scales; for example, when a pedestrian decelerates and ultimately stops to make a sharp turn or to avoid collisions with other pedestrians; a commuter stays at home, walks to a bus stop, waits for a bus, and stays at its office to work; and a person may relocate and find a new home to stay with its life events. Methodologically, partitioning a trajectory based on staying behaviour can be simply achieved by introducing a Distance-Threshold (Th_d). If a distance of each segment in a trajectory is less than Th_d , then the segment is assigned as *STAY* and a trajectory is partitioned by the segment. If consecutive segments are assigned to *STAY*, those segments are considered as one sub-trajectory in order to differentiate short and long staying behaviours.

For each trajectory partition ($TR_{par(i)}$), multi-dimensional vectors are calculated to characterize the sub-trajectory. The vector values include total duration (d_t), total horizontal distance (d_x), total vertical distance (d_y), total two-dimensional distance (d_{2D}), velocity vector on x-axis (v_x), velocity vector on y-axis (v_y), and velocity (v), horizontal beeline distance (d_{sx}), vertical beeline distance (d_{sy}), two-dimensional beeline distance (d_{s2D}), area of minimum bounding box (mbb), and sum of cosine of turning angle between two consecutive segments (sct). All of these vector values are then normalized with mean equals to 0 and variance equals to 1.

To reduce the dimensionality of multiple vectors of sub-trajectories, PCA is employed. PCA is a multivariate statistical technique to the dimensionality of a dataset consisting of interrelated variables by finding a new set of variables, i.e., Principal Components (PCs), which is smaller than the original set of variables but still containing most of the information in the original dataset. Eigenvalues of PCs measure the amount of variation, and this study uses PCs if their eigenvalues are greater than 1. PC scores of each sub-trajectory for each PC (Eigenvalue ≥ 1) are computed, and then they are used as a new input dataset for sub-trajectory clustering.

To classify sub-trajectories for extracting local movement behaviours, the K-means clustering algorithm developed by Hartigan & Wong (1979) is applied. To estimate the optimal value of k in K-means clustering, clustering algorithms are run with different

values of k (min:2, max:20), and the optimal value of k is selected by the Gap Statistic (Tibshirani, Walther, & Hastie, 2001).

The Space-Time Kernel Density Estimation (STKDE) (Brunsdon et al. 2007) and volume rendering technique (Levoy 1988; Nakaya & Yano 2010) are used for visualising cluster density distribution in space and time. The interactive approach of volume rendering is achieved using an open source visualization software, ParaView (Henderson 2007).

3. Results

As a case study to examine the proposed trajectory data mining framework, data regarding pedestrian evacuation dynamics was analyzed. The trajectory data was generated by an ABM based on the social force model (Helbing and Molnár, 1995). In its simplest form, there are three forces formulated as follows.

$$m_i \frac{dv_i}{dt} = m_i \frac{v_i^0(t)e_i^0(t) - v_i(t)}{\tau_i} + \sum_{j \neq i} f_{ij} + \sum_w f_{iw}$$

The first force is a driving force toward a desired destination described by a pedestrian i of mass m_i , of desired velocity v_i^0 , of desired direction e_i^0 , and of actual velocity v_i with a certain characteristic time τ_i . The second force is a repulsive force, $\sum_{j \neq i} f_{ij}$, describing the interaction effects with other agents j ($j \neq i$), and the third force is a repulsive force, $\sum_w f_{iw}$, to avoid walls and obstacles. Pedestrians in this basic form of the social force model walk unidirectionally, i.e., each pedestrian travels between an origin and a destination. This is too simplistic, so to overcome the deficiency, the idea of multiple waypoints is implemented. In the algorithm, each pedestrian i owns a sequenced list of waypoints and walks toward the first waypoint in the list. When it reaches at the waypoint within a certain buffer zone described by a two-dimensional vector $bZ(bx, by)$, the waypoint is removed from the list and the pedestrian walks toward the first waypoint in the new list until reaching the final destination.

In this study, pedestrian evacuation dynamics on a diagonal corridor was simulated. In the simulation, pedestrians evacuate from North, West, and South corridors to an East exit. Table 1 represents initial settings for model environment and parameters used for the social force model. To analyze trajectory data of simulated pedestrian evacuation dynamics, locations (x,y) of pedestrians and corresponding time stamps were output at every one second (=30 frames). As a result of the Gap Statistic, we obtained five sub-trajectory clusters as the optimal k value.

Figure 1 illustrates the clustering result of sub-trajectories using the Distance-Threshold partitioning approach ($k=5$). Figure 2 presents the cluster profiles describing movement characteristics within clusters. The vertical axis represents independent variables for corresponding cluster IDs ($k=5$) and the horizontal axis shows the average of normalized value of independent variables within a cluster. Figure 3 visualises sub-trajectory cluster density distributions in space and time estimated by STKDE. This explains when and where a particular pattern of movement behaviour occurred.

These results showed that sub-trajectories of Cluster 1 and 2 are identified as smooth evacuation behaviours because both have higher average velocity values and continuous trajectories without staying or stopping. In addition, these clusters are found beneath Cluster 4 near the intersection area and on the East corridor in the STKDE map indicating that pedestrians who reached at the corner of the intersection earlier have successful evacuation. On the other hand, Cluster 4 and 5 are partitioned by Cluster 3 that represents staying or stopping behaviours near the corners of the intersection. This explains the evacuation bottleneck due to the overcrowding.

Model environment	Number of pedestrians	120
	Area width	800
	Area height	700
	Simulation Tick	1 frame
Parameters for social force model	Pedestrian's mass m_i	1
	Pedestrian's desired velocity v_i^0	1.3
	Characteristic time τ_i	2

Table 1. Settings of pedestrian evacuation model.

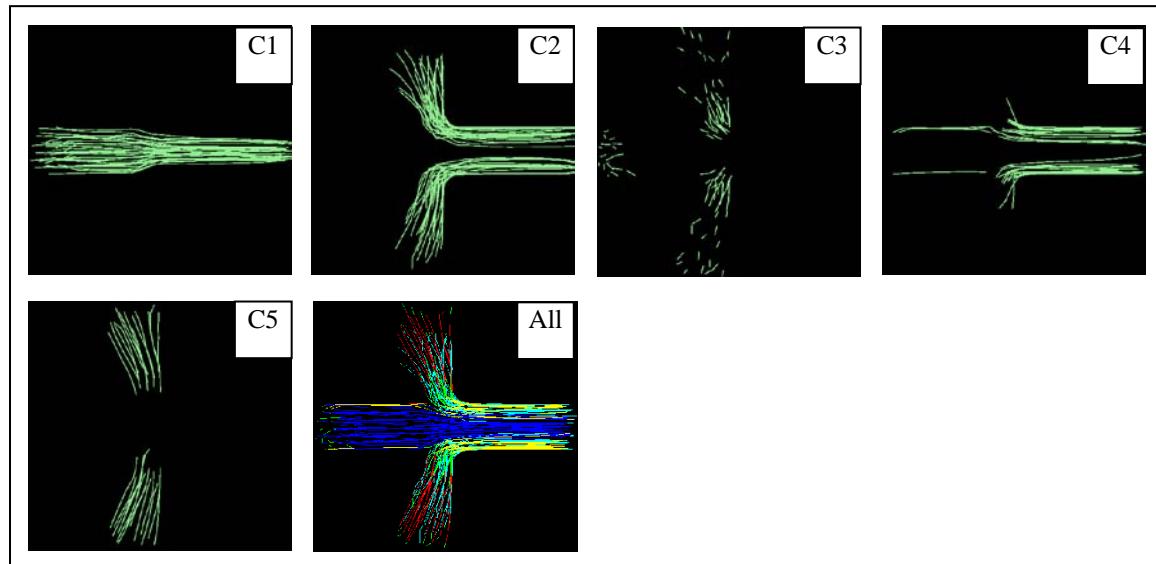
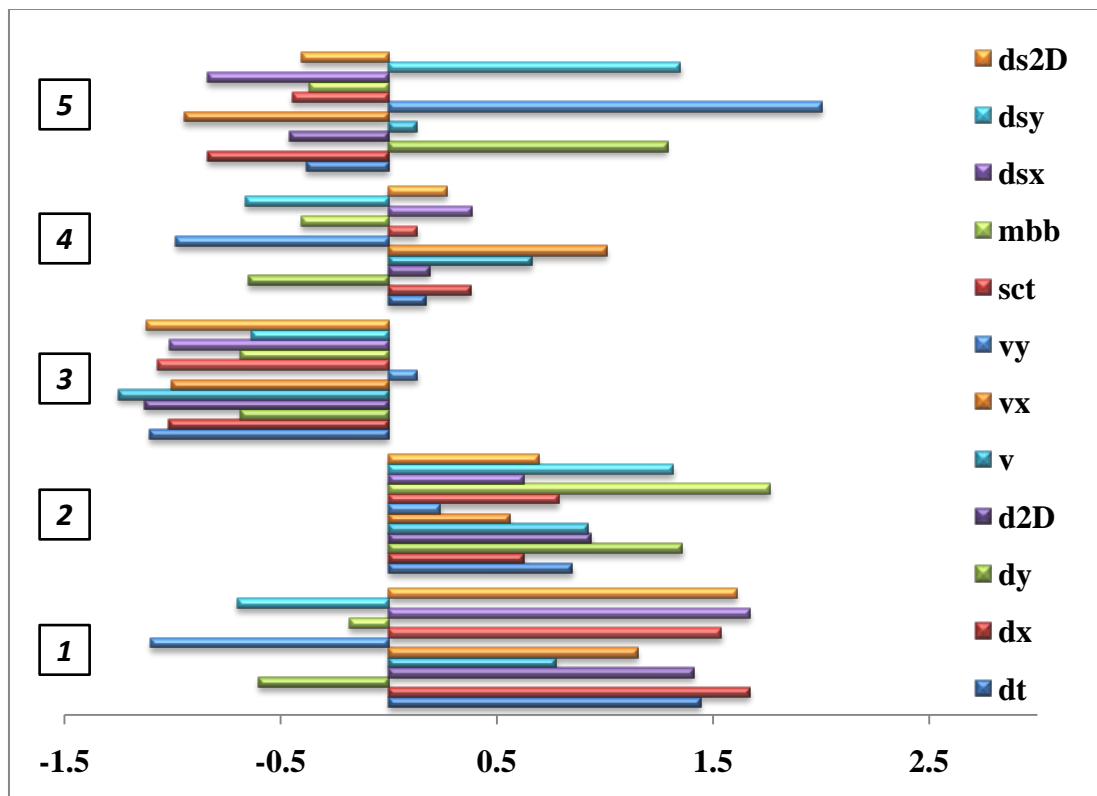


Figure 1. 2D images of sub-trajectories by each cluster ($k=5$)

Figure 2. Sub-trajectory cluster profiles ($k=5$)

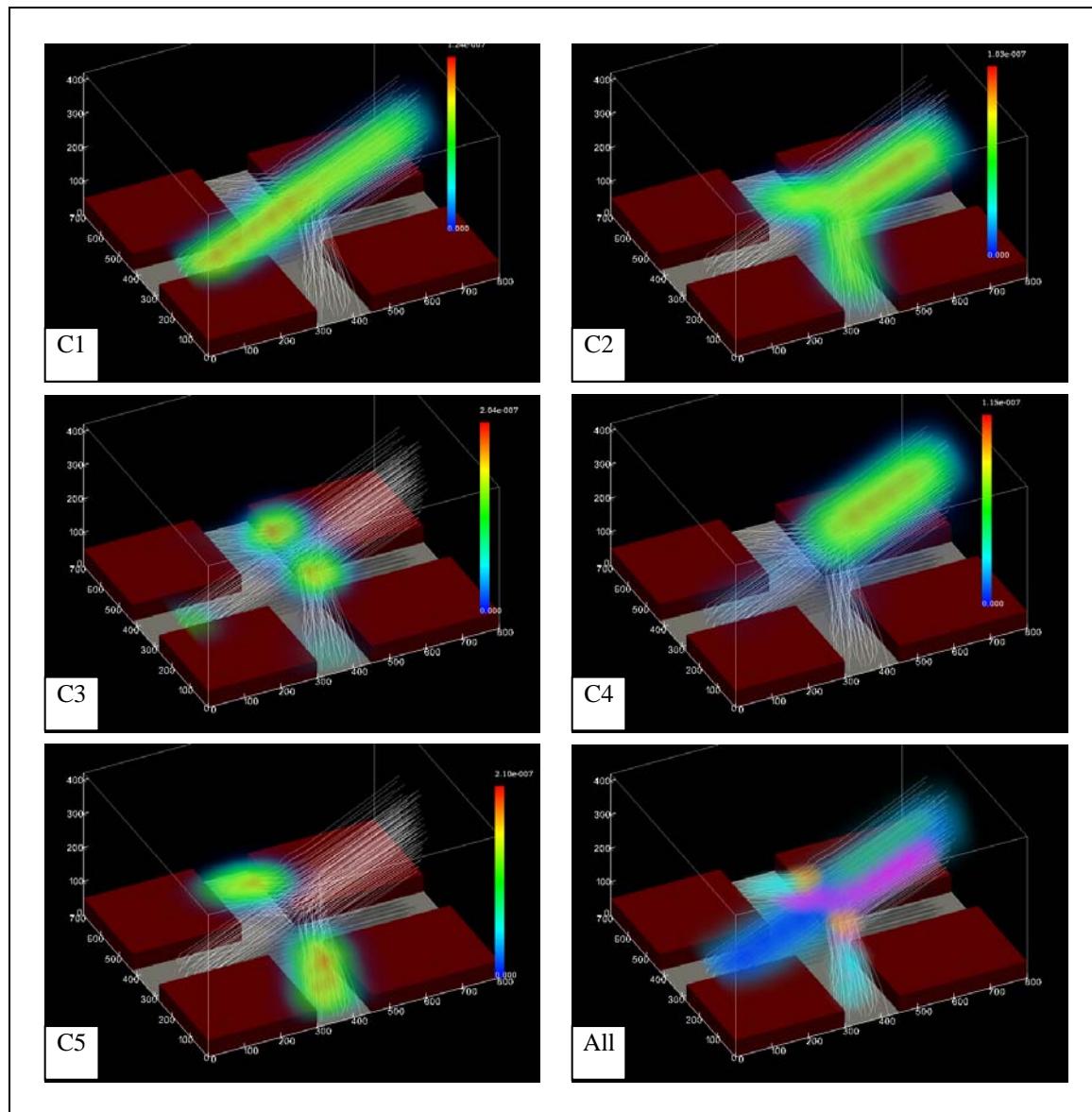


Figure 3. Sub-trajectory cluster distributions in space and time ($k=5$)

4. References

- Andrienko, N., & Andrienko, G. (2011). Spatial Generalization and Aggregation of Massive Movement Data. *IEEE Transactions on Visualization and Computer Graphics*, 17(2), 205-219.
- Brunsdon, C., Corcoran, J., & Higgs, G. (2007). Visualising space and time in crime patterns: A comparison of methods. *Computers, Environment and Urban Systems*, 31, 52-75.
- Gaffney, S., Robertson, A., Smyth, P., Camargo, S., & Ghil, M. (2006). *Probabilistic Clustering of Extratropical Cyclones Using Regression Mixture Models*. Technical Report, UCI-ICS 06-02, University of California, Irvine.

- Guo, D., Liu, S., & Jin, H. (2010). A Graph-based Approach to Vehicle Trajectory Analysis. *Journal of Location Based Service*, 4(3), 183-199.
- Helbing, D., & Molnár, P. (1995). Social force model for pedestrian dynamics. *Physical Review E*, 51, 4282-4286.
- Henderson, A. (2007). *ParaView Guide, A Parallel Visualization Application*. Kitware Inc.
- Lee, J. G., Han, J., & Whang, K. Y. (2007). Trajectory clustering: A partition-and-group framework. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, (pp. 593-604). Beijing, China.
- Levoy, M. (1988). Volume rendering: Display of surfaces from volume data. *IEEE Computer Graphics & Applications*, 8(3), 29-37.
- Nakaya, T., & Yano, K. (2010). Visualising crime clusters in a space-time cube: an exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS*, 14(3), 223-239.
- Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a dataset via the Gap Statistics. *Journal of the Royal Statistic Society: B*, 2, 411-423.

An Adaptive-Velocity Time-geographic Density Estimator

J. A. Downs¹, M. W. Horner²

¹Department of Geography, University of South Florida, Tampa, FL USA
Email: downs@usf.edu

²Department of Geography, Florida State University, Tallahassee, FL USA
Email: mhorner@fsu.edu

1. Introduction

Time geography, theorized by Hägerstrand (1970) and extended by Miller (2005a), provides an elegant mathematical framework for analyzing the movements and interactions of spatial objects. The basic elements of time geography—the space-time path and prism—provide the foundation for a number of analyses in GIScience. Notable examples include: quantifying interactions among humans (Miller 2005b, Neutens et al. 2007b), understanding human spatial behavior (Schwanen and Kwan 2008, Shaw and Yu 2009), quantifying spatial uncertainty (Neutens et al. 2007a), measuring accessibility (Novak and Sykora 2007), and modelling traffic flows (Raubal et al. 2007, Kuijpers et al. 2010).

While time-geographic techniques are well established in GIScience, recent work has aimed to transform the discrete mathematics of the space-time prism into a statistical formulation that can serve as the basis of a 'probabilistic' time geography. For example, Winter and Yin (2010a,b) quantified probability distributions for individual objects located within space-time prisms according to random-walks. They found this probability distribution was uneven, suggesting probabilistic space-time cones can be more informative than discrete ones. In other work, Downs (2010) developed a fixed-velocity time-geographic density estimator which computes a continuous probability density surface of a mobile object's location. This is accomplished given a set of control points in a space-time path and a specified maximum velocity parameter. This technique is useful for quantifying and visualizing the spatial distribution of a variety of moving objects, such as pedestrians or animals. Both of these studies highlight the need for continued development of probabilistic time-geographic techniques.

This paper builds upon the work of Downs (2010) by developing an adaptive-velocity time-geographic density estimator. First, fixed-velocity time-geographic density estimation is reviewed in order to discuss its limitations and to provide the mathematical foundation for the new technique. Second, the variable-velocity version is formulated. Finally, for comparison, both methods are applied to a sample tracking dataset that documents the movements of a pedestrian.

2. Fixed-velocity time-geographic density estimation

Time-geographic density estimation (TGDE), detailed by (Downs 2010), incorporates the fundamental elements of time geography with statistical density estimation. TDGE is used to generate a continuous probability density surface for a moving object over time given a set of observed control points. The technique operates in a manner analogous to kernel density estimation except instead of applying a kernel to each data point (Silverman 1986), a distance-weighted geo-ellipse function is fit to each

consecutive pair of control points in a space-time path. A geo-ellipse denotes all potentially reachable locations during the time interval between consecutive locations; in other words, the geo-ellipse is a geometric footprint of the space-time prism for a specified time period.

Mathematically, TGDE can be formulated as:

$$\hat{f}_t(x) = \frac{1}{(n-1)[(t_n - t_1)v]^2} \sum_{i=1}^{n-1} G\left(\frac{\|x - x_i\| + \|x_j - x\|}{(t_j - t_i)v}\right),$$

where $\hat{f}_t(x)$ is the time geographic density estimate at any point x in a map, n denotes the number of control points which are indexed consecutively as i and j , t records the timestamp, v is the object's maximum velocity, and G is a distance-weighting function of the geo-ellipse. The maximum velocity is assumed to be constant over the tracking duration and must be specified by the user. Its value can be determined by either the theoretical or the observed maximum velocity of the object. The geo-ellipse function must be a discrete distance-weighting operator, such as a linear decay function, in order to ensure that no intensity value is computed at locations where the object could not have been located given constraints imposed by the control points. The geo-ellipse function operates on the distance between an evaluation point x and each of two consecutive control points, which is divided by the maximum distance the object could have travelled during the time interval given its maximum velocity. The sum of weighted distances at each x is then multiplied by 1 divided by $n-1$ geo-ellipses times the square of the maximum possible travel distance. Once intensities are computed for all locations in the map in this manner, a continuous probability density surface of the object's spatial position during the tracking interval is realized. Since, this method assumes the maximum velocity is constant throughout the tracking interval, we now refer to the method as fixed-velocity TGDE.

3. Adaptive-velocity TGDE

A potential limitation of fixed-velocity TGDE is that it assumes the maximum velocity of the object is constant over the tracking duration. However, in practice, the speed of an object is likely to vary over time, depending on its behavior, the situation, or other factors. As such, time-geographic density estimates could be improved if lower maximum velocities could be specified for tracking intervals where the object travelled at less than maximum speed. We propose adaptive-velocity TGDE to accomplish this task. The formulation is similar to that for fixed-velocity TGDE except for two changes. The Constant from the previous formula is replaced with the maximum velocity for each space-time path segment for each pair of control points. Additionally, the maximum length of the space-time path is computed by summing the maximum path segment lengths as computed from the maximum velocities and the corresponding elapsed times. Otherwise, the density estimates for variable-velocity TGDE are computed in the same manner as the fixed-velocity version.

4. Application to Pedestrian Tracking Data

This section compares fixed-and adaptive-velocity TGDE using pedestrian tracking data. The methods are illustrated using a hypothetical tracking dataset of 60 points. The first 20 points represent a running movement, the second 20 points are clustered in a single location of relative inactivity, and the last 20 points represent walking.

Figure 1 shows a resulting fixed-velocity time-geographic density surface for a sample of 60 control points that record movements of a pedestrian; note the shaded area delineates the object's potential path area (Miller 2005a) or all locations it could have potentially visited during the time duration. The approximate trajectory of the pedestrian is illustrated using a space-time path constructed by connecting adjacent points with straight-lines.

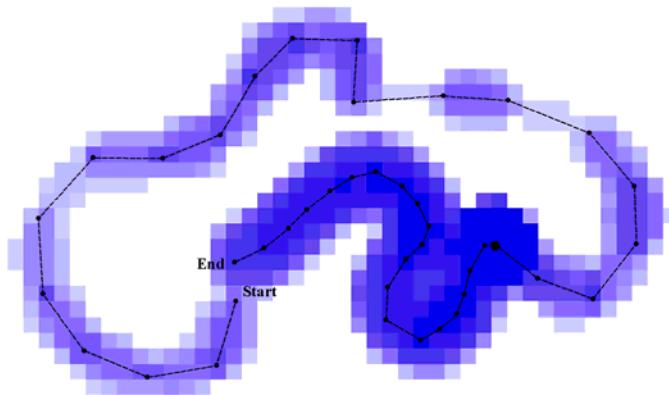


Figure 1. Fixed-velocity time-geographic density surface

Figure 2 illustrates adaptive-velocity TGDE for the pedestrian data, using different maximum velocities for each of the three portions of the space-time path. This example illustrates how the adaptive-velocity TGDE surface more precisely delineates the possible areas where the pedestrian was located during the tracking interval as compared to the fixed-velocity surface. While the computed density surfaces are identical for the first portion of the trajectory, there are clear differences in the latter two segments, where the potential path area is much narrower. In situations like these, where maximum velocities can be specified in greater spatial detail, adaptive-velocity TGDE can produce more accurate probability density surfaces than its fixed-velocity counterpart. This research further evaluates both techniques in the context of mapping pedestrian movements using GPS tracking data collected for students walking on a university campus.

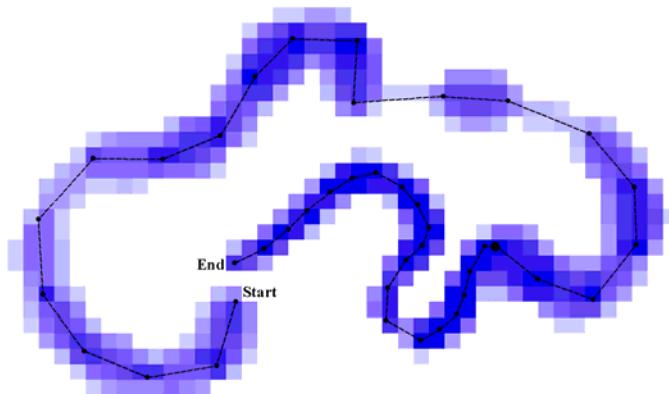


Figure 2. Adaptive-velocity time-geographic density surface

Acknowledgements

Portions of this research are supported by grants made to the authors by the National Science Foundation (NSF). The contents of this abstract represent the views of the authors and not those of NSF.

References

- Downs JA, 2010, Time-geographic density estimation for moving objects. *GIScience Proc, Lecture Notes in Computer Science*, 6292: 16-26.
- Haagerstrand, T., 1970, What About People in Regional Science, *Papers of the Regional Science Association* 24: 7-21.
- Kuijpers, B., H. J. Miller, T. Neutens and W. Othman, 2010, Anchor Uncertainty and Space-Time Prisms on Road Networks. *International Journal of Geographical Information Science*, 24(8): 1223.
- Miller HJ, 2005a, A measurement theory for time geography. *Geographical Analysis* 37: 17-45.
- Miller, H.J., 2005b, Necessary space-time conditions for human interaction. *Environment and Planning B-Planning & Design*, 32:381-401.
- Neutens, T., Witlox, F., De Weghe, N.V. and De Maeyer, P., 2007a. Human interaction spaces under uncertainty. *Transportation Research Record*, 28-35.
- Neutens, T., Witlox, F., De Weghe, N.V. and De Maeyer, P.H., 2007b. Space-time opportunities for multiple agents: A constraint-based approach. *International Journal of Geographical Information Science*, 21:1061-1076.
- Novak, J. and Sykora, L., 2007. A city in motion: Time-space activity and mobility patterns of suburban inhabitants and the structuration of the spatial organization of the Prague metropolitan area. *Geografiska Annaler Series B-Human Geography*, 89B:147-168.
- Raubal, M., Winter, S., Tessmann, S. and Gaisbauer, C., 2007. Time geography for ad-hoc shared-ride trip planning in mobile geosensor networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62:366-381.
- Schwanen, T. and Kwan, M.P., 2008, The Internet, mobile phone and space-time constraints. *Geoforum*, 39:1362-1377.
- Shaw, S.L. and Yu, H.B., 2009, A GIS-based time-geographic approach of studying individual activities and interactions in a hybrid physical-virtual space. *Journal of Transport Geography*, 17:141-149.
- Silverman BW, 1986, Density estimation for statistics and data analysis. London: Chapman Hall.
- Winter, S. and Z. Lin, 2010a, 'The Elements of Probabilistic Time Geography'. Forthcoming in *GeoInformatica*. doi 10.1007/s10707-010-0108-1
- Winter, S. and Z.-C. Yin, 2010b, 'Directed Movements in Probabilistic Time Geography', *International Journal of Geographical Information Science*, 24(9): 1349 - 1365.

Dynamic Planning Of Ambulance Location in Leicestershire

Emeka Chukwusa¹, Alexis Comber², Chris Brunsdon³

University Of Leicester, Department of Geography, University Road, Leicester. LE1. 7RH.

Telephone: +44(0)1162523823

Fax: +44(0)1162523854

Emails: ec102@le.ac.uk , ajc36@le.ac.uk , cb179@le.ac.uk

1. Introduction

Health facility location planning is an important for improving accessibility, coverage and in the case of ambulances, response times. Location-Allocation problems concerns the spatial locations of resources (Brandeau and Chiu, 1989) and involves strategic decision-making in relation to where to locate, how to allocate and what time constraints to consider to ensure adequate coverage of demand. For example, a typical ambulance location problem involves making decisions on where to locate ambulances or paramedic facilities to minimise response times (Sasaki et al., 2009; Comber et al., in press). In order to solve this problem, several location modelling techniques are described in the literature. These techniques coupled with GIS spatial analysis, have proven to be very efficient in addressing location-allocation problems (e.g: Kumar, 2004; Oppong and Hogson, 1994). Though previous studies have applied these techniques, one limitation in their application is the assumption that demands for public health facilities are static and fixed. This is contrary to real life scenarios where demands vary generally spatially at various time of the day, as people move from their place of residence to other locations (e.g journey to work).

Given this limitation, it is important to develop approaches that locate emergency medical services strategically in order to account for spatio-temporal variations in demand. Optimising ambulance and paramedic service location by considering such variations will improve ambulance response times to emergency cases. Ambulance response times are critical for patient survival especially for severe emergency cases, where time to pre-hospital treatment is critical (Snyder et al., 2007).

This paper addresses this gap by using a modified *P-median* model to optimise ambulance service locations for a spatio-temporally varying population based on journey to work data in Leicestershire.

2. Approach

The objective of this work involves identifying the optimum location of ambulances by using a modified *P-median* model while accounting for any spatio-temporal variation in demand. The *P-median* selects subset of facilities known as *P* facilities from a set of candidate facilities that minimises the aggregate travel or time between demand points and nearest facility locations (Fotheringham et al., 1995). The classical *P-median* model first espoused by ReVelle and Swain (1970) was modified to account for spatio-temporal variation by incorporating a spatial variation (x_i, y_i) and time component (t_n). See equation 1.

$$\text{Minimise } Z = \sum_{i=1}^m \sum_{j=1}^n a_{i(x_i,y_i)t_n} * d_{i(x_i,y_i)t_n} * j \quad (1)$$

$I....m.$ = Set of demand locations (centroid points).

$J....n.$ = Set of ambulance station location (20 ambulance stations).

$(x_i, y_i).$ = Location coordinates (showing spatial variation in demand).

t_n = Time (showing temporal change in demand).

d_{ij} = Shortest distance between demand and ambulance station locations

a_i = weight of demand node i at time t_n .

A typical location-allocation problem involves selecting optimum location choices from a pool of candidate location and allocating demand to these points. In this study, the pool of candidate location consists of 20 ambulance stations in Leicestershire, with a choice of selecting 12 ambulance stations to allocate to 583 demands. Finding solution for this type of problem is computationally difficult because the solution search space is large. For example, choosing 12 ambulances from a set of 20 ambulances requires a solution search space of $12!/12!(20-12)!$ possible solution.

Deriving solution for this problem involve the application of heuristics. Teitz and Bart heuristic was applied to solve the *P-median* problem. It is interchange heuristics that selects a set of initial random solutions and improves their outcome by swapping until there are no further improvements in the solution of the objective function (Teitz and Bart, 1968).

The modified *P-median* model helps to answer certain questions that arise in spatial health planning such as for example: i) where should ambulances be located when demand is spatiotemporally varying? ii) What ambulances resources should be allocated to certain areas to ensure that demand is covered at a particular time? iii) Where are the best places to locate an ambulance or paramedic facility to minimise the response time (distance) prior to an emergency event at specific time?

3. Results

Figures 1 and 2, illustrates the optimum location points of limited number of ambulances for day and night time population respectively. The location model suggests a better way to allocate ambulance resources during the day and night to ensure optimal coverage and minimise ambulance response time.

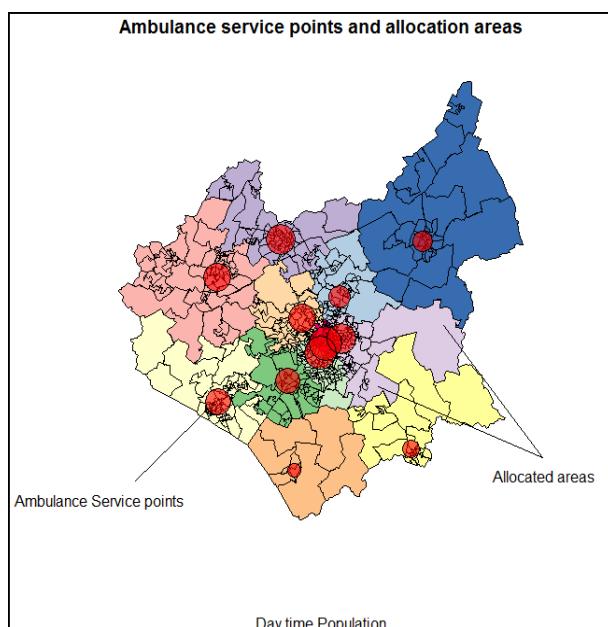


Figure 1.0 Ambulance location points and allocation areas:
Day time population in Leicestershire

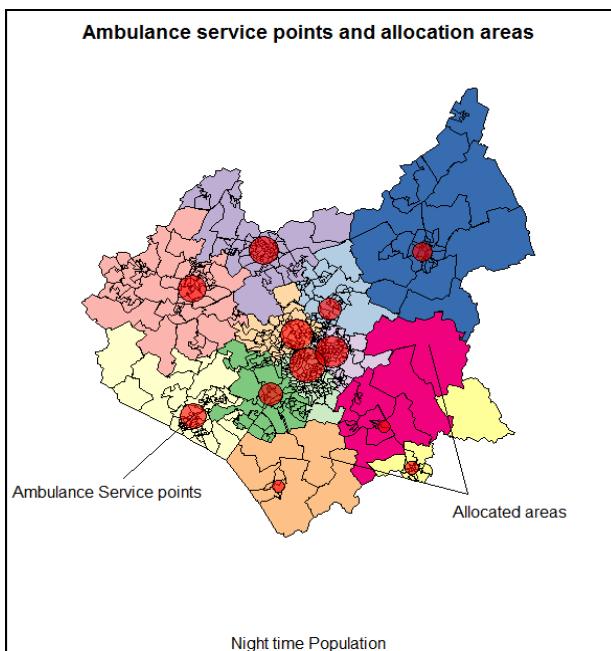


Figure 2.0 Ambulance location points and allocation areas:
Night time population in Leicestershire

Optimum locations for ambulances are represented with proportional circles. Each circle is located on the region they serve, with their sizes relative to the proportion of population (demand) they cover. For example, a larger circle signifies more demand allocations. In addition, lower super output areas (LSOAs) with similar hue are served by the same ambulance. The results will be described and discussed in more detail during the conference presentation.

4. References

- Brandeau, M.L. & Chiu, S.S., (1989) An overview of representative problems in location research. *Management*, 35(6), pp.645-674.
- Comber, A.J., Sasaki, S., Suzuki, H. and Brunsdon, C., (in press). A modified grouping genetic algorithm to select ambulance site locations. *International Journal of Geographical Information Science*. doi:10.1080/13658816.2010.501334
- Fotheringham,A. S, Curtis A, Densham , P. J (1995) The zone definition problem and location-allocation modelling. *Geographical Analysis* 27:60–77.
- Kumar, N., (2004) Changing geographic access to and Locational efficiency of health services in two Indian districts between 1981 and 1996. *Social Science & Medicine*, 58, pp.2045-2067.
- Oppong, J. R. & Hodgson, M. J (1994). Spatial accessibility to health facilities in Suhum district, Ghana, *Professional Geographer* 46 (2) 199±209.
- ReVelle, C.S. & R.W. Swain (1970) Central facilities location. *Geographical Analysis*, 2: 30-42.
- Sasaki, S., Comber, A.J., Suzuki, H. and Brunsdon, C., (2010). Using genetic algorithms to optimise current and future health planning - the example of ambulance locations. *International Journal of Health Geographics*, 9: 4. doi:10.1186/1476-072X-9-4

Snyder, D.E., White, R.D. & Jorgenson, D.B. (2007). Outcome prediction for guidance of initial resuscitation protocol: Shock first or CPR first. *Resuscitation*, 72, 45–51.

Teitz, M. B., and P. Bart. (1968). Heuristic Methods for Estimating the Generalized Vertex Median of a Weighted Graph. *Operations Research* 16:955–6.

Locating-allocation schools using metaheuristics and GIS

R. Zurita-Milla, Md. S. Arifin and O. Huisman

University of Twente, Faculty of Geo-Information Science and Earth Observation (ITC)

Hengelosestraat 99, 7500 AA Enschede, The Netherlands

Phone: +31 (0)53 4874 367

Fax: +31 (0)53 4874335,

E-mail: {zurita-milla, arifin21412, huisman}@itc.nl

1. Introduction

Location-allocation (LA) problems deal with the search of optimal locations for one or more facilities with respect to the spatial distribution of the demand which, in turn, should be allocated (assigned) to these facilities (Church, 1999).

LA problems have been an active research topic for more than a century, since Alfred Weber formulated his “Theory of the Location of Industries” in 1909. The extension of Weber’s work resulted in one of the most well known types of LA problems, the so-called p-median problem. This problem consists on finding p facility locations such that the total sum of the distance between the demand and the facilities is minimised.

Since Weber’s original work, LA problems have attracted a lot of attention in the scientific literature as their solution is relevant for optimally placing both public and private facilities like hospitals, supermarkets, distribution centres, and so forth (Li and Yeh, 2005; Sasaki et al., 2010). However, finding optimal solutions to LA problems has proven to be far from trivial as they are NP-hard (Murray, 2010).

LA problems are made up of several components: the type and number of facilities, the nature of the demand, and the geographic space where the problem should be solved (e.g. continuous vs. networks). Because of the computational complexity and the diversity in types of LA problems, numerous algorithms and optimization methods have been developed (Murray, 2010). Basically, LA “solvers” can be categorized into 3 classes: 1) exact methods, which only exist for simple problems or that require complete counting of all possibilities, 2) heuristics, which are fast methods to find “reasonable” solutions, and 3) metaheuristics, which are robust methods that find better solutions than the ones found by heuristics as they are usually not trapped in local optima.

This work explores the use of two well known metaheuristics, namely genetic algorithms and simulated annealing, to solve LA problems. In addition, it makes use of geographic information system (GIS) to store, to pre-process and to visualize the spatial datasets that are typically needed in LA problems. The following two sections briefly review each of these metaheuristics, and a case study is presented in the final section to illustrate our implementation.

1.1 Genetic algorithms

Genetic algorithms (GAs) were introduced by Holland (1975). These are based on the evolutionary idea of the “survival for the fittest”. As a consequence, terms like individual, population, reproduction, selection, crossover and mutation are commonly used in this search technique.

A population is a set of individuals, each of which represents a potential solution to the problem. The fitness of individuals is judged by the value of the objective function that they yield, where the objective function is the function that we seek to optimize. Selection, mutation and crossover are three main operators that control the evolution of the population until a stop criterion is met. The best individual in this final population represents the solution to the problem (e.g. the value that minimizes a given function).

GAs are efficient in finding near optimal solutions for complex optimization problems, which explain why they have been applied in many disciplines (Goldberg 1989). Hosage and Goodchild (1986) presented one of the first papers using GAs to solve LA problems in a GIS context.

1.2 Simulated annealing

Simulated annealing (SA) was inspired by the process of crystals forming during the cooling of metals/minerals (Kirkpatrick et al., 1983 and Černý, 1985). Its methodology is based on neighbourhood search. An initial solution is first proposed, and then the simulated annealing continues to seek for better solutions through iteration. For each iteration, SA randomly chooses a new solution if its fitness value is better than that of the previous solution. If the fitness value is not better, SA can still select it according to a probability function that depends on the cooling temperature.

SA is a serious competitor to GAs and it is worth to compare their results as both metaheuristics are derived from analogy with natural systems and deal with optimization problems of same type (Sivanandam and Deepa, 2008).

2. Case study and implementation

The work is illustrated with data from the city of Enschede (The Netherlands). It draws on public primary school data to establish a simple and generic school planning scenario. This planning scenario aims to evaluate the location of the current public primary schools (e.g. finding “the worst” current locations, if one school should be closed-down) and to propose the construction of new facilities or the relocation of existing ones.

In The Netherlands there are no official school districts. This means that children can attend whichever school that they prefer. Most parents send their children to the closest school unless they have some preference for an educational method.

From a practical point of view, the school planning scenario is based on the following available spatial data:

- location and capacity of public primary schools
- demographic data for each neighbourhood, and
- topologically consistent main road network data

For the implementation of the system a tool was built by loosely coupling free and open source programs/libraries. In particular, we used C# as development environment, the GA library developed by Samir (2006) and the SA code of Hamdar (2008). This is because, although GIS offers many core processing and analysis tools for LA (like buffers, network analyst, measuring distances, and so on), to our knowledge no GIS package offers a comprehensive, flexible and efficient “solver” for this kind of problem. The gvSIG GIS package was used to compute the origin-destination matrix (road network distances from the origin –demand points- to the destination –schools-).

The main characteristics of our implementation can be summarized as follows:

- it is possible to load shapefiles to delineate the study area and the basic administrative units (neighbourhoods in our case) as well as to spatially represent the location of the existing schools and, if available, the spatial distribution of the demand.
- it is possible to randomly generate demand and facility points in each administrative unit; where the demand points represent the children living in each neighbourhood and the facility points indicate the new potential locations for primary schools.
- It is possible to use Euclidian and network based distances.
- It is possible to customize the objective function to either make it a capacitated p-median problem (minimize sum travelled distances) or to make it a multi-objective problem (e.g. minimize travelled distance while maximizing the matching of school preference of each child and the type of school).
- It is possible to load a “mask” to exclude certain city areas from the analysis. For instance, demand and facility points cannot occur on parks, lakes, industrial areas, etc.
- It is possible to export the results as a shapefile.

After the optimization, three main outputs are provided (c.f. Figure 1):

1. a spider graph visualization with the selected schools and their allocated demand points is provided.
2. the total travelled distance for each selected schools. This information is valuable to plan future schools in areas where this distance exceeds a given threshold.
3. The value of the objective function at each iteration (generation) and a log-file with the input configuration and the CPU time used. This can be used to compare results between metaheuristics but also to optimize their configuration parameters and to study the sensitivity and the uncertainty of the results.

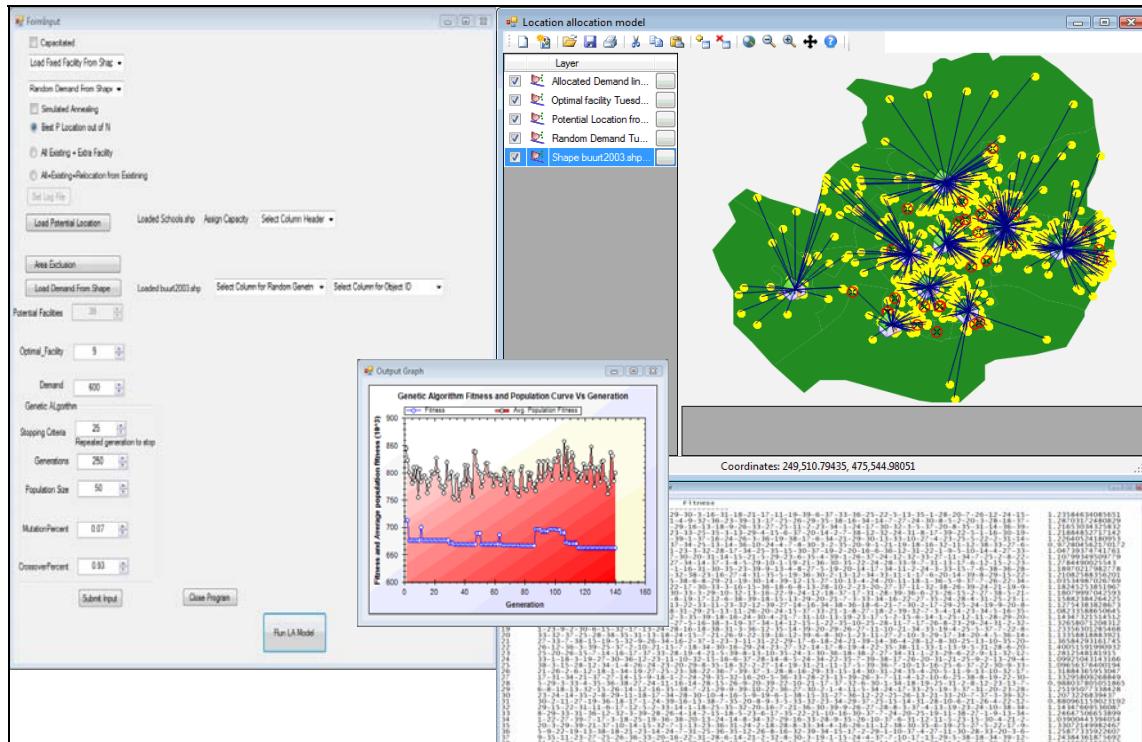


Figure 1. A screenshot of the software

3. References

- Černý V, 1985, Thermodynamic approach to the travelling salesman problem: an efficient simulation algorithm. *Journal of optimization theory and applications*, 45:41-51
- Church RL, 1999, Location modeling and GIS. In: *Geographical information systems - Volume 1*, Longley PA, Goodchild MF, Maguire DJ and Rhind DW (editors). John Wiley & Sons, West Sussex, UK, 293-303.
- Goldberg DE, 1989, *Genetic Algorithms in Search, Optimisation and Machine Learning*. Addison-Wesley. Reading, USA.
- Hamdar A, 2008, Simulated Annealing - Solving the Travelling Salesman Problem <http://www.codeproject.com/KB/recipes/simulatedAnnealingTSP.aspx> (accessed 17-01-2011).
- Holland JH, 1975, *Adaptation in Natural and Artificial Systems*. The University of Michigan Press. Ann Arbor, USA.
- Hosage CM and Goodchild, MF, 1986, Discrete space location-allocation solutions from genetic algorithms. *Annals of Operations Research*, 6:35-46.
- Li, X. and Yeh, AG-O, 2005, Integration of genetic algorithms and GIS for optimal location search. *International Journal of Geographical Information Science*, 19:581-601.
- Murray A., 2010, Advances in location modeling: GIS linkages and contributions. *Journal of Geographical Systems*, 12: 335-354.
- Kirkpatrick S, Gelatt CD and Vecchi MP, 1983, Optimization by Simulated Annealing. *Science*, 220: 671-680.
- Samir S, 2006, TSP (Travelling Salesman Problem) using Genetic Algorithm http://www.codeproject.com/KB/cs/TSP_Solved_using_GA.aspx (accessed 17-01-2011).
- Sasaki S, Comber A, Suzuki H, and Brunsdon C, 2010, Using genetic algorithms to optimise current and future health planning - the example of ambulance locations. *International Journal of Health Geographics*, 9:4
- Sivanandam SN and Deepa SN, 2008, *Introduction to Genetic Algorithms*, Springer, Berlin, Germany.

A New Variable for Spatial Accessibility Measurement in Social Infrastructure Planning

Y. Li¹, A. J. Brimicombe²

^{1,2}Centre for Geo-Information Studies
University of East London, University Way, London E16 2RD
Telephone: 0044 2082232603, 2352
Fax: 0044 2082232918
Email: y.li@uel.ac.uk, a.j.brimicombe@uel.ac.uk
<http://www.uel.ac.uk/geo-information>

1. Introduction

The quantitative analysis and modelling of networks and their flows as processes of spatial organisation came sharply into focus in the 1960s (e.g. Chorley & Haggett 1967) and, along with spatial cognition and navigation/wayfinding, remain on the research agenda for GIScience (e.g. Duckham et al. 2003). Accessibility commonly refers to the ease with which something or somebody can be reached, which from a spatial perspective also implies nearness (distance) or ease of travel. In this study, carried out in the context of social infrastructure planning in the UK, equitable access to facilities across a region shows a consistent match between supply and demand and can be taken as an indicator of good planning (Rosero-Bixby 2004, HUDU 2007). Thus measures of spatial accessibility are a means of analysing inequalities within the organisation of social services (Waters 2000). To provide a variable that captures accessibility at UK census Output Area (OA) level as an input to analyses of inequalities and to inform social infrastructure planning, an approach has been developed and tested which is less computationally intensive and overcomes some of the disadvantages of conventional approaches.

2. Spatial Accessibility in Social Infrastructure Planning

Social infrastructure in the UK normally includes: healthcare, education, community facilities, emergency and other essential services. In order to ensure that these services are delivered effectively and comprehensively, social infrastructure should be well planned for new developments, regeneration and in rationalising the efficient use of available resources. Within a planning area, inequality in local social services is relative, as it is increased by imbalances in supply and demand across space. Thus relative spatial accessibility needs to be measured rather than in absolute terms. Meanwhile, social infrastructure planning in UK is mainly carried out at a local scale, for example by Local Authorities and Primary Care Trusts (PCTs - soon to be superseded by GP consortia). Details of spatial accessibility in relation to relevant social-economic variables are thus desirable by small area geography.

From a GIS perspective, conventional approaches to accessibility normally include concentric buffers and polygons of network drive/walk times around individual or groups of facilities. This puts the point of origin for any accessibility at the facility being accessed and is therefore a supply-side view. One problem with such approaches is transforming the zones thus produced into a variable that can be attached to administrative units for further analysis. A demand-side view starts with

where people live and estimates the distance required to access their nearest facilities. Recently, considerable effort have been made on travel cost calculation in order to achieve accurate and comprehensive measures, which at the same time lead to more complex and less compatible approaches. Furthermore, the use of networks for either view requires assumptions to be made not only of travel speeds, but mode of travel as the available network may differ. Some inappropriate assumptions might result in a poor understanding of local accessibility. For example, a study of accessibility to health services in Liverpool was based on the public transport network, but a survey showed that only 19% respondents took public transport to visit a GP. Approaches based on these two views can be seen in various measures for catchment profiling, travel impedance modelling and gravity modelling (Guagliardo 2004, Liu & Zhu 2004).

In current research, spatial accessibility measurement has been improved significantly with increasing complex methods and growing amount of detailed information. However, challenges remain to establish robust and flexible approaches for various applications and rapid ‘what-if’ analyses, which can balance a range of factors. In this study, a new variable is developed to measure the spatial accessibility by small area geography for social infrastructure planning from the perspective of where people live. The new variable leads to less computational burden than network-based methods whilst correlates well with them. By using updated time series data, local organisations are able to monitor and analyse spatial accessibility locally in a timely manner.

3. Average Weighted Distance by Small Area Geography

A new variable – called average weighted distance (AWD) - is tested for measuring the relative inequalities in spatial accessibility to local social services/facilities. It is based on the average distance of residents to a nearest facility by small area geography and further weighted by population distribution. No assumption is made for the new variable about travel mode. OA is chosen as the basic geographic unit, as it can be associated with many demographic and socio-economic variables.

Population distribution can exert significant influence on spatial accessibility measures (Langford et al. 2008). However there is a lack of up-to-date inter-census information about population distribution by small area geography. For each postcode unit, OS CodePoint provides a geographic delivery centroid as well as the number of domestic deliveries. This is used as a proxy of population distribution within an OA.

AWD is developed to measure the relative spatial accessibility for social infrastructure planning. It intends to compare the ease of travel rather than the exact travel cost. For rapid appraisals, a measure with less computational load will be preferred, if it correlates well with more complex distance measures. In GIS there are three measures of distance that are readily computable: Euclidean, Manhattan and network distance. The network distance can be taken as is (network geometry), or further refined to reflect impedances such as speed limit, congestion level (attributed network). Our study shows very strong correlation between Euclidean and network distances at OA level for both urban and rural districts (see case studies below). Euclidean distance is therefore adopted for AWD. The general principle for AWD is expressed in equation (1). Fig. 1 illustrates how AWD is derived within an OA.

$$OA_F = \frac{\sum (W_{ddp}^P \times D_F^P)}{N_{ddp}^{OA}} \quad (1)$$

where:

OA_F = OA average weighted distance to a facility

D_F^P = Euclidean distance from postcode centroid to nearest facility

W_{ddp}^P = weight equal to the number of domestic delivery points within each postcode

N_{ddp}^{OA} = total number of domestic delivery points for the OA

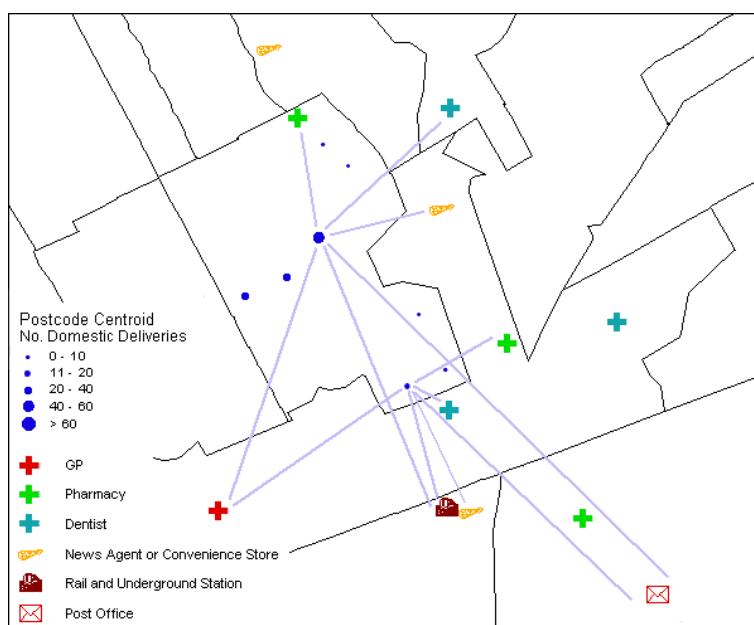


Figure 1. Deriving AWD for an OA (boundaries and postcodes Crown Copyright).

4. Case Studies

Two case studies have been carried out to investigate the feasibility of the new approach. Two case study areas are selected: Haringey is an inner London borough with high density urban population and Uttlesford is a rural district with small towns and dispersed villages and hamlets. Pharmacies were selected as the facility for modelling accessibility. To avoid boundary problems in the calculations, a buffer area is included for both districts. OS Integrated Transport Network is used for the network.

As illustrated in Fig. 2, Euclidean distances can be rapidly calculated using GIS whilst network distances need more editing and higher computational loads. The exact lengths are of course different for average weighted Euclidean and network distances. The median percentage difference is 28% (130m) for Haringey and 21% (670m) for Uttlesford. However, there are strong correlations between these two types of distance measures where coefficient is 0.936 for Haringey and 0.981 for Uttlesford (see Fig 3 for regression models). Such strong correlation supports the concept that AWD based on Euclidean distance can provide an acceptable relative measure, which aims to develop a straightforward measure for rapid assessments. A visual comparison of the results is given in Fig 4. This shows the relative inequalities in accessibility to a pharmacy. Thus if a local GP consortium wishes to evaluate filling the gaps in accessibility to commercial pharmacies, it can rapidly ‘what-if’ model additional pharmacies attached to existing surgeries.

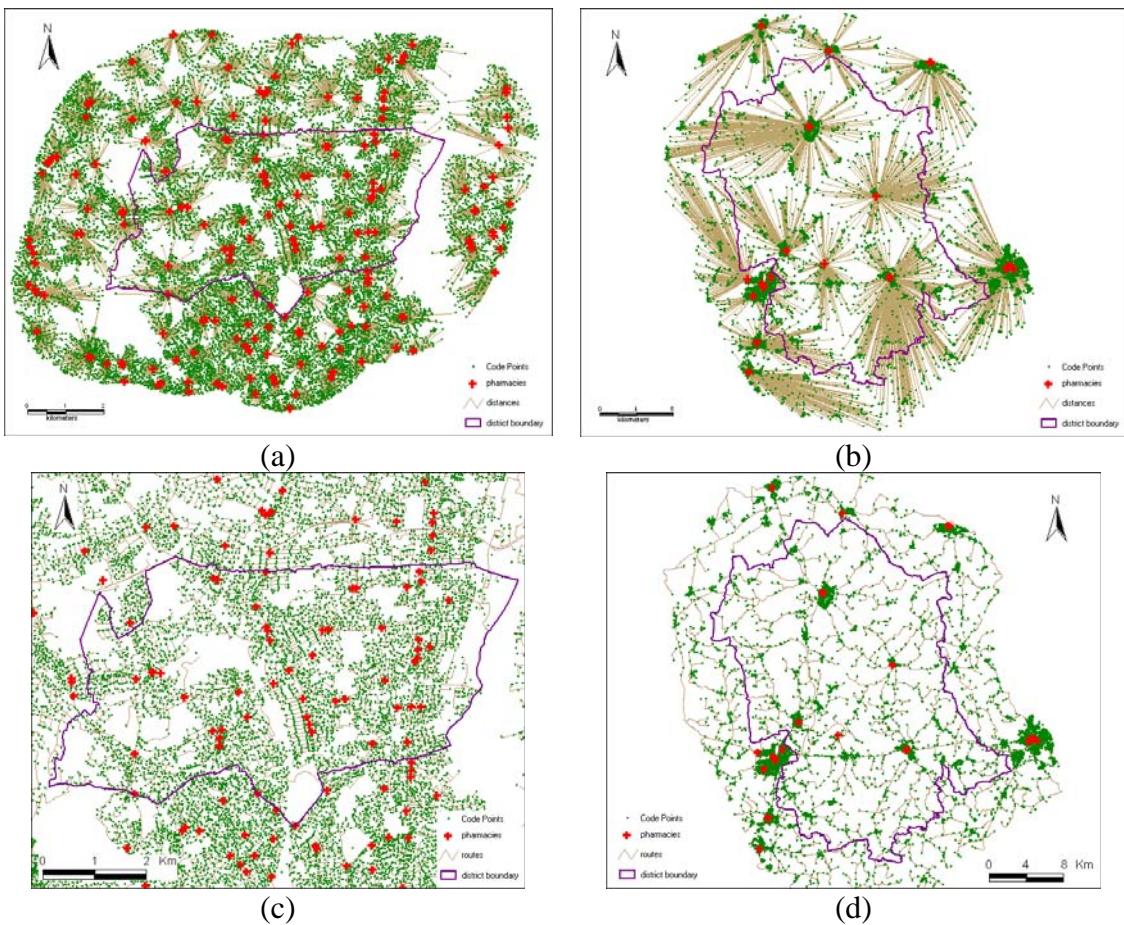


Figure 2. Distances from Code Points to nearest pharmacies (a) Euclidean distances for London Borough of Haringey, (b) Euclidean distances for Uttlesford District, (c) network distances for London Borough of Haringey, (d) network distances for Uttlesford District (data Crown Copyright).

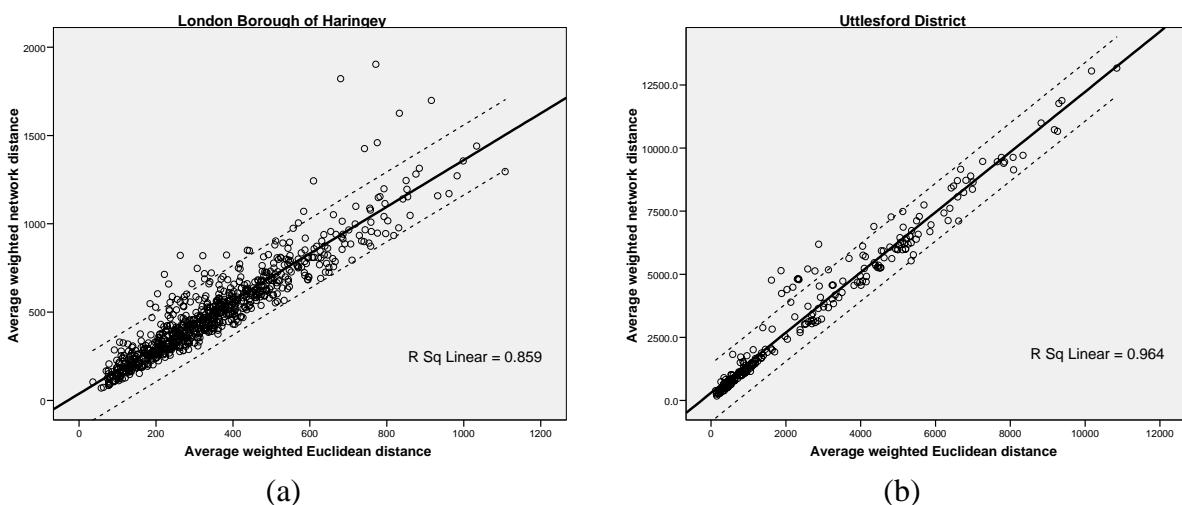


Figure 3. Regression between average weighted Euclidean and network distances, (a) London Borough of Haringey and (b) Uttlesford District.

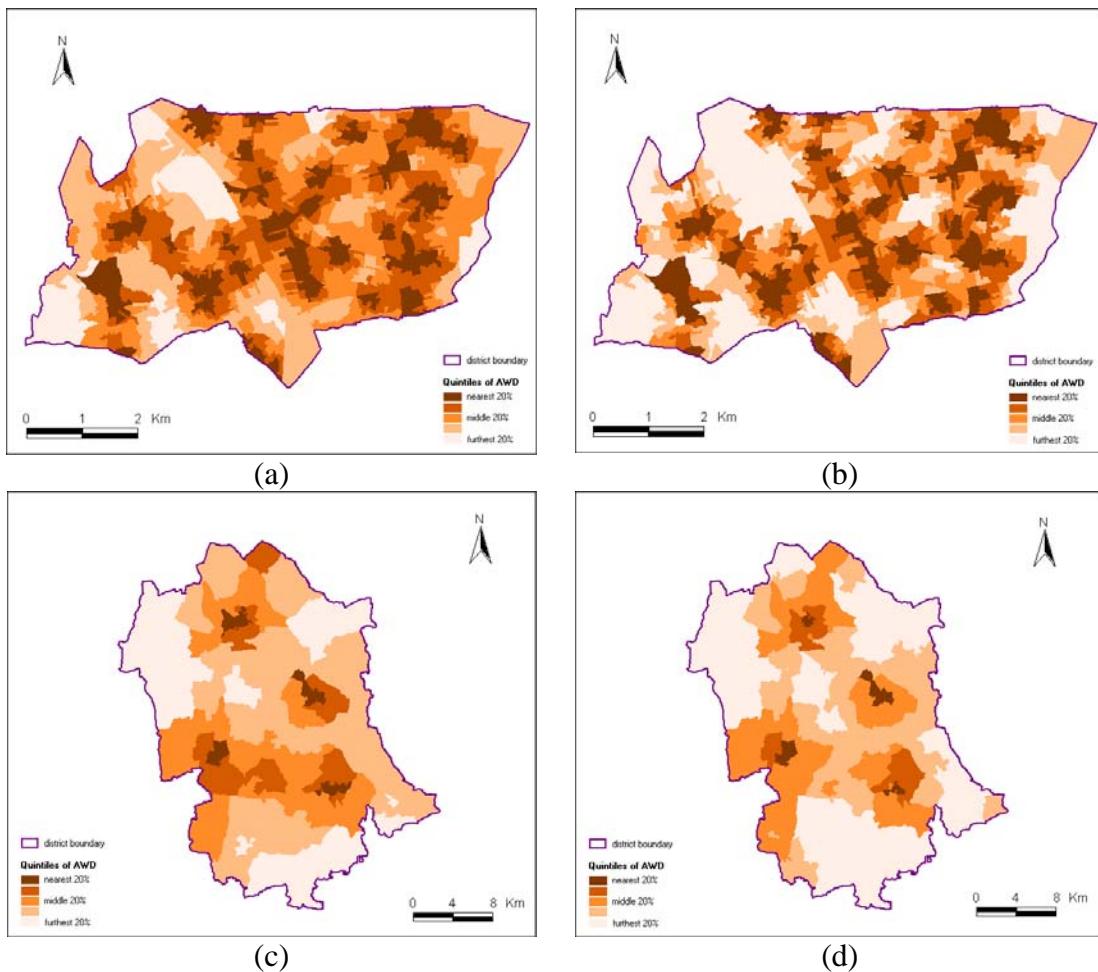


Figure 4. Patterns of local accessibility in quintiles (a) average weighted Euclidean distance for London Borough of Haringey, (b) average weighted network distance for London Borough of Haringey, (c) average weighted Euclidean distance for Uttlesford District, (d) average weighted network distance for Uttlesford District (boundaries Crown Copyright).

5. Conclusion

A new variable – AWD - is developed to measure and analyse spatial accessibility by small area geography. It will support rapid assessments of inequalities and ‘what-if’ analyses in local social infrastructure planning. The approach can use both Euclidean distance and network distance using postcode centroids as the atomic spatial unit. However, it is found that these two approaches have a high correlation and therefore similar patterns of relative inequality. The Euclidean distance approach has less computational load and is generally applicable, particularly where rapid ‘what-if’ analyses are required for decision support in a planning context. Local organisations are then able to interpret and further analyse relative local spatial accessibility for specific services/facilities as well as monitor changes in accessibility over time.

6. Acknowledgements

The authors wish to thank the British Academy for funding this study.

7. References

- Chorley RJ and Haggett P, 1967, *Models in Geography*. Methuen, London.
- Duckham M, Goodchild MF and Worboys MF, 2003, *Foundations of Geographic Information Science*. Taylor & Francis, London.
- Guagliardo MF, 2004, Spatial accessibility of primary care: concepts, methods and challenges. *International Journal of Health Geographics*, 3(1):1-13.
- HUDU, 2007, *Health and Urban Planning Toolkit*. NHS Healthy Urban Development Unit, London.
- Langford M, Higgs G, Radcliffe J and White S, 2008, Urban population distribution models and service accessibility estimation. *Computers, Environment & Urban Systems*, 32:66-80.
- Liu S and Zhu X, 2004, Accessibility Analyst: an integrated GIS tool for accessibility analysis in urban transportation planning. *Environment and Planning B: Planning and Design*, 31: 105-124.
- Rosero-Bixby L, 2004, Spatial access to health care in Costa Rica and its equity: a GIS-based study. *Social Science & Medicine*, 58:1271-1284.
- Waters HR, 2000, Measuring equity in access to health care. *Social Science & Medicine*, 51:599-612.

Temporal limits to urban growth modelling

Gargi Chaudhuri¹, Keith C. Clarke²

¹University of California Santa Barbara, 1832, Ellison Hall, Department of Geography
Telephone: (001) 805 452 3824
Fax: (001) 805-893-2578
Email: gargi@geog.ucsb.edu

²University of California Santa Barbara, 1832, Ellison Hall, Department of Geography
Telephone: (001) 805 308-2836
Fax: (001) 805-893-2578
Email: kclarke@geog.ucsb.edu

1. Introduction

A model is a representation of reality, so sensitivity testing is obligatory to establish credibility in complex models (Clarke, 2004). The success of a model is defined by how well its simulated maps match with the known maps (Pontius and Schneider, 2001). Sensitivity testing becomes more important when the models are used for decision making purposes. The present study uses a popular land use change model called SLEUTH to investigate the temporal sensitivity of the forecast maps. The objective of the study is to investigate – (a) the trend of uncertainty of the forecasted images from immediate future to distant future; and (b) the impact of prediction date range on the accuracy of the output images.

SLEUTH is a cellular automata model of land use change and has been applied extensively in geographic simulation of future planning scenarios (Clarke and Gaydos, 1998; Clarke et al., 2007). The model is a tightly coupled blend of two cellular automata models: Clarke's Urban Growth Model (UGM) and the Deltatron land use change model. It uses weighted maps of slope, land-use, exclusion, urban extent over time, transportation, and a hill-shaded backdrop used for visualization as inputs, divides the study area into square cells, and applies from year to year in sequence a set of CA rules that determine whether or not cells will change from non-urban to urban land use or among other land uses. Behavior of the CA is regulated by five parameters that control diffusive growth, outward spread, the degree of new center creation, and the influence of roads on the growth pattern.

2. Study Area and Data

Gorizia is a small town on the Isonzo River at the foothills of the Italian Alps, astride Italy's northeastern border with Slovenia. The study area covers 23 sq.kms of area covering the Italian city of Gorizia and its Slovenian counterpart called Nova Gorica (including the settlements of Solkan and Nova Gorica in the north to Sempeter in the south).

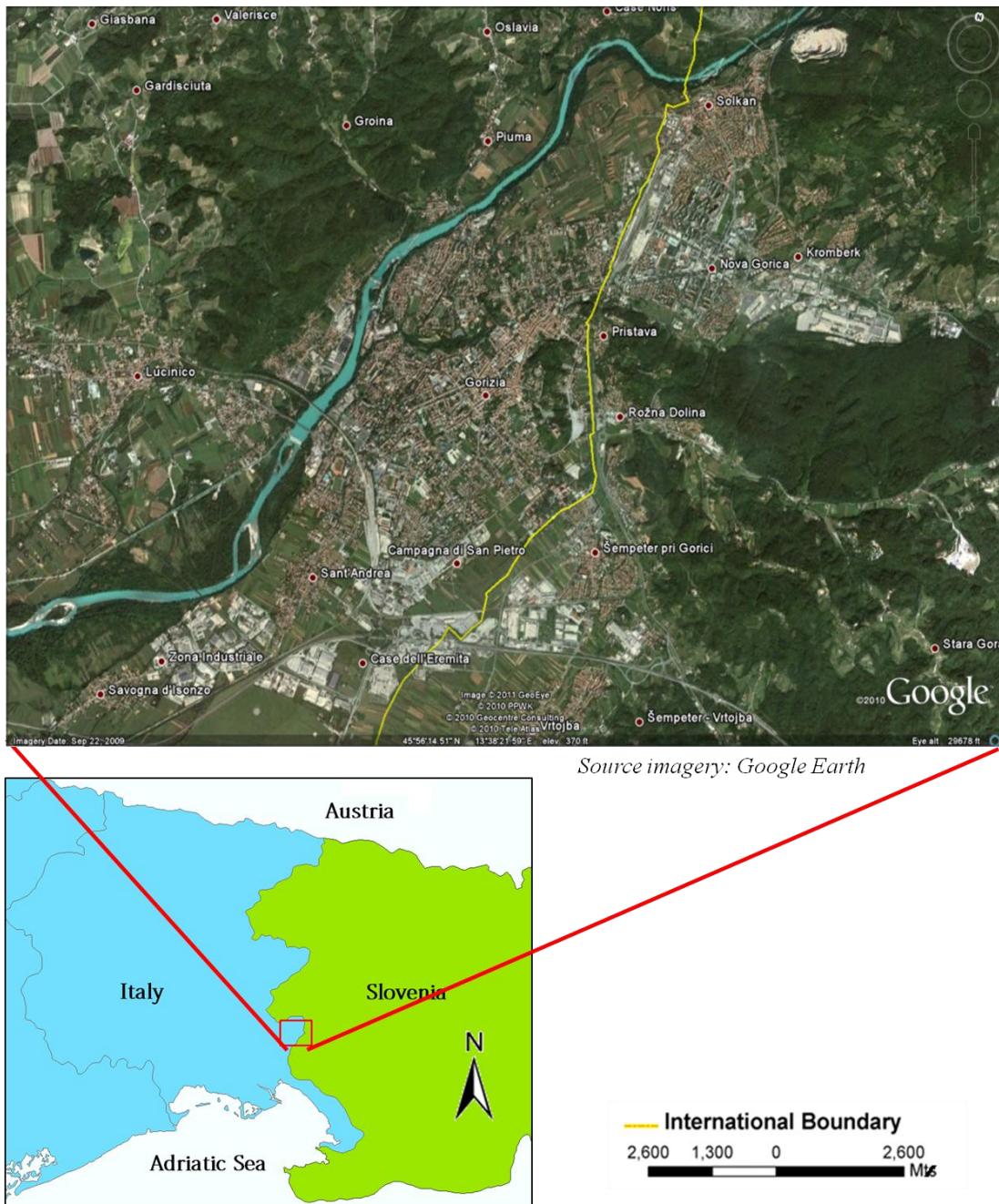


Figure 1: Gorizia, Italy and Nova Gorica, Slovenia - the study area

For modeling purposes, SLEUTH requires a minimum of a slope layer, hillshade layer (for background visualization), 2 land use layers, 4 urban layers, 2 transportation network layers and an exclusion layer. The exclusion layer is included in the model to restrain urban growth in areas where it is not possible to urbanize (for example, water bodies and protected areas). Details about the data preparation can be found in the model website (<http://www.ncgia.ucsb.edu/projects/gig/v2/About/dtInput.htm>). Images from Landsat 5 TM 1985, 1991, Landsat 7 ETM+ 1999 and Aster 2004 were used as urban layers, 1985 and 2004 images were used for the land use layer and 1968 and 1998 maps for the transportation layer. The transportation network was weighted according to the class of the road. The classified images of Landsat 5 TM 2005, 2006, 2007, 2009 and 2010 were used for validation of the predicted images. Classification accuracy of the data is provided below:

Image	Overall accuracy (%)	Kappa Coefficient
1985	86.20	0.80
1991	64.14	0.47
1999	75.51	0.63
2004	84.40	0.75
2005	88.40	0.83
2006	86.34	0.79
2007	88.89	0.83
2009	91.56	0.87
2010	91.19	0.87

Table 1: Classification accuracy of input data

3. Methodology

After rigorous calibration of the SLEUTH model using the data, the best fit parameter values were used to run a prediction to 2040 for the region. To fulfill the first objective, the accuracy of the forecast images (urban area) of 2005, 2006, 2007, 2009 and 2010 were validated against the observed map of those years. For the second objective, the prediction was run with the same set of parameters for 3 additional prediction ranges: 1970 – 2040, 1985 – 2040, and 2000 – 2040. The forecasted images of 2010 from all the predicted ranges were then validated

with the observed map of 2010 and compared against each other. Presently, the model uses the last input urban layer as the start date for prediction run. The urban and road images used to initialize growth during prediction are those with dates equal to, or greater than, the prediction start date. If the prediction start date is greater than any of the urban dates, the last urban file on the list was used. Similarly, if the prediction start date is greater than any of the road dates, the last road file on the list was used. The prediction run terminated at the prediction stop date (http://www.ncgia.ucsb.edu/projects/gig/v2/About/data_files/scenario_file.htm).

Validation was performed independently, using *Kappa* coefficient analysis in the Map Comparison Kit (MCK), (RIKS and MNP – RIVM, 2004) (Visser and Nijs, 2006). The *Kappa statistic* is computed from a confusion matrix derived from cell-by-cell comparison of the observed map and the predicted map (Hagen-Zanker and Martens, 2008). *Kappa* is based on the percentage of agreement is corrected for the fraction of agreement that can be expected by pure chance. *Klocation* (Pontius, 2000) compares the actual success space to the expected success rate relative to the maximum success space given that the total number of cells of each category does not change (Pontius, 2000). *Khisto*, according to Hagen (2002a), is an alternative expression for the similarity of the quantitative model, based upon the total number of cells taken in by each class and can be calculated directly from the histograms of two maps. *Kappa*, *KLoc* and *KHisto* are connected through the multiplicative relation: $Kappa = KLoc * KHisto$ (Visser and de Nijs, 2006). According to Landis and Koch (1977) the kappa indices can be discretely interpreted in the following way: < 0.00 = poor; 0.00 – 0.20 = slight; 0.21 – 0.40 = fair; 0.41 – 0.60 = moderate; 0.61 – 0.80 = substantial; 0.81 – 1.00 = almost perfect.

4. Results and Discussion

In the first test, we compared the accuracy of predicted urban images from 2005 till 2009 to find out the trend of uncertainty over time. The results show that all types of Kappa values, from the years 2005 to 2009, are within the class range 0.81 – 1.00 that means the prediction results are ‘almost perfect’.

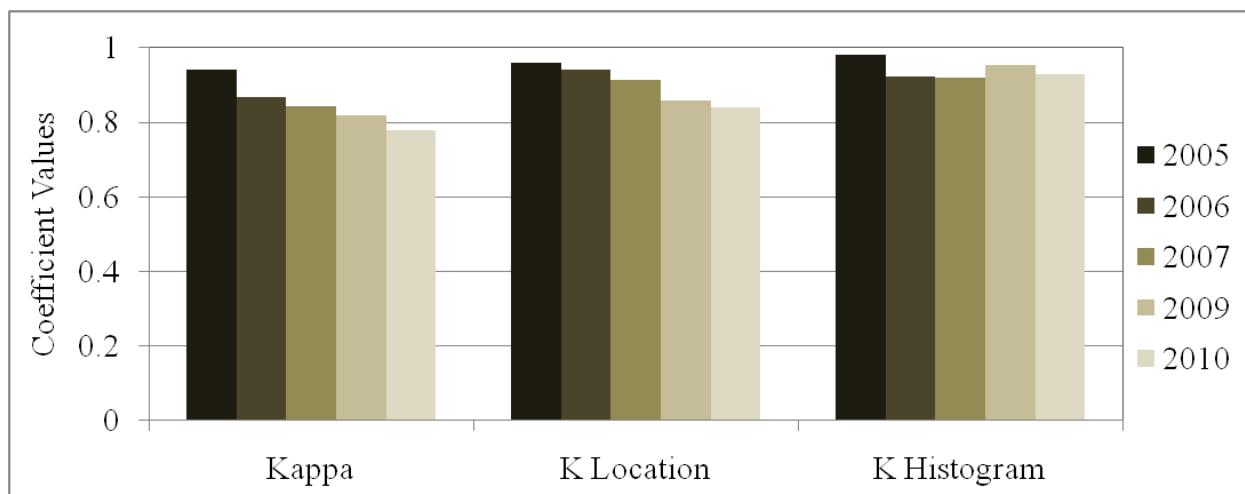


Figure 2: Decrease in accuracy of the predicted images over time

Only the Kappa coefficient of 2010 goes down to 0.78 which is in the class of ‘substantial’ accuracy though Kloc and Khisto for that year remains in the ‘almost perfect’ class. As expected, 2005 has the highest level of agreement whereas 2010 has the least. Interestingly, for Kappa and Kloc there is a decreasing trend of accuracy but for Khisto after the immediate decrease in 2006, it remains almost similar until 2010 with only a little deviation. Figure 2 also shows that the quantitative similarity between the predicted and the observed images is higher than the locational similarity of the corresponding years. It can be seen that the overall Kappa coefficient value decreases at a constant rate as we go more in the distant future, thus it can be claimed that the forecasted images till 10 years from the predicted start date are still within the tolerable levels of accuracy. Beyond 10 years, the future prediction becomes more uncertain.

In the second test, we can see that with the change in prediction start date, the accuracy of the 2010 forecast map decreases. So the prediction start date has significant impact of the accuracy of the predicted images.

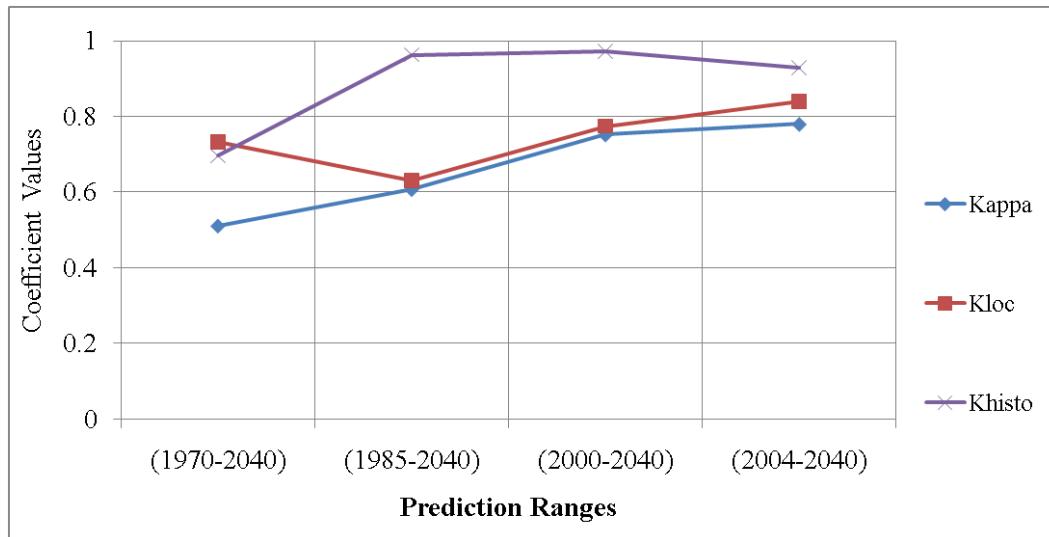


Figure 3: Trend of accuracy of the 2010 image when the prediction start dates are altered

A comparative analysis of the 2010 forecasted map across all prediction date range show that the ones starting from 1970, 1985 and 2000 overestimate the percentage of urban pixels in respective 2010 images whereas the 2004 prediction start date underestimates the percentage of urban pixels.

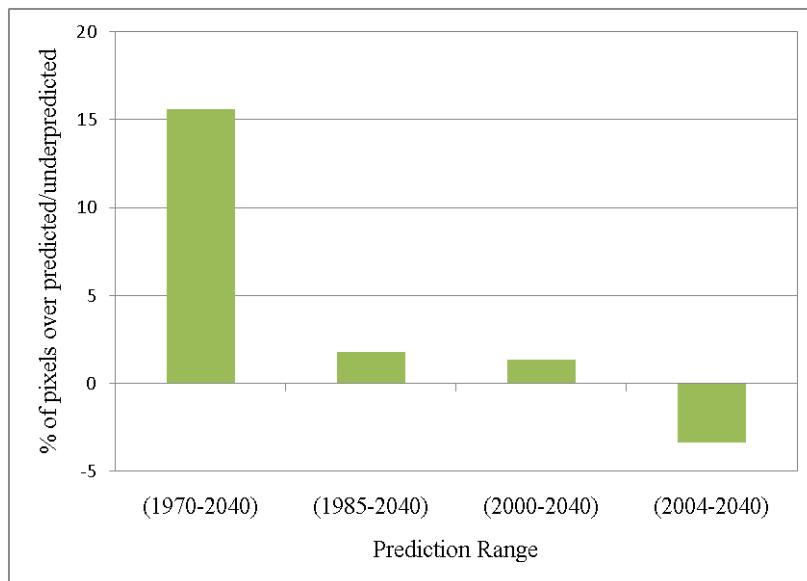


Figure 4: Overestimation/Underestimation of the percentage of urban pixels

At present, each model run is completed in a Monte Carlo fashion (http://www.ncgia.ucsb.edu/projects/gig/v2/About/data_files/scenario_file.htm). For calibration, measurements of simulated data are taken for years of known data, and are averaged over the

number of Monte Carlo iterations. The averaged values are compared to the known data, and multiple coefficient measures are calculated, the product of which gives us the Optimum SLEUTH Metric (OSM) (Dietzel and Clarke, 2005). The set of coefficient values having the highest OSM from the final calibration are considered as the ‘best-fit values’ and were used to initiate each simulation in a prediction run along with the SLEUTH images and a random seed value. After a simulation is complete, the initializing seed that began that simulation was reset and a new simulation was run (<http://www.ncgia.ucsb.edu/projects/gig/v2/About/bkStrPrediction.htm>). Thus in the first three cases, when the prediction start dates are in the distant past, with the average rate of growth of 20 years, the model overestimated urbanization for 2010 whereas when the prediction is from the ‘present year’ the model underestimated urbanization in 2010. This shows that in reality that region experienced significant fluctuations in the rate of urbanization and at present the growth rate is higher than the average.

The SLEUTH land use change model uses structure and form of the land use classes to capture the dynamics of the land use system. Goldstein et al, (2004) estimated that SLEUTH can be run for forecasts of urban extent for a time period as long as the historical data, in their case 70 years.

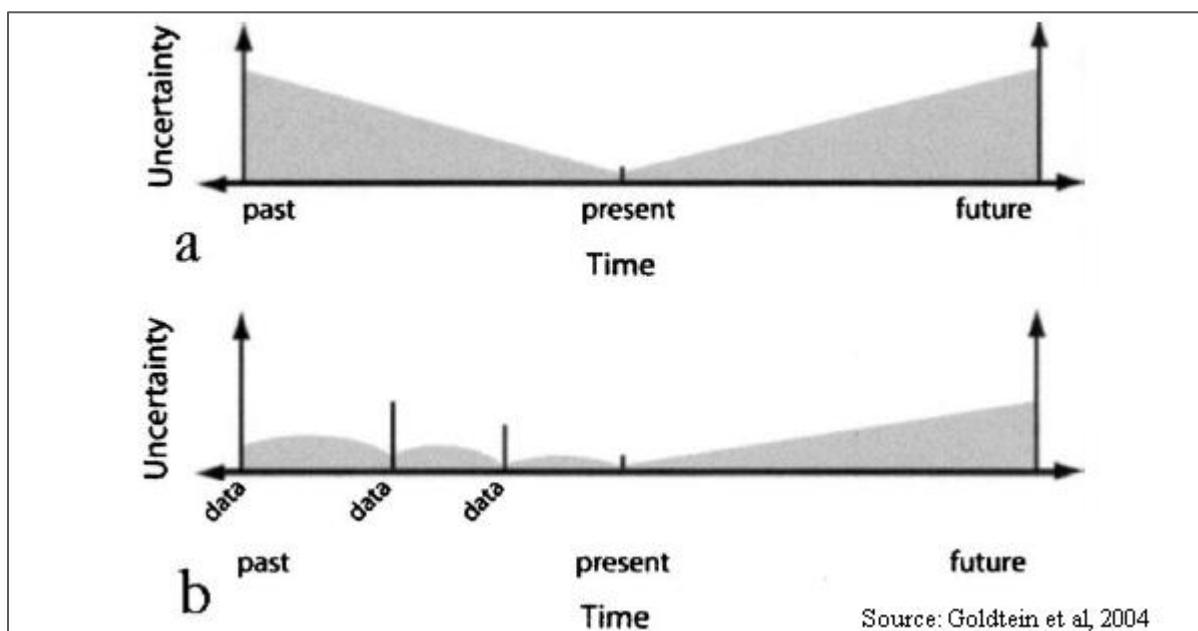


Figure 5: Uncertainty in urban models over time. In (a), only data for the present are included. For (b), three historical data sets are included in the modeling.

On the other hand, Candau (2000) showed that the accuracy of output data will increase with maximum amount of input data from immediate past. For long term prediction, these results can only be true if the process creating the pattern remains constant throughout the prediction range, which is unlikely. The present study shows that to maintain a reasonable level of accuracy, it is best to forecast the near future than the distant one, which resonates in time the first law of Geography by Tobler - '*Everything is related to everything else, but near things are more related to each other*'.

5. Acknowledgment

This study is supported by the UCTC Dissertation Research grant (2009)

6. References

- Candau J. T., 2002, Temporal Calibration Sensitivity of the SLEUTH Urban Growth Model, Master of Arts Thesis, University of California, Santa Barbara.
- Clarke, K. C., 2004, The limits of simplicity: Toward geocomputational honesty in urban modeling, Geodynamics, Atkinson, P., Foody, G., Darby, S., and Wu, F. (Eds) Florida, CRC Press
- Clarke, K. C., Gaydos, L., 1998. Loose-coupling a cellular automaton model and GIS:long-term urban growth prediction for San Francisco and Washington/Baltimore. International Journal of Geographical Information Science, 12, 699 – 714.
- Clarke, K. C, Gazulis, N., Dietzel, C. K., Goldstein, N. C., 2007. A decade of SLEUTHing: Lessons learned from applications of a cellular automaton land use change model. In: Fisher, P. (Ed.), Classics from IJGIS. Twenty Years of the International Journal of Geographical Information Systems and Science. Taylor and Francis, CRC. Boca Raton, FL, pp. 413-425.
- Goldstein, N. C. and Clarke, K. C., 2004, Approaches to simulating the ‘March of Bricks and Mortar’, Computers, Environment and Urban Systems, 28:125 – 147
- Hagen, A., 2002a, Multi-method assessment of map similarity, In: Ruiz, M., Gould, M., Ramon, J. (Eds.), Proceedings of the Fifth AGILE Conference on Geographic Information Science, Palma, Spain, pp. 171-182.
- Hagen, A., and Uljee, I., 2003, MAP COMPARISON KIT, User manual, RIKS report. Available from: <http://www.rivm.nl/milieu/modellen> or <http://www.riks.nl/MCK/>.
- Hagen-Zanker, A. and Martens, P., 2008, Map Comparison Methods for Comprehensive Assessment of Geosimulation Models, Computational Science and Its Applications – ICCSA 2008, Lecture Notes in Computer Science, 5072: 194-209
- MCK Reader: Methods of the Map Comparison Kit, 2005, compiled by Research Institute for Knowledge Systems (RIKS)

Pontius Jr., R. G., 2000, Quantification error versus location error in comparison of categorical maps, *Photogrammetric Engineering and Remote Sensing*, 66(8): 1011-1016

Pontius Jr., R. G., and Schneider, L. C., 2001, Land-cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA, *Agriculture, Ecosystems and Environment*, 85(1-3): 239-248

Visser, H. and de Nijs, T., 2006, The Map Comparison Kit, *Environmental Modelling and Software*, 21: 346 – 358

Landis, J.R., Koch, G.C., 1977. The measurement of observer agreement for categorical data, *Biometrics*, 33: 159 – 174.

High-Resolution Estimation and Analysis of Transport Accessibility in the City

Itzhak Benenson¹, Amit Rosenthal¹, Karel Martens²

¹Department of Geography and Human Environment, Tel Aviv University,
Ramat Aviv, 69978 Tel Aviv, Israel, benny@post.tau.ac.il, amitros2@post.tau.ac.il

²Institute for Management Research, Radboud University Nijmegen,
PO Box 9108, 6500 HK, Nijmegen, The Netherlands, k.martens@fm.ru.nl

1. Introduction

Accessibility, understood as the ability of people to reach and participate in activities (Garb and Levine 2002), is increasingly identified as a key criterion to assess transport and land use policies (Bristow et al. 2009). Comparison of car and transit accessibility is considered more important than ever, from an environmental and social perspective (Benenson et al. 2009).

Most large-scale accessibility analyses in the literature measure accessibility at the level of neighborhoods or traffic zones and use rather rough estimates of travel time (e.g., Shen 1998; Blumenberg and Ong 2001; Hess 2005; Kawabata and Shen 2006; Kawabata 2009). This may be sufficient for car accessibility, but for transit accessibility an accurate assessment of travel time requires geo-information at the resolution of buildings and road segments in order to accurately incorporate access, egress and waiting times in the measurement of total travel time.

2. Urban.Access2, the tool for high-resolution analysis of transport accessibility

We present Urban.Access2 – the novel and practically applicable tool for a high-resolution comparative analysis of accessibility at a large scale. Urban.Access2 is a GIS-SQL-server application for estimating car-based and transit-based accessibility to employment and other land uses. It combines GIS-abilities of the initial version of Urban.Access (Benenson et al, 2010) with the novel ability of estimating accessibility at resolution of separate building for large urban areas with the population of several millions and thousands public transportation lines of different kinds. We are not aware about similar tools for accessibility analysis.

Urban.Access2 enables a detailed representation of the traveler origin and destination at resolution of separate building and calculations of the travel times by transit and car and thus makes it possible to adequately compare accessibility levels by transport modes. The critical advantage of the approach and the software is its high performance that is based on representation of public transportation lines and travels as tables of relationships and constructing accessibility maps with the help of the optimized SQL transactions with the MS SQL 2008 server. The GIS component of the system is employed for one-time heavy spatial pre-processing at the initial stage of analysis and for the presentation and analysis of the resulting accessibility maps (Figure 1).

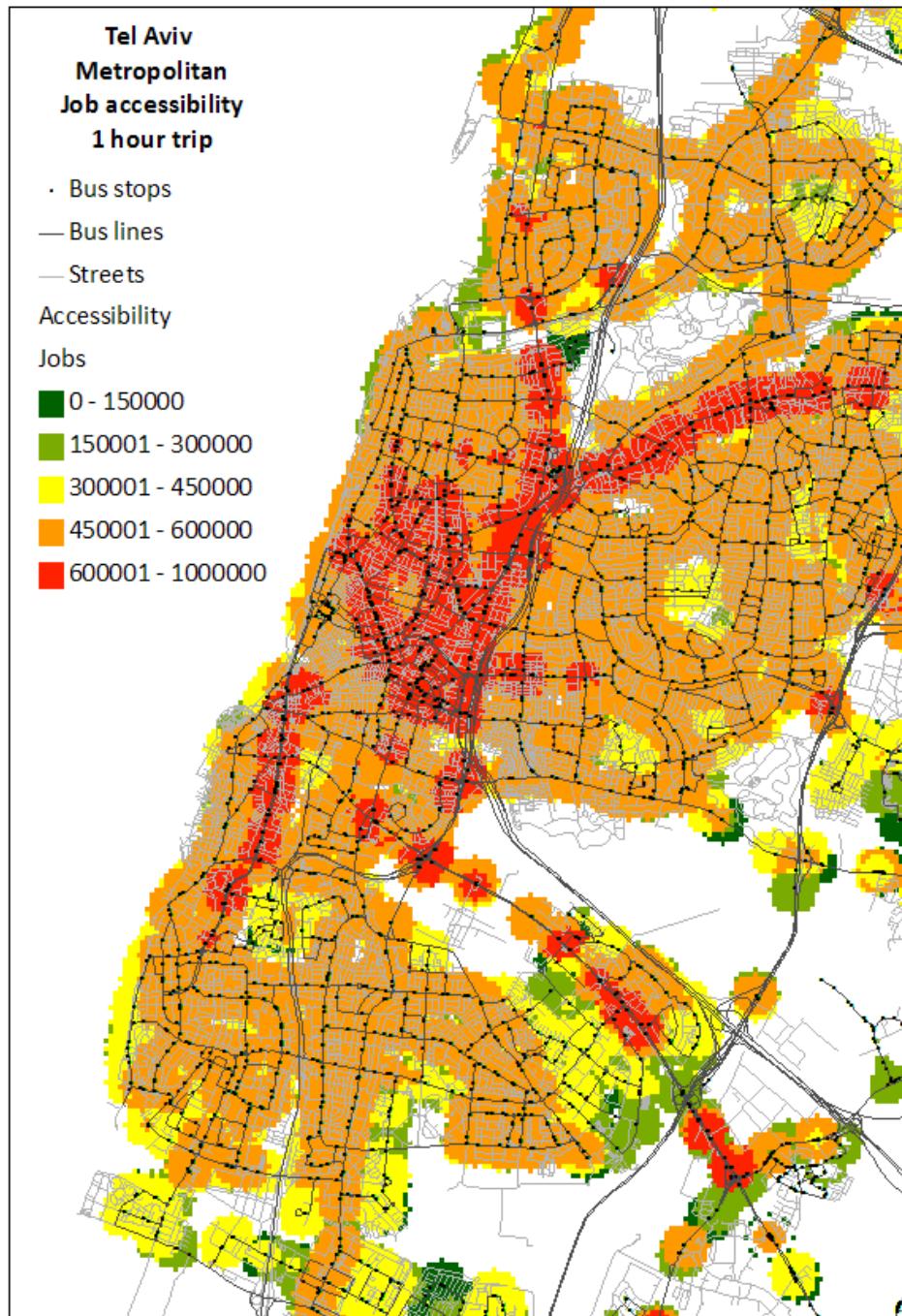


Figure 1. The example of Urban.Access2 output – transit accessibility of jobs

3. Application of Urban.Access2 for analysis of the public transportation system in Tel Aviv metropolitan area

The application of Urban.Access2 to the public transportation system in Tel Aviv metropolitan area, with its population of 3 million, and 100,000 origins and destinations, 300 bus lines and 3500 bus stops reveals large gaps between car-based and transit-based

accessibility. At average, transit accessibility of the Tel-Aviv metropolitan urban area is 7 – 10 times lower and of the jobs there 3 – 4 times lower than accessibility with the car.

The results found for the Tel Aviv metropolitan area show large gaps between car-based and transit-based accessibility, which are similar to those found in USA, which is notorious for its poor transit network. For instance, Hess (2005) finds a car/transit job accessibility ratio of 1.7 to 8.2. Based on the Urban.Access2 application, we find even lower values for many areas than the lowest value observed by Hess (2005). We argue that this is not the result of a poorer transit system in the Tel Aviv area, but rather of a more detailed description of travel by transit in the Urban.Access2 application. Given the counter-intuitive finding that the results for Tel Aviv, with its relatively dense bus network, are largely comparable to those found in US cities, we conclude that a more detailed representation of travel by transit results in larger accessibility gaps. These large gaps can be ascribed to the distinction between direct trips and trips with transfers, to a detailed analysis of transit travel time at the level of individual addresses, and to the inclusion of both the estimated waiting time at the outset of a trip, and the walking and waiting times related to (bus) transfers, in the accessibility index.

The larger gaps point to a greater need for adequate policy responses, both for reducing car dependence as well as for creating a more equitable transport system. In this milieu we analyze the new program of public transportation restructuring that has been started in Tel Aviv metropolitan in 2010 and estimate its advantages and limitations.

4. References

- Benenson, I., K. Martens, Y. Rofé, A. Kwartler, 2010, Public transport versus private car GIS-based estimation of accessibility applied to the Tel Aviv metropolitan area, DOI 10.1007/s00168-010-0392-6
- Benenson I, Martens K, Rofé Y and Kwartler A, 2009, Public transport versus private car: GIS-based estimation of accessibility applied to the Tel Aviv metropolitan area, *89th Annual Meeting of the Transportation Research Board*. Washington DC, USA.
- Blumenberg EA and Ong P, 2001, Cars, buses, and jobs: welfare participants and employment access in Los Angeles, *Transportation Research Record: Journal of the Transportation Research Board*, 1756: 22-31.
- Bristow G, Farrington J et al, 2009, Developing an evaluation framework for crosscutting policy goals: the Accessibility Policy Assessment Tool. *Environment and Planning A*, 41(1):48-62.
- Garb Y and Levine J, 2002 Congestion pricing's conditional promise: promotion of accessibility or mobility? *Transport Policy*, 9(3):179-188.
- Hess DB, 2005, Access to employment for adults in poverty in the Buffalo–Niagara region. *Urban Studies*, 42(7):1177-1200.
- Kawabata M, 2009, Spatiotemporal dimensions of modal accessibility disparity in Boston and San Francisco, *Environment and Planning A*, 41(1): 183-198.
- Kawabata M and Shen Q, 2006, Job accessibility as an indicator of auto-oriented urban structure: a comparison of Boston and Los Angeles with Tokyo, *Environment and Planning B: Planning and Design*, 33(1): 115-130.
- Shen Q, 1998, Location characteristics of inner-city neighborhoods and employment accessibility of low-wage workers, *Environment and Planning B: Planning and Design*, 25(3): 345-365.

Agent-based Modeling of Sustainable Urban Development along the Mid-Section of Silk Road in Northwest China

Yichun XIE and Xiaojin TAN

Department of Geography and Geology, Eastern Michigan University, Ypsilanti,
Michigan 48197, yxie@emich.edu

An accelerated urbanization has taken place in economically advanced regions of the east coast of China, particularly where the influence of market reforms and globalization have been most strongly felt. However, less is known of the urban development in West China along the ancient silk-road (so called West Yellow River Corridor - WYRC). The urban system there was originated with the thriving of silk trade and military installations, and was promoted as an important inland industrial and transportation base before China's economic reforms and has been reenergized in recent years as an important corridor to relay economic miracles from the east to the west. However, the urban growth has taken a unique path deviating from what is commonly seen in the coastal China. The urban growth has been driven by historical heritage and political favoritism, but constrained by water shortage and harsh natural environment because this region is geographically located in the arid and semi-arid hinterlands of the Eurasia Continent. The trajectory of urban growth in this region reflects dynamic complexity of coupled human and natural systems.

This paper attempts to construct a multi-agent-based model (MABM) to explore the emergence of urban system in recent three decades in WYRC and the associated environmental constraints and socioeconomic forces underlying the urbanization process. The WYRC-MABM is structured to consist of three groups of agents, water agents (WAs), developer agents (DAs), and policy agents (PAs). WAs account for available water resource and are quantified as the current water quotas through the field survey, which is often adopted in regional studies of resource management. DAs allocate urban growth on the basis of land availability and cost and population growth. PAs study sustainable urban development, promote rational water utilization, and predict economic and demographic growths. The complex interactions of WAs, DAs, and PAs are maximized on the basis of the economic base theory in simulation. The total available water resources and the water quotas by major economic sectors at present were calculated from the field survey. The water quotas and their change rates by economic sectors over the period of 2000 – 2030 were estimated on the basis of policy study of regional development goals and future trends of water usages and changes. Finally the water quotas and their change rates were integrated with regional models of demographic and economic predictions to compute sustainable economic development by economic sectors, and then to derive total population and urban population. Therefore, a sustainable urban growth under the limitation of insufficient water supplies in He-Xi was determined.

The simulation results over the period of 2000 – 2030 reveal that a sustainable urban growth could be achieved under the limitation of insufficient water supplies only if the urbanization level would be allowed to rise to 35.14% by 2030. This figure is just at the average urbanization level of China in 2001. Thus it is a challenging task for China's governments to control urban growth in WYRC.

Key words: China, agent-based modeling, arid, cellular automata, Eurasia, regional development, semi-arid, sustainability, urbanization, water

Using Agent-Based Complex Systems to model impacts of policy decisions on Climate Change scenarios

Marta Vallejo

School of Mathematical & Computer Science.

Heriot-Watt University.

EH14 4AS Edinburgh

Telephone: +44 (0) 131 451 4393

mv59@hw.ac.uk

March 29, 2011

1 Introduction

The present abstract proposes the use of Agent-Based Complex Systems (ABCS) to model the impact of policy decisions on Climate Change. ABCS is a powerful and innovative technique of representation capable of modelling and analysing the behaviour of the key actors in climate policy as well as the consequences of their decisions.

1.1 Background

Climate Change poses a problem that should be tackled taking into consideration the interaction between science and policy. Meanwhile scientists' expectations about the emergence of the problem [7], the scope, the duration and its plausible consequences are not completely defined. This is because the processes involved in climate are not totally understood and the level of uncertainty is remarkable. Based on these premises policy-makers try to design a series of measures to mitigate causes and adapt societies to these possible scenarios. However the mechanism, the speed and the priority in which this political and social process should be developed is, in fact, an unknown factor.

Globalisation is a political convergence process where the power of decision of each state is gradually less significant. This lack of autonomy determined by structural constraints entails a general tendency where most policies converge to one global policy. This is a crucial fact to understand the driving forces that play a role in the creation of the policies.

The formulation of regulations can be based on two main goals:

- **Social or ideational:** responding to preferences like: social benefits, voters, interest groups...

- **Economic:** based on economic models like DICE [6] in which the problem is defined as a cost-benefit trade-off and is reduced to a monetary problem. Currently there exists a clear dominance of economic versus ideational points of view. However, this approach by itself can be considered flawed for analysing the problem of Climate Change [5]

The study of how these two goals can be combined and what consequences result from this selection can shed light on how policies should be developed in the future and hence identify the effects for climate.

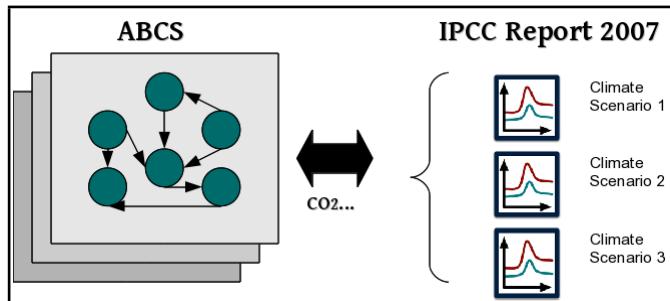
Individually climate and how policies are designed, involve an enormous level of complexity. If an observer tries to contemplate and understand how both systems are influenced by each other it is highly important to analyse the role of human beings. In order to understand the development of legislations and agreements and in which way they influence and are influenced by Climate Change, I propose the creation of a model coupling social and environmental areas.

Agent-based models (ABM) and Cellular Automata (CA) have been used to understand the interconnections, interdependences and feedbacks among a set of individual entities. The capability of representing human decision-making to achieve a determined goal and the emergence of new properties in a spatial and temporal changing environment, make it a good model for our purpose. ABM has been used in multiple disciplines, concretely in ecology [1, 2] and social sciences [3] and the application of these techniques are increasingly more common.

Nevertheless the underlying characteristics of a climate and policy decision-making system requires dealing with a large quantity of different patterns and behaviours. Cellular Automata has the drawback that it can deal with only one pattern. The consequence of this is that an individual pattern is not rich enough to describe the behaviour of a complex system. It is therefore compulsory to search for a way of representing a higher level of complexity, taking into account the problem that is: the more representation capability, the more complexity in structure and computation.

2 The Proposed Model

I propose to increase the representational power of ABM-CA using Agent-Based Complex Systems (ABCS) & Pattern-Oriented Modelling POM [4]. Both techniques constitute a more powerful and comprehensive point of departure. Taking observed patterns in nature as a starting point[8], the created model should be capable of mimicking the internal organisation of the system.



Instead of representing all the knowledge in a unique physical layer the system will be made up of a set of layers. Each layer represents a different area of interest and can interact with other layers in order to reproduce complex phenomena. These layers could cover:

- Geography: the physical division of countries with characteristics like population, level of development...
- Economy: The boundaries among countries are more and more fuzzy due to globalisation. Countries tend to group and behave according to a common set of rules given by treatments, agreements.
- Each region can have different areas of influence according to possible factors like language, history, political regime, religion...
- Ecology: climate, biodiversity, ecology concern....

Each heterogeneous agent has associated a set of characteristics according to the different layers where it is located. Each level can show independent or grouped behaviour that can be conditioned by the characteristics of other layers. Its goal can be subdivided into different parts according to these layers and could be modified by other agents according to its level of autonomous behaviour. In turn, its decisions can have different levels of repercussion in other agents with respect to its area of influence.

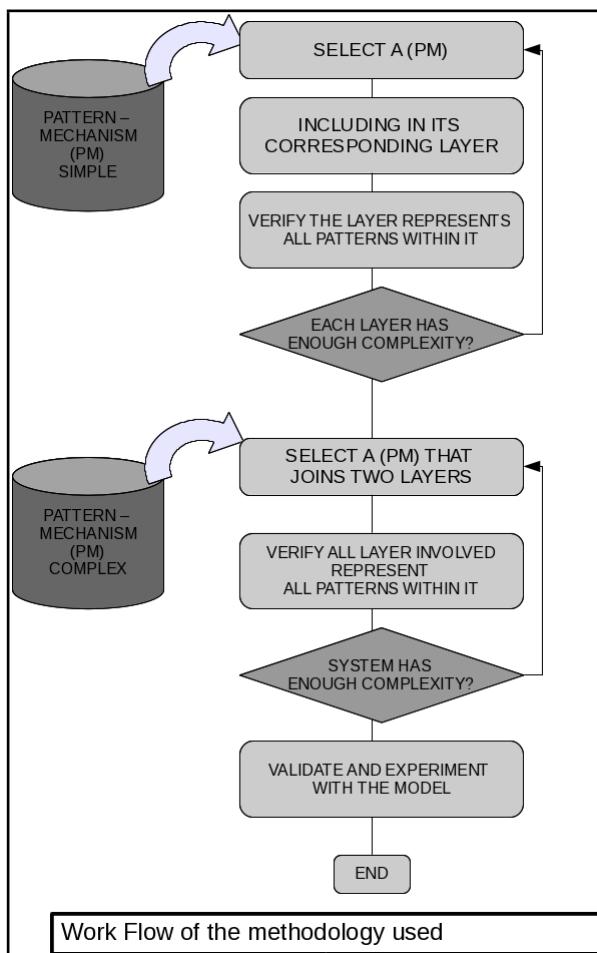
2.1 Potential Benefits

With the outcomes gathered from the agents in form of CO_2 emissions, the system will be able to evaluate future climatic responses studying the scenarios constructed by the IPCC [7]. Scenarios are defined by a range of greenhouse gas thresholds in the atmosphere. If our model concludes that in 10 years the increment of CO_2 will give rise to a 1.8 degree growth in temperature, then we are in the scenario called B1 and the IPCC can give us information about what are the most plausible consequences for the environment. With these data the system can correct its behaviour relaxing or strengthening the regulations in order to control the future effects of Climate Change. With this exchange of information the system will trace the steps for possible future scenarios. Exploring different suitable paths can help us to understand the essence of the interconnection between these two systems.

2.2 Methodology

Reality cannot be totally formalised. The key goal of modelling is to find the appropriate level of aggregation that allows us to synthesize an effective solution. Each additional degree of freedom represents a remarkable increment in the efforts to structure and understand the model.

The designed architecture allows us to start from a coarse model that represents a simple pattern observed in nature and studies possible mechanisms that give rise to these patterns. Each layer will be treated independently in a preliminary phase as an independent component of the model. By an incremental approach the model is refined and rechecked for each pattern included. When a layer's definition process finishes, an interconnection phase will be carried out by the inclusion in the model of patterns that interconnect more than one layer. Deciding the layers involved we can expand the model until achieving the complexity desired.



Attributes determine the decision-making behaviour of agents along with their strategies and goals. The application of reinforcement learning or genetic algorithms techniques will be used to measure closeness between actions and

goals and minimise this distance.

2.3 Testing

This phase will be accomplished from two points of view:

- Sensitivity and error analysis. This will be carried out by the analysis of simple and hierarchical relationships between input and output parameters of the model.
- Exploring the model structure by performing a set of comparative experiments about realistic and unrealistic scenarios from the past, focusing in the way that the outcomes are determined by the assumptions considered and how the model is simplified and refined.

3 Acknowledgements

This project is funded by Heriot-Watt's Environment & Climate Change Theme.

References

- [1] F Bousquet and C Le Page. Multi-agent simulations and ecosystem management: a review. *Ecological Modelling*, 176(3-4):313 – 332, 2004.
- [2] Donald L. DeAngelis and Wolf M. Mooij. Individual-based modeling of ecological and evolutionary processes1. *Annual Review of Ecology, Evolution, and Systematics*, 36(1):147–168, 2005.
- [3] Nigel Gilbert and Klaus G Troitzsch. *Simulation for the Social Scientist*. Open University Press, 2005.
- [4] Volker Grimm, Karin Frank, Florian Jeltsch, Roland Brandl, Janusz Uchmanski, and Christian Wissel. Pattern-oriented modelling in population ecology. *Science of The Total Environment*, 183(1-2):151 – 166, 1996. Modelling in Environmental Studies.
- [5] Scott Moss, Claudia Pahl-Wostl, and Thomas Downing. Agent-based integrated assessment modelling: the example of climate change. *Integrated Assessment*, 2:17–30, 2001. 10.1023/A:1011527523183.
- [6] William D. Nordhaus. *Managing the Global Commons. The Economics of Climate Change*. The MIT Press (Cambridge, Mass.), October 1994.
- [7] S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, KB. Averyt, M. Tignor, and HL. Miller. Ipcc (2007) summary for policymakers. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, 2007. Cambridge University Press.
- [8] Jacob Weiner. On the practice of ecology. *Journal of Ecology*, 83(1):153–158, February 1995.

An Agent-Based Model of Woodland Caribou Habitat-Selection in West Central Alberta: A Behavioural and Ecological Approach

C. A. D. Semeniuk¹, M. Musiani², M. Hebblewhite³, S. Grindal⁴, D. J. Marceau⁵

¹Department of Geomatics Engineering, University of Calgary, 2500 University Dr., Calgary, Alberta, Canada, T2N 1N4
Telephone: 1-403-220-8794
Fax: 1-403-284-1980
Email: semeniuc@ucalgary.ca

²Faculties of Environmental Design and Veterinary Medicine, University of Calgary
Telephone: 1-403-220-2604
Fax: 1-403-284-4399
Email: mmusiani@ucalgary.ca

³College of Forestry and Conservation, University of Montana, Montana, US, 59812
Telephone: 1-406-243-6675
Fax: 1-406-243-4557
Email: mark.hebblewhite@cfc.umt.edu

⁴ConocoPhillips Canada, Alberta Canada, T2P 2H7
Telephone: 1-403-233-3261
Email: scott.d.grindal@conocophillips.com

⁵Department of Geomatics Engineering, University of Calgary
Telephone: 1-403-220-5314
Fax: 1-403-284-1980
Email: dmarceau@ucalgary.ca

1. Introduction

Alberta woodland caribou (*Rangifer tarandus*) are classified as threatened in Canada, and a local population in the province's Foothills Region, the Little Smoky herd, is at immediate risk of extirpation due, in part, to anthropogenic activities such as oil, gas, and forestry that have altered the ecosystem dynamics. The Alberta government resultantly recommends the assessment and management of cumulative effects on caribou, as well as the identification and provision of adequate habitat (amount and type), to allow for caribou persistence (ASRD 2010). While much is known about caribou ecology, the behavioural mechanisms by which resource-extraction industries contribute to caribou population decline are less clear. Traditional approaches to studying wildlife-human-environment interactions do not typically consider individual-level information, account for complexities, or integrate cross-scale and cross-discipline data and methods, resulting in a great loss in predictive or explanatory power (An et al. 2005). To address these issues, we have developed a spatially explicit, agent-based model (ABM) to simulate winter habitat selection and use of woodland caribou in the face of intense land use by resource-extraction industries in west-central Alberta.

1.1 Theoretical background

The knowledge of species distribution is a vital component in wildlife conservation and management. Such information aids in quantifying animal–habitat relationships, describing and predicting differential space use by animals, and ultimately identifying habitat that is important to an animal (Beyer et al. 2010). In determining critical habitat for population persistence it is crucial to understand the mechanisms that drive animal movement and habitat selection as it is the adaptive responses of individuals to environmental conditions that give rise to population- and community-level phenomena from which researchers can more effectively identify habitat quality. ABMs provide an ideal tool for this important characteristic of ecological systems because they are designed to explore the mutual relationship between the adaptive behavior of individuals and the system-level properties (Grimm et al. 2007).

As such, ABMs can readily incorporate two critical ecological theories involved in habitat selection: animal movement ecology and behavioural ecology. The movement paths of wildlife result from the dynamic interplay of the internal state of the organism, its motion capacity, its navigation capacity, and the external environment (Kolyoak et al. 2008, Revilla and Wiegand 2008). Since agents are also given fitness-maximizing goals and can trade off competing strategies to find optimal solutions to the problems they face (i.e., behavioral ecological theory), this enables the understanding of the processes that govern movement, distribution, and selection, and therefore predict how animals might respond to habitat loss and other environmental change (Grimm et al. 2007).

The theoretical foundation of our ABM is that caribou agents are goal driven and will make realistic, optimizing tradeoffs between factors constraining fitness: energy reserves, resource distribution and abundance, energetic cost of movement, and predation risk and disturbance. The decision with the highest payoff will drive animal movement and ultimately determine habitat selection. By bestowing ecological processes and evolutionary principles on caribou agents, we can identify the selection of resources that are truly related to animal fitness.

2. Methodology

2.1 Woodland caribou system

We have chosen to simulate our caribou ABM during the winter, as over-wintering caribou face the energetic costs of food availability, environmental conditions, predator avoidance, and disturbance. Specifically, the availability of terrestrial lichen, the main food source for Alberta woodland caribou in winter, is constrained by specific habitat requirements (Dzus 2001); minimizing costs in winter appears important for female caribou, at times at the expense of increased predation risk (Johnson et al. 2002); and caribou reduce predation pressure by increasing the distance between conspecifics, other ungulates, and wolves, or by avoiding the habitat in which they are found (DeCesare et al. 2010). Finally, woodland caribou may also associate industrial features with high predation risk, or equally, perceive industrial activities as being of high predation risk in their own right, thus experiencing energetic costs in industrial-feature avoidance (Bradshaw et al. 1998).

2.2 Model calibration and evaluation

The main data source used for the model calibration is a database composed of radio-collared GPS location data of caribou (Hebblewhite and Musiani 2010). A total of 5225 location points were obtained for 13 female individuals over the course of winter (November-April) 2004-2005. Using the GPS point samples, the spatiotemporal trajectory of each caribou was built and stored within an ArcGIS database, from which patterns were extracted and used to calibrate the movement behaviours of caribou: turn angles and daily step distributions. Other sources of biological information necessary for the caribou ABM parameterization include caribou agents' internal state variables, bioenergetic functions, spatial memory, and learned decision-making processes. The values for these variables were obtained from an extensive literature review, and include other published agent-based models of ungulate movement and migration (Bennet and Tang 2006, Metsaranta 2008). Data used to spatially represent the environment comprise remote sensing and GIS datasets of the Little Smoky region. These data have also been collated in an ArcGIS database and include vegetation and land cover maps, elevation data, road networks, seismic lines, and locations of wells, circa 2005 (McDermid et al. 2009, DeCesare et al. 2010). The area of interest covers 2400 km² and represents the official political and biological range delineation of the Little Smoky herd by the Alberta Fish and Wildlife Division (ASRD 2010).

The ABM model is implemented in the agent-based modelling platform NetLogo (Wilensky 1999), and verified for proper programming functioning through progressive debugging and uncertainty testing. The simulated behaviours and habitat selection of the agent caribou are to be evaluated by comparing these data with actual caribou patterns extracted from the empirical data and literature sources that are not used in the calibration process. This approach uses multiple patterns, each describing a certain characteristic aspect of the real system, to verify parameterization and guide model selection (Wiegand et al. 2003). Patterns used to evaluate the caribou ABM comprise caribou net displacement, fixed kernel density estimators, use of landscape cover, and distance to industrial features.

2.3 Conceptualization and implementation of the ABM

The ABM model is composed of the following elements: (1) cognitive caribou agents, (2) decision-making heuristics that act to optimize the agent's self interest (e.g., reproduction, bioenergetics), (3) learning rules and adaptive processes (e.g., memory of high quality and low risk sites), and (4) a non-agent environment (e.g., the spatially-explicit GIS-based caribou range).

Three raster maps of 45 m resolution are used to represent the physical environment where the agents are located and interact with each other: a land-use map representing vegetation and land cover, a digital elevation map, and an industrial-features map. This resolution is chosen to represent the size of the foraging patch of caribou (Bailey and Provenza 2008). A virtual grid is overlaid on these three maps to provide an environment to the agents and allow their movement from one cell to the next cell (one of its eight neighbours). The environment is assigned three main characteristics: forage availability, energetic travel cost, and predation risk/disturbance.

The caribou are represented as cognitive agents: they have a mental representation of their environment; they can plan their activities, and have a memory of profitable

patches. The goal of the caribou is to maximize its fitness throughout the winter (six months), which is translated in the model as energy reserves. Each time step in the model represents one half hour. A decision model encompassing the behavioural rules of the caribou agents has been built (Figure 1) wherein the animal must assess its current energetic state and make a decision about its next action dependent on whether: 1) it has met its daily energetic requirements, 2) its current actions will result in enough accumulated energy to birth viable offspring, and 3) the risk of predation in its immediate surroundings is tolerable. The assortment of actions the caribou can choose range from resting to foraging, moving locally to a safer location, moving locally to forage, and taxiing to a new location or one from memory, either at low energetic cost or at low risk of predation.

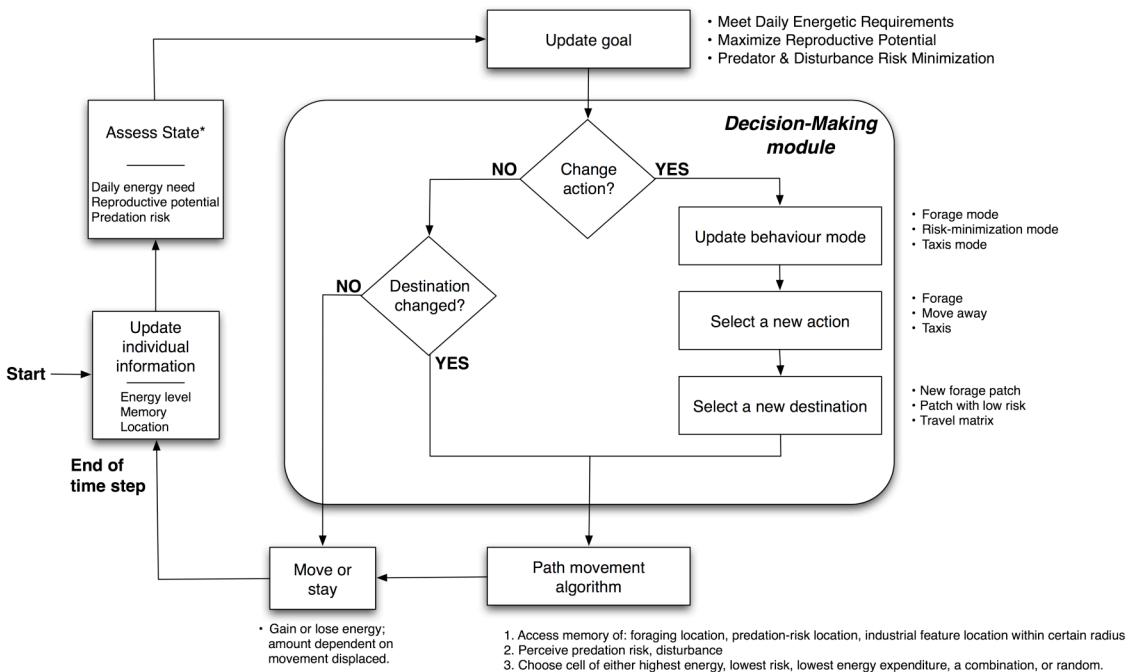


Figure 1. Conceptual behavioural decision model of caribou agents

2.4 Model simulations

To determine the most probable mechanism of habitat selection, we shall be simulating alternative caribou fitness-maximizing goals, for instance, whether caribou in winter attempt to meet their daily energetic requirements, maximize their reproductive potential, avoid predation, or a combination. These strategies are governed by caribou threshold tolerances for energetic reserves and willingness to accept risk/disturbance (Figure 2). The underlying mechanism is to be confirmed by verifying which resultant simulations of movement patterns most closely match actual caribou distributions and other extracted patterns. The results we shall be presenting will elucidate the most common strategy caribou use to select their habitat, thus offering insight into *why* caribou are choosing the habitats they use, and consequently, the habitat types that are most important in maximizing caribou bioenergetics and fitness.

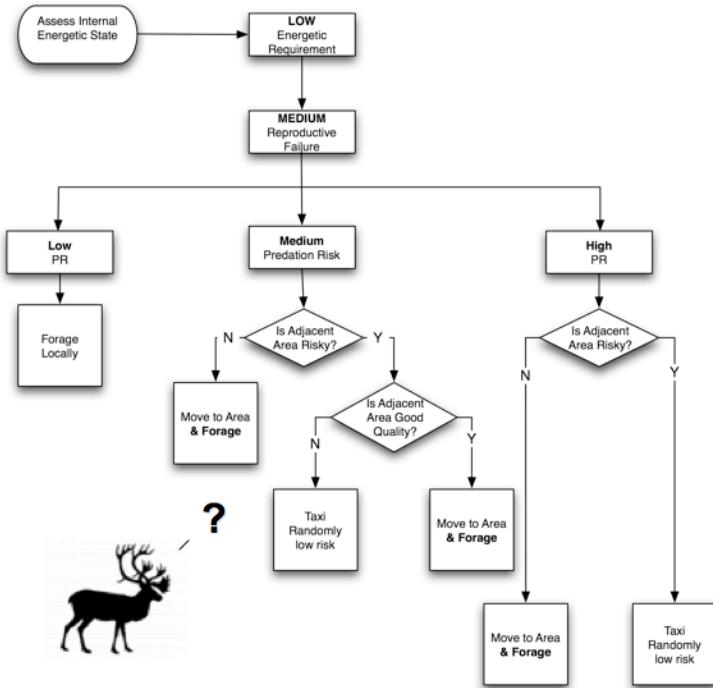


Figure 2. Example of behavioural trade-off flow chart of caribou agents

3. Conclusion

By capitalizing on the utility of ABMs to accommodate behavioural and ecological theory, we aim to show that carefully designed mechanistic models can be used to understand and to predict the consequences of individual behavioral responses to environmental conditions for population-level phenomena such as habitat selection and use. Our model findings will have benefits for conservation and industry-management purposes, serving as an applied, science-based decision tool for managing potential effects of resource extraction activities on valued resources.

4. Acknowledgements

This project is funded by the MITACS Accelerate Program in collaboration with ConocoPhillips Canada and two University Technologies International Scholarships awarded to C. Semeniuk. Support is also provided by the Schulich Research Chair awarded to D. Marceau. We would like to thank Greg McDermid and Nick DeCesare for their invaluable assistance in providing data to the project.

5. References

- Alberta Sustainable Resource Development and Alberta Conservation Association (ASRD), 2010, Status of the Woodland Caribou (*Rangifer tarandus caribou*) in Alberta: Update 2010. Alberta Sustainable Resource Development. Wildlife Status Report No. 30 (Update 2010). Edmonton, AB. 88 pp.
- An L, Linderman M, Qi J, Shortridge A, and Liu J, 2005, Exploring complexity in a human-environment system: an agent-based spatial model for multidisciplinary and multiscale integration. *Annals of the Association of American Geographers*, 95:54–79.
- Bailey DW and Provenza FD, 2008, Mechanisms determining large herbivore distribution. In F. van Langevelde and H.T.T. Prins (ed.) *Resource Ecology Spatial and Temporal Dynamics of Foraging*. Wageningen University Resource Ecology Group & Frontis, Wageningen, the Netherlands. pp.7-28.
- Bennett DA, and Tang W, 2006, Modeling Yellowstone's northern range elk herd as adaptive, spatially aware, and mobile agents. *International Journal of Geographical Information Science*, 20:1039–1066.
- Beyer HL, Haydon DT, Morales JM, Frair JL, Hebblewhite M, Mitchell M, and Matthiopoulos J, 2010, The interpretation of habitat preference metrics under use-availability designs. *Philosophical Transactions of the Royal Society B*, 365: 2245-2254.
- Bradshaw CJA, Boutin S, Hebert DM, 1998, Energetic implications of disturbance caused by petroleum exploration to woodland caribou. *Canadian Journal of Zoology*, 76:1319-1324.
- DeCesare NJ, Hebblewhite M, Robinson HS, and Musiani M, 2010, Endangered, apparently: the role of apparent competition in endangered species competition. *Animal Conservation*, 13:353-362.
- DeCesare NJ, Peters W, Robinson HS, Weckworth B, Semeniuk CAD, Musiani M, McDermid G, Hebblewhite H, 2010, Summary of GIS layers available for spatial analyses by the Canadian Rockies Caribou Project. Geographic Information System (GIS) Work Plan, September 2010.
- Dzus E, 2001, Status of the Woodland Caribou (*Rangifer tarandus caribou*) in Alberta. Alberta Environment, Fisheries and Wildlife Management Division, and Alberta Conservation Association, Wildlife Status Report No. 30, Edmonton, AB. 47 pp.
- Grimm V, Stillman R, Jax K, Goss-Custard J, 2007, Modeling adaptive behavior in event-driven environments: temporally explicit Individual-based Ecology. In: Bissonette J, Storch I (eds). *Temporal Dimensions of Wildlife Ecology: Wildlife Responses to Variable Resources*. Springer, pp. 59-73.
- Hebblewhite M and Musiani M, 2010, Linear features, forestry and wolf predation of caribou and other prey in west central Alberta. Petroleum Technology Alliance Canada (PTAC) Caribou Research Final Report.
- Holyoak M, Casagrandi R, Nathan R, Revilla E, Spiegel O, 2008, Trends and missing parts in the study of movement ecology. *Proceedings of the National Academy of Sciences*, 105, 19060-19065.
- Johnson CJ, Parker KL, Heard DC, Gillingham MP, 2002, A multiscale behavioral approach to understanding the movements of woodland caribou. *Ecological Applications*, 12:1840-1860.
- McDermid GJ, Hall RJ, Sanchez-Azofeifa GA, Franklin SE, Stenhouse GB, Kobliuk T, and LeDrew EF, 2009, Remote sensing and forest inventory for wildlife habitat assessment. *Forest Ecology and Management* 257:2262-2269.
- Metsaranta JM, 2008, Assessing factors influencing the space use of a woodland caribou *Rangifer tarandus* caribou population using an individual-based model. *Wildlife Biology*, 14: 478-488.
- Revilla E and Wiegand T, 2008, Individual movement behavior, matrix heterogeneity, and the dynamics of spatially-structured populations, *Proceedings of the National Academy of Sciences*, 105:19120-19125.
- West Central Caribou Landscape Planning Team (WCCLPT), 2008, West Central Caribou Landscape Plan. Submitted to the Alberta Caribou Committee.
- Wiegand T, Jeltsch F, Hanski I, and Grimm V, 2003, Using pattern-oriented modeling for revealing hidden information: a key for reconciling ecological theory and application. *Oikos*, 100: 209–222.
- Wilensky U, 1999, NetLogo. <http://ccl.northwestern.edu/netlogo/>. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Agent Based Modelling and GIS for Community Resource Management: Acequia-based Agriculture

S. Wise¹, A. T. Crooks²

¹George Mason University, Room 379, Research 1 Building, MS 6B2, Fairfax, VA 22030, USA
 Telephone: (001) 703 993 4640
 Fax: (011) 703 993 9290
 Email: swise5@gmu.edu

²George Mason University, Room 379, Research 1 Building, MS 6B2, Fairfax, VA 22030, USA
 Telephone: (001) 703 993 4640
 Fax: (011) 703 993 9290
 Email: Andrew Crooks acrooks2@gmu.edu

1. Introduction

In northern New Mexico, water is a scarce and precious commodity. A traditional local system of water management involves landowners collectively maintaining and managing ditches which distribute water among the properties. This system of physical ditches and organization together are known as an acequia, and the landowners who maintain it are called parciante. The water carried by the ditches is a shared resource, and the complex management system of the acequia has evolved to avoid Hardin's Tragedy of the Commons with regard to natural resources (Hardin 1968).

Despite the historical strengths of acequias, parciante are increasingly pressured to convert farmland into residential space. Any effort to protect this traditional form of agriculture relies an understanding of how the different parts of the system interact and how rigorous the system is to perturbation. This simulation seeks to model land use in acequia-dependent Taos, New Mexico. As an example of the area, fig. 1, displays a map of the tracts of land associated with this form of agriculture. The goal is to construct a tool that will allow a researcher or policy-maker to understand and interact with the acequia system in an intuitive fashion. Fig. 2, is an example of the simulation interface.

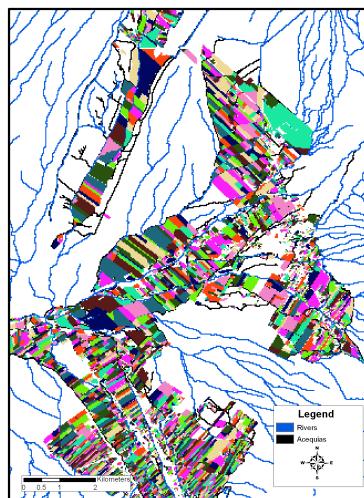


Figure 1. Tracts of land owned by parciante in Taos, New Mexico.

2. Methodology

The simulation is a spatially explicit agent-based model programmed in Java using the MASON Simulation Toolkit (Luke et al, 2005). It is made up of modules which capture the physical, economic, and social processes that impact land use patterns in the Taos area. The model includes a series of maps showing the spatial environment, graphs which track statistics about parciantes and urbanized versus agricultural land use, and an interface which allows to user to hide layers of information or modify the parameters of the environment mid-run.

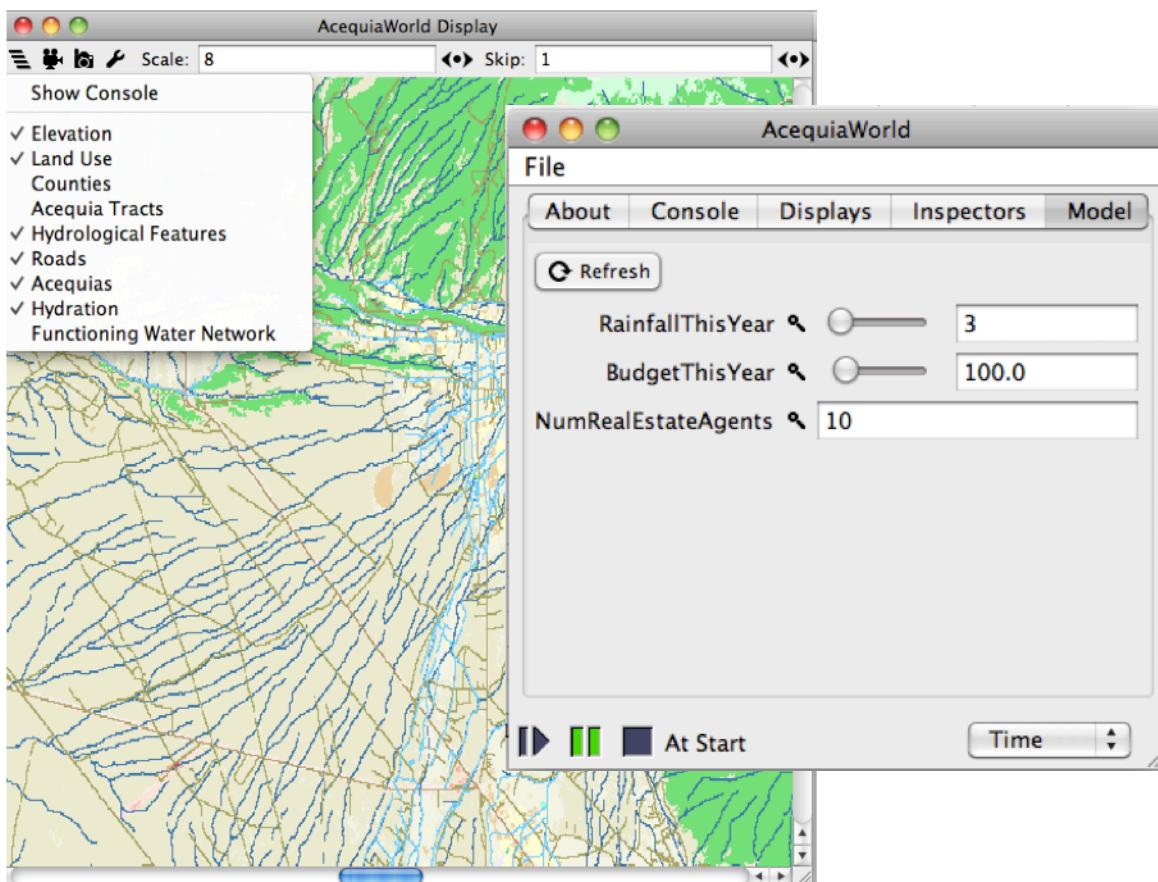


Figure 2. A sample run with user interface.

2.1 Data

Data utilized within the model as shown in fig. 3, comes from the work of Michael Cox (2010) and are supplemented with GIS information from the USGS's EarthExplorer. The area of study is the county of Taos, New Mexico and its surrounding area. The information was processed into $30m^2$ raster grid cells. The land use classification utilized here is described in Homer's work about the National Land Cover database (2007).

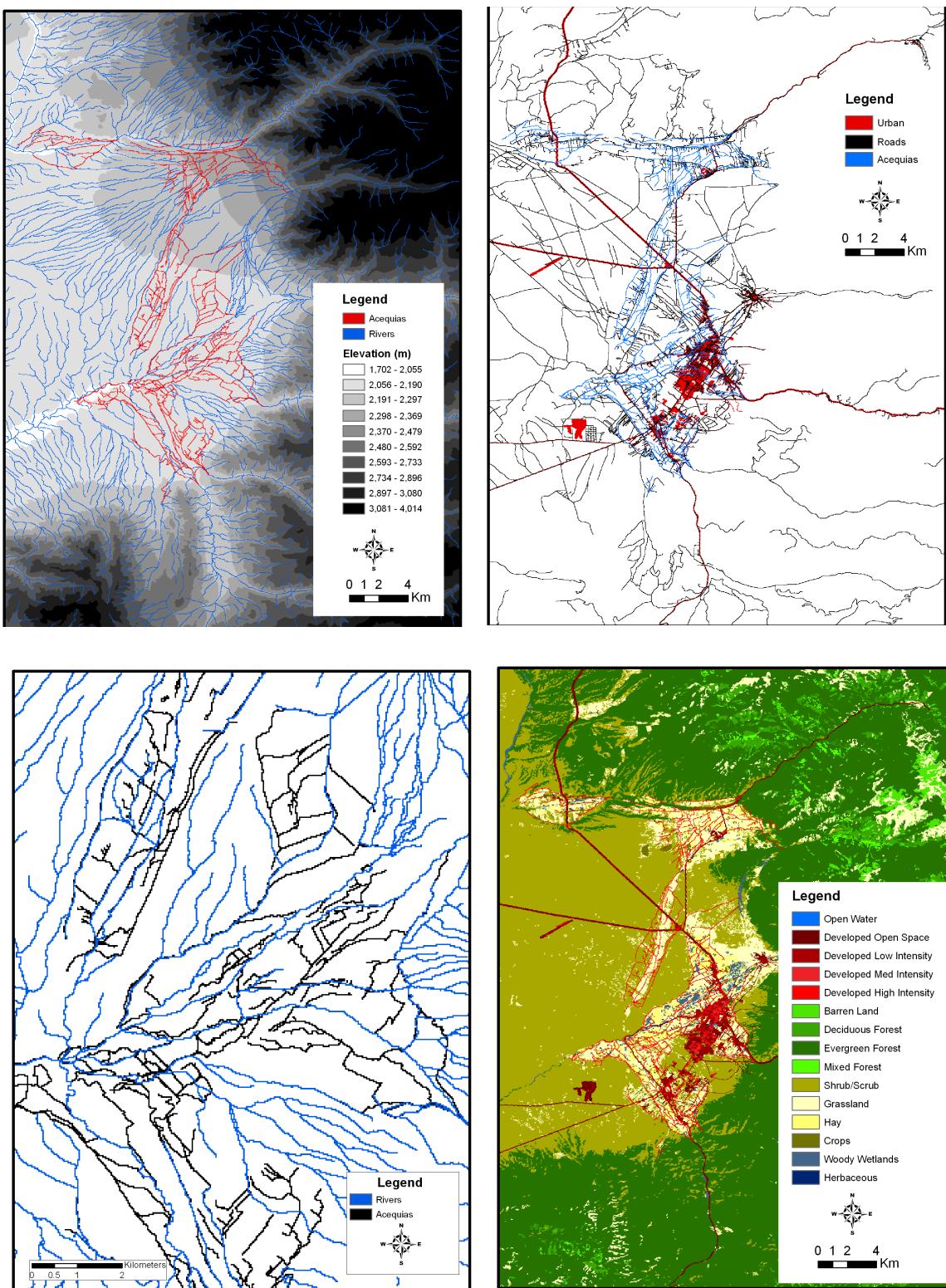


Figure 3. GIS data used within the model include (in clockwise order) elevation, urbanization, land use data, and waterways.

2.1 Simulation

Acequias are a complex system, and the importance of low-level dynamics makes it difficult to understand the macro-trends in local development. In this simulation, water, land, parciantes, and real estate agents are all simulated to try to explain the turnover of agricultural land into urbanized, residential space as the process is impacted by the ever critical question of access to water.

The propagation of water through the environment is accomplished by a network of rivers and acequias. Acequia links are special because acequias build up sedimentation - or decay the weight of the relevant water system link - every year unless they are maintained. A fraction of the water that flows through the decayed link is lost. An unmaintained section of acequia can thus eventually cut off all "downstream" nodes from access to hydration.

Land is a passive object, in that it is irrigated by the water network and cultivated by Parciantes. Parciantes can choose to grow various kinds of crops on their parcels of Land tiles, and the income from a parcel of tiles P on which a crop C is planted is given by equation 1.

$$I_p = C_{\text{price}} * P_{\text{size}} \quad (1)$$

The income derived from various crops is user-determined. As Parciantes hold multiple units of land as part of their land parcel, it is possible for one Parciante to plant a different crop on each of his units of land.

Parciante agents represent the individual acequia owners who make land use and acequia maintenance choices in the real world. In the context of the simulation, Parciantes choose whether to maintain their acequias, plant crops, and sell their land. They have a number of attributes, including a set of Land tiles, a sum of money, and a 'strategy'. The money attribute reflects agent resources, and if its value dips below a certain level the Parciante is forced to sell his land to any bidder. Money is expended when the agent helps maintain his acequia: the cost is a function of the length of the acequia A and the number of Parciantes N as shown in equation 2.

$$C = A_{\text{length}} / N \quad (2)$$

To reflect the importance of cultural heterogeneity and personal preference in these decisions, Parciante agents are further endowed with a 'strategy' that defines their approach to land use decisions. One example of such a strategy is the Traditionalist, who values his land and will hold onto it as long as his money holds out. A different strategy, the Sheep, leads the Parciante to observe what his neighbours are doing, and emulate the behaviour of the majority.

Real Estate Agents represent the rising demand for housing in the area, and their goal is to buy and develop as much land as possible. These agents are endowed with a certain budget and make offers to individual Parciantes. The offer prices follow the model of Filatova et al. (2009), that given an agent with budget B bidding on a parcel with transport cost T (proportional to the distance of the parcel from the road), the utility U is a function of the amenity A (here, the parcel size) and normalized distance from the

economic centre P weighted by a factor b. The offer price O is thus determined by equations 3-5.

$$Y = B - T \quad (3)$$

$$U = A^\alpha P^\beta \quad (4)$$

$$O = Y U^2 / (b^2 + U^2) \quad (5)$$

If the Parciante agent accepts the Real Estate Agent's offering price, the parcel of land is urbanized and the Parciante removed from the simulation. The distribution of Real Estate Agent budgets is determined by the user and can be modified mid-run. Fig. 4, sketches out the interactions between the agents in the simulation.

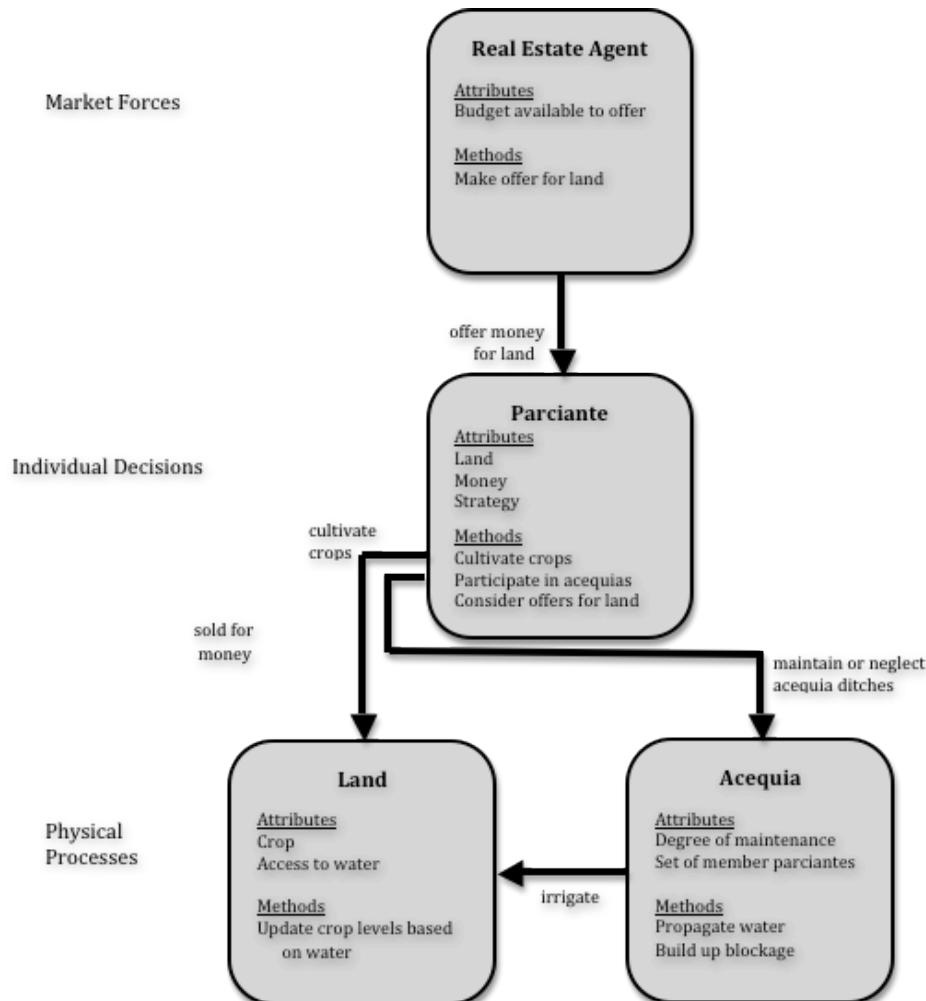


Figure 4. A flowchart of interactions between different kinds of agents.

3. Results

We have only begun to explore the vast range of experiments that the model presented here can represent. Verification has been completed and advanced validation is underway and will be presented at the conference.

The model was built in an iterative fashion, ensuring that each of the submodules demonstrates the appropriate behaviour over the space of inputs it can accept. The water network, for example, was tested by deactivating all of the Parciantes and allowing the acequias to go unmaintained for several decades as shown in fig. 5. As for validation, the land use output of the model looks reasonable upon inspection, but we plan to extend the validation process further by comparing it with the real land use patterns of the area derived from land cover data collected in 2008 using the Map Comparison Kit (Visser and de Nijs, 2006).



Figure 5. A view of the water network link weights at the beginning (left) and after 30 years of neglect (right). Darker links are stronger.

4. Conclusion

The model presented here uses empirical GIS data to build a realistic model of a complex socio-physical system. By representing the interacting physical, economic, and social processes, the interconnected nature of the acequia system is more precisely represented. It is the authors' hope that this model will be used by researchers seeking to answer questions about the rigorouslyness of this community resource management system, its specific strengths and critical weaknesses, and how to protect this traditional way of life.

5. Acknowledgements

The authors would like to acknowledge of the Department of Computational Social Science at George Mason for providing support for this research, Michael Cox for

sharing his data on the Taos, New Mexico acequia system with us, and John Paul Gonzales of the Santa Fe Institute, who is an acequia parciante.

6. References

- Cox, M., 2009. *Exploring the dynamics of social-ecological systems: the case of the Taos valley acequias*. Thesis (PhD). Indiana University.
- Filatova T, Parker D and van der Veen, A, 2009, Agent-based urban land markets: Agent's pricing behaviour, land prices and urban land use change. *Journal of Artificial Societies and Social Simulation* 12 (13).
- Hardin G, 1968. The tragedy of the commons, *Science* 162 (3859).
- Homer C, Dewitz J, Fry J, Coan M, Hossain N, Larson C, Herold N, McKerrow A, VanDriel J, and Wickham J, 2007, Completion of the 2001 national land cover database for the conterminous United States. *Photogrammetric Engineering & Remote Sensing*. 73(4): 337-341.
- Luke S, Cioffi-Revilla C, Panait L, Sullivan K and Balan G, 2005, MASON: A multi-agent simulation environment. *Simulation: Transactions of the society for Modeling and Simulation International*. 82(7): 517-527.
- Visser H and de Nijs T, 2006, The map comparison kit. *Environmental Modelling and Software*, 21(3): 346-358.

SAFEPED: Agent-Based Environment for Estimating Accident Risks at the Road Black Spots

Gennady Waizman, Eilon Blank-Baron, Itzhak Benenson

Department of Geography and Human Environment, Tel Aviv University, Ramat Aviv, 69978 Tel Aviv, Israel,
gwaizman@gmail.com, eilon.blancbaron@gmail.com, benny@post.tau.ac.il

1. Introduction

Police data on the number of traffic accidents clearly point to the "Black Spots", where the accident rate remains high in months and years. However, road safety research is still far from understanding why certain road locations are risky.

Essentially, we lack the knowledge of how pedestrians and drivers interact when facing a potentially dangerous traffic situation and, most important, the integrated framework that relates the data on human behavior to the real-world traffic situations.

So far, road safety is studies with the general purpose traffic simulation models extended towards including conflict statistics. This approach, however, is inherently limited. The dynamic road safety model should incorporate the variables that are critical for road incidents but superfluous for simulating general traffic: the characteristics of mechanical and functional characteristics of vehicles and in-vehicle systems and, especially, the rules of drivers' and pedestrians' behavior, including drivers and pedestrians awareness and reaction to each other (Gettman and Head, 2010).

We present safety oriented high-resolution spatial micro-simulation model of car and pedestrian traffic that enables direct simulation of the road accidents and associated risks.

2. SAFEPED Simulation Environment

To represent the dynamic reality at the Black Spot and merge it with the experimental data on drivers' and pedestrians' behavior we have developed SAFEPED - Multi-agent environment for spatially explicit microscopic 3D simulation of the Black Spot dynamics.

SAFEPED serves as a testbed for evaluating experimentally estimated drivers' and pedestrians' behavioral rules and estimating accident risks in various traffic situations. It aims at analyzing disadvantageous environmental design at the Black Spot and assessing alternative architectural solutions there.

The major features of the SAFEPED are as follows:

- SAFEPED agents are autonomously behaving pedestrians and drivers who see and estimate the 3D-movement of the other agents and react in response once in 0.04 sec;
- SAFEPED agents see each other in 3D and behave based on the 3D visibility
- The user defines the properties and goals of movement of the drivers and pedestrians participating in the traffic episode;
- The rules of agents' behavior are based on the experimental data, obtained from the videos, including the videos captured at the investigated Black Spot.
- SAFEPAD is validated based on the multi-view video footages.

During the simulation, SAFEPED records the full life-history of every agent, including all crash and near-crash episodes. The user can analyze the crash and near-crash statistics, rewind and replay the simulation starting from any moment of time, observe accidents from various viewpoints, including the viewpoints of the crash participants (Figure 1). The user can also intervene into the model dynamics by taking the full control over one or more agents. To represent the risks of the accidents, SAFEPAD applies indicators describing the conflicts between traffic participants, such as Time-to-Collision (TTC) and Post Encroachment Time (PET) (Morsink, 2008).

The paper presents the SAFEPED and the results of investigation of several accident scenarios.

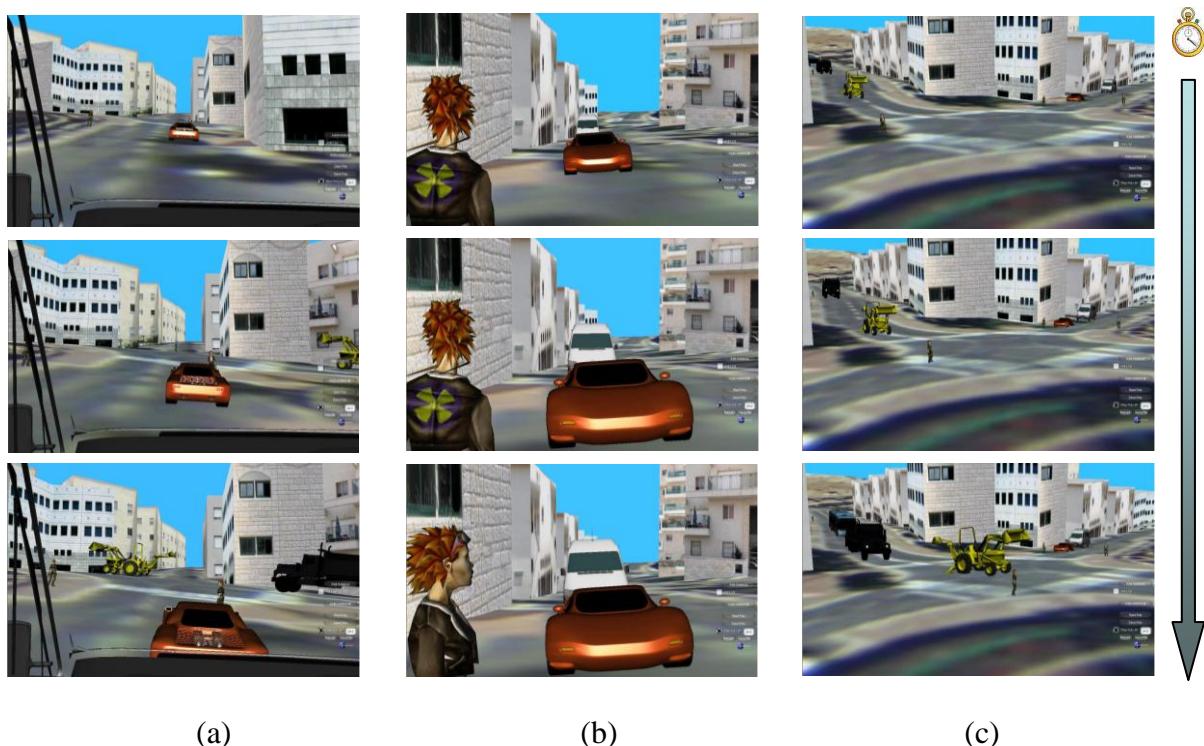


Figure 1. SAFEPAD traffic episode from the viewpoint of (a) drivers participating in the episode; (b) outer pedestrian located close to the episode; (c) outer pedestrian located far from the episode

3. References

- Gettman, D. and Head, L., 2003, Surrogate Safety Measures From Traffic Simulation Models, *Report No. FHWA-RD-03-050, Federal Highway Administration, Washington, DC*
 Morsink, P., Wismans, L., Dijkstra, A. 2008, Micro-Simulation For Road Safety Impact Assessment Of Advanced Driver Assistance Systems, *Seventh European ITS Congress, Geneva, paper 2676*

Polyline averaging using distance surfaces: a spatial hurricane climatology

K. N. Scheitlin¹, V. Mesev², J. B. Elsner³

¹Longwood University, Department of Biological & Environmental Sciences, Farmville, VA 23909, USA
 Telephone: (1) 434-395-2970
 Fax: (1) 434-395-2652
 Email: scheitlink@longwood.edu

²Florida State University, Department of Geography, Tallahassee, FL 32306, USA
 Telephone: (1) 850-645-2498
 Fax: (1) 850-644-5913
 Email: vmesev@fsu.edu

³Florida State University, Department of Geography, Tallahassee, FL 32306, USA
 Telephone: (1) 850-644-8374
 Fax: (1) 850-644-5913
 Email: jelsner@fsu.edu

1. Introduction

The US Gulf states are constantly hit by hurricanes, causing widespread damage resulting in economic loss and occasional human fatalities. Much of the existing research is focused on devising models that predict where hurricanes are likely to hit in the future. These models typically create what are called *hurricane climatologies* of given geographic areas, and are built from data on storm intensity, size, and return rate from past events, in addition to climatic variables that are known to be favourable for hurricane activity (Jagger et al. 2001, Elsner and Jagger 2006, Landsea et al. 2006). However, hurricane climatologies frequently omit information on the spatial patterns of hurricane movement. In other words, data on the linear tracks that hurricanes take are seldom incorporated. A more complete *spatial* hurricane climatology would improve our understanding of the temporal frequency of hurricane events and hence their propensity for further hurricane occurrence.

1.1 Hurricane tracks

The hurricane tracks that would be used to build a spatial hurricane climatology are recorded by the National Hurricane Center (NHC) in Miami, Florida. They are initially plotted as individual points, which can later be connected in chronological order to form a linear track. From a geocomputational standpoint, these tracks are essentially one-dimensional polylines, fixed with starting and end points corresponding to the hurricane's origin as an organized storm and ultimate dissipation, respectively. And as linear polylines the tracks can be analyzed when searching for spatial patterns, in particular the relative frequency. One way to measure this is to calculate an *average track* by combining an entire set of polylines that represent past hurricanes for a given geographic area. The average track would represent not only the most frequented path of past hurricanes but also a probable path for future hurricane activity.

1.2 Average track

The calculation of an average hurricane track using polylines may at first appear trivial. Yet it is absent from the meteorological literature. One geocomputational technique is to first create *distance surfaces* for each polyline that represents an individual hurricane track. Distance surfaces are isotropic surface interpolations that represent regular Euclidean distance intervals and are very much akin to regular-interval buffering techniques in most GIS software packages (Hirata 1995). The procedure is illustrated by fig 1. Three polylines have a range of [0,1] and a value x at every $0.01y$ – this is where $x=y+a$, and where a is normally distributed with mean of 0 and standard deviation of 0.1 (fig 1a). The Euclidean distance maps are then generated for each of the polylines, resulting in the calculation of isolines or buffer zones that represent inverse distance radiating from the entire length of the polylines in units of x (fig 1b).

The next stage is aggregation of all three polyline distance surfaces, and involves the calculation of composite isolines based on the average of the three—an inverse distance weighted (IDW) spatial analysis procedure implemented by a union function (fig 1c). The outcome is that every point on the composite map is the product of the average value of all three distance surfaces; mathematically, a value of some number v indicates that the three polylines are, on average, v units of x away from that point. For instance, shorter distance values have the most agreement between the polylines, and a value of zero means that all of the polylines intersect at that point.

The last stage is for a new polyline to be digitized along the center of the isoline with the lowest composite distance—this is considered the average polyline (fig 1d). And is essentially the least-cost route, if the composite distance map was to be viewed as a typical cost-density surface. A decision must also be made on the range of the polyline; i.e. the start and end points--what approximate linear range of y values does the average polyline make sense? This is not always a straightforward case with physical phenomena such as hurricanes where their precise beginning and end are less well-defined and variable.

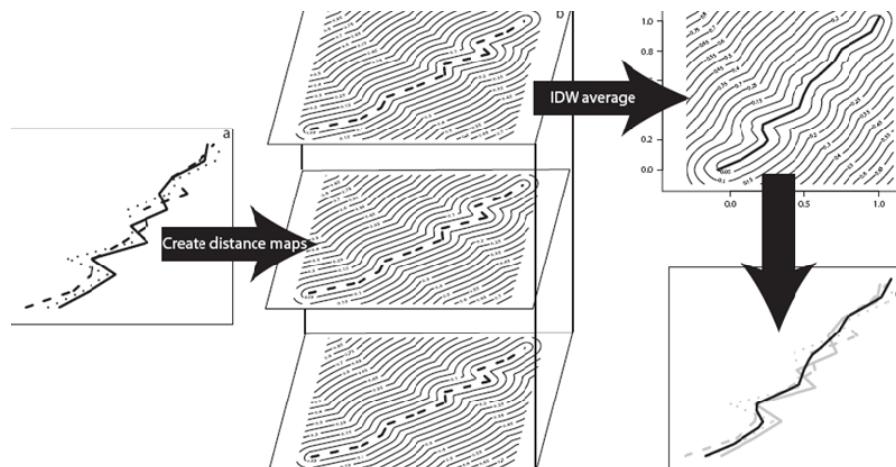


Figure 1. Polyline averaging using distance surfaces: a) three polylines representing hurricane tracks, b) distance surfaces for each polyline where isolines represent radiating distances from the polyline in units of x , c) inverse-distance weighted average distance map, d) average polyline (black), and the original set of polylines (grey).

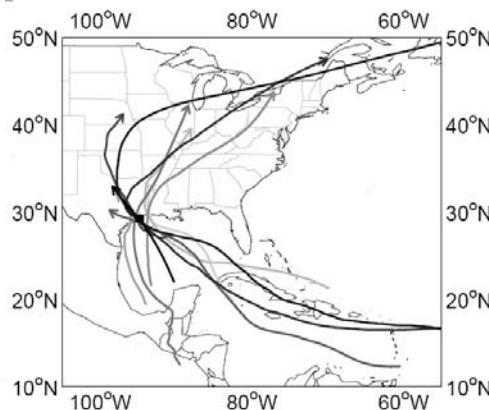


Figure 2

Distance mapping allows all tracks to be averaged, regardless of intersections or parallel points (similar latitudes). Moreover, the procedure works for two or more lines in the same spatial plane. We explored kernel density estimation using points (remember, the hurricane tracks are polylines connecting hourly observations) but because hurricanes move at different speeds points are recorded at irregular intervals (they tend to cluster for slower hurricanes). Using only latitude and longitude was also unacceptable, because hurricanes have different start and end points, and some even cross one line of latitude multiple times.

2. Hurricane track averaging: the case of Galveston

The averaging method can be illustrated by an example from the city of Galveston in Texas, USA (fig 2). Known for its vulnerable location, ten major hurricane tracks (wind speeds $>50 \text{ m s}^{-1}$) that passed within 100 km between 1851 and 2009 were accessed from the HURDAT dataset (NHC) and from a qualitative archival collection (Chenoweth 2006, Scheitlin et al. 2010). Distance surfaces were generated for each of the ten tracks and then combined to calculate the average (fig 3). In addition, the IDW assumption is modified by a distance-decay parameter to adjust for diminishing importance with increasing distance (darker tracks in fig 2 represent hurricanes that passed closer to the city). The formula for the average distance map $D(\mathbf{s})$ is now:

$$D(\mathbf{s}) = \frac{\sum_{k=1}^{10} w_k D_k(\mathbf{s})}{\sum_{k=1}^{10} w_k} \quad (1)$$

where $w_k = \frac{1}{d(e, t_k)}$ and $d(e, t_k)$ is the great-circle distance from the track to Galveston and $D_k(\mathbf{s})$ is the distance surface for track k .

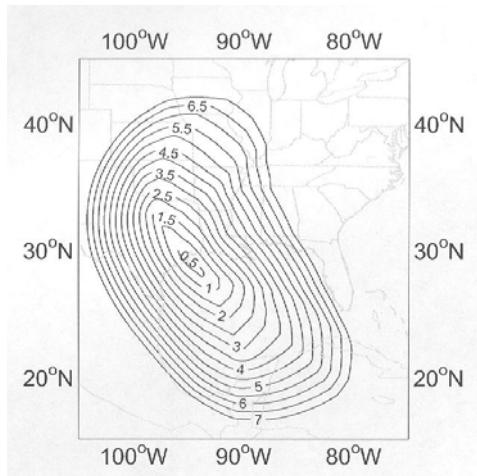


Figure 3

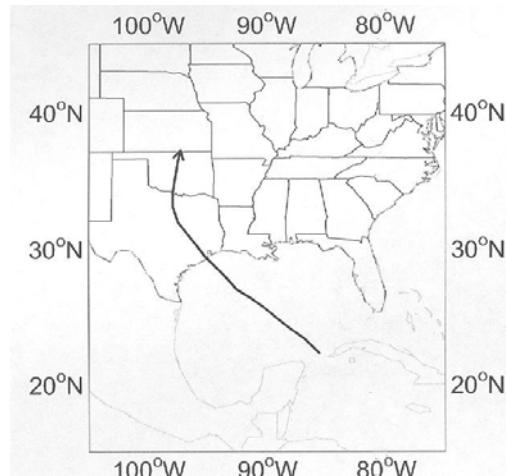


Figure 4

3. Conclusions

The Galveston example produced a single distance weighted average track (fig 4) that represents the most likely path future hurricanes may take. This is important for many reasons. Firstly, it marks the most vulnerable coastal strips that may be damaged by hurricanes; this in itself is priceless for helping develop effective mitigation policies. Secondly, from a computationally reductionist standpoint, an average track, like all averages, condenses multiple tracks into a convenient single-track summary. And like all averages, the more observed frequencies the stronger the relevance and confidence of the average. In a sense it is also a means for visualizing simple trends from complex data. Thirdly, the averaging technique adds the spatial component to existing hurricane climatologies and predictive models, and this in turn provides a more complete framework with which to study the climatic factors responsible for fuelling hurricane activity. And lastly, the calculation of an average may also be used to augment, strengthen and to complete historical records by helping fill gaps in missing years.

Possible improvements to the polyline averaging method include estimating some degree of error, where uncertainty information would be obtained from polyline similarity testing (Kuijpers et al. 2006). This is where similarity between the whole set of polylines would provide information regarding the relevance of an average polyline. If the polylines have little similarity then the average polyline may not have as much physical relevance. Once an average polyline is created, polyline similarity testing can provide information about the difference between the average polyline and the original set of polylines. Calculating the standard distances from the average polyline to the polyline set is another way to obtain similar information. Such testing should also help determine a reasonable range of values within which the average polyline is finally digitized.

In terms of portability, polyline averaging serves an immediate purpose for summarizing hurricane tracks. But it also may be relevant for other severe atmospheric phenomena as well as animal migratory patterns, and maybe even traffic and pedestrian flows.

4. References

- Chenoweth M, 2006, A reassessment of historical Atlantic basin tropical cyclone activity, 1700–1855. *Climatic Change*, 76, 169-240.
- Elsner JB and Jagger TH, 2006, Prediction models for annual U.S. hurricane counts, *Journal of Climate*, 19, 2935-2952.
- Hirata T, 1995, A unified linear-time algorithm for computing distance maps, *Information Processing Letters*, 58, 129-133.
- Jagger TH, Elsner JB and Niu X, 2001, A dynamic probability model of hurricane winds in coastal counties of the United States, *Journal of Applied Meteorology*, 40, 853-863.
- Kuijpers B, Moelans B and Van de Weghe N, 2006, Qualitative polyline similarity testing with applications to query-by-sketch, indexing and classification, *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*, ACM, New York, NY, USA, 11-18.
- Landsea CW, Harper BA, Hoarau K and Knaff JA, 2006, Can we detect trends in extreme tropical cyclones? *Science*, 28, 452-454.
- Scheitlin KN, Elsner JB, Malmstadt JC, Hodges RE and Jagger TH, 2010, Toward increased utilization of historical hurricane chronologies, *Journal of Geophysical Research*, 115, D03108, doi:10.1029/2009JD012424.

Using Fine Resolution Population Data and Spatial Interaction Modeling to Estimate Risk from Airborne Toxic Releases

J. F. Conley¹, R. N. Stewart²

¹Department of Geology & Geography, West Virginia University,
330 Brooks Hall, 98 Beechurst Ave., Morgantown WV, USA, 26506
Telephone: (1) 304-293-6352
Fax: (1) 304-293-6522
Email: Jamison.Conley@mail.wvu.edu

²Geographic Information Science & Technology Group, Oak Ridge National Laboratory,
PO Box 2008 MS 6017, Oak Ridge, TN, USA, 37831
Telephone: (1) 865-574-7646
Fax: (1) 865-241-9272
Email: stewartm@ornl.gov

1. Introduction

The close connection between pollution concentration, exposure, toxicity, and public health (US EPA 1989, 1997) makes estimating a pollution source's impact on a community a vital task in environmental health. Two common methods of estimating this impact are 1) considering locations in the same areal unit as a pollution site as exposed (e.g., Croen *et al.* 1997) and (2) considering locations within a buffer around the pollution site as exposed (e.g., Kearney and Kiros 2009).

Three concerns arise in these methods. Each has been addressed individually, but no national-scale study has addressed all three. The first concern is the Modifiable Areal Unit Problem (Openshaw 1984), which has been addressed by using raster cells instead of census units (Mennis 2002, Mohai and Saha 2006). Second, what is the mathematical function relating distance from a site and that site's impact? Many studies treat this as a flat surface; if a person is within the buffer, they are at risk. Other studies measure risk as the distance to the nearest site (Mohai and Saha 2006). This distance-based approach recognizes that a site's impact decays with distance. However, a simple distance measure ignores the volume and toxicity of the chemicals released at that site. Therefore, we want a model incorporating the distance from the pollution site and the volume and toxicity of the chemicals released. Only Cutter *et al.* (2001) use such a model within the environmental justice literature. Spatial interaction models from economic geography (Sen and Smith 1995) can meet these requirements. The third concern is anisotropy. Most distance-based studies treat all directions as equivalent, although this is unrealistic, as prevailing winds carry pollutants further downwind. Studies using site-specific atmospheric models (e.g. Fisher *et al.* 2006) account for this, but no nationwide study accounts for anisotropy.

This research combines a raster-based analytical approach with different anisotropic spatial interaction models to more accurately estimate the population at risk. We use LandScan USA © data from Oak Ridge National Laboratory (Bhaduri *et al.* 2007), which estimates daytime and nighttime population of the continental United States at a 90-meter grid resolution. We calculate the risk at each LandScan grid cell using a series of

increasingly complex spatial interaction models in which each added step addresses an additional concern.

2. Data

Our health data are lung cancer age-adjusted mortality rates from 1990 through 2006 provided by the US National Cancer Institute's Surveillance Epidemiology and End Results (SEER) database. Our pollution data are airborne releases of lung carcinogens identified in the EPA's IRIS (Integrated Risk Information System), HEAST (Health Effects Assessment Summary Tables), and PPRTV (Provisional Peer Reviewed Toxicity Values) databases. We include all releases in the continental United States from 1987 through 1996 in the US Environmental Protection Agency's Toxic Release Inventory (TRI) database. The TRI data period ends before the health data period to reflect a lag time between chronic exposure to toxic chemicals and lung cancer development. We use covariates listed in table 1 from the 1990 US Census and other sources. The final dataset is the nighttime population from the LandScan USA dataset.

percent with no high school degree	smoking rate
percent with college degree	physicians per 1000 residents
poverty rate	spatial indicator: South
unemployment rate	spatial indicator: Midwest
percent non-white	spatial indicator: West
percent male	spatial indicator: Appalachian

Table 1. Demographic and behavioral covariates used in regression analyses.

3. Methods

We develop and analyze four risk datasets with increasing complexity.

1) The first set uses simple models given in equations 1-4, which respectively give a buffer, Cutter *et al.* (2001)'s decay formula, a power-based spatial interaction model, and an exponential spatial interaction model. In the equations, l_{iz} is the impact of pollution site i on location z , r_i is the release volume at site i , d_{iz} is the distance from site i to location z , and α , θ , and T are model parameters. We evaluated a range of values for each parameter.

$$l_{iz} = \begin{cases} r_i^\alpha : d_{iz} < T \\ 0 : d_{iz} \geq T \end{cases} \quad (1)$$

$$l_{iz} = \begin{cases} r_i^\alpha \left[1 - \frac{d_{iz}^{-\theta}}{T^{-\theta}} \right] : d_{iz} < T \\ 0 : d_{iz} \geq T \end{cases} \quad (2)$$

$$l_{iz} = \begin{cases} r_i^\alpha (d_{iz})^{-\theta} : d_{iz} < T \\ 0 : d_{iz} \geq T \end{cases} \quad (3)$$

$$l_{iz} = \begin{cases} r_i^\alpha \exp(-\theta d_{iz}) & : d_{iz} < T \\ 0 & : d_{iz} \geq T \end{cases} \quad (4)$$

We incorporate anisotropy by using wind speed and direction to transform the distance values in these equations. The three transformations are shown in fig. 1. This initial dataset assumes a constant wind speed and westerly direction for the entire country. Mathematical details are available from the first author. We test these three transformations and no transformation.

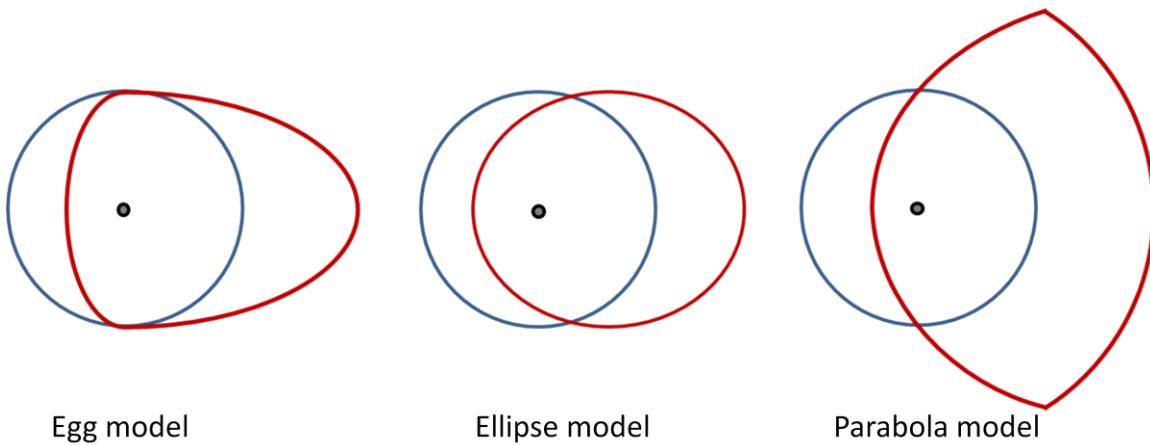


Figure 1. Three models of anisotropy: the blue circles are the untransformed isotropic distances and the red shapes are the transformed distances.

- 2) The second set uses Inhalation Unit Risk values from the IRIS, HEAST and PPRTV databases to assign chemical-specific α parameters.
- 3) The third set uses climate data from the National Climate Data Center to vary speeds and directions for prevailing winds by location.
- 4) The fourth set uses chemical weight values from the IRIS, HEAST and PPRTV databases to assign chemical-specific θ parameters.

Each set of risk estimates are analyzed in multivariate ordinary least squares (OLS) and spatial error regressions using the mortality rates as the dependent variable and the covariates and spatial interaction risk estimates as independent variables. The regressions are conducted in *R*.

4. Results and Discussion

Table 2 shows the AIC values and parameterizations of the OLS and spatial error regressions that minimize the AIC. Parameters values with an asterisk were subsequently modified on a chemical-by-chemical basis. As the table shows, the best parameterization was consistent across both regression models and all four datasets.

	<i>Set 1</i>	<i>Set 2 (add toxicity)</i>	<i>Set 3 (vary winds)</i>	<i>Set 4 (add molecular weights)</i>
OLS AIC	22361.63	22823.70	22806.41	22794.90
Spat. Err. AIC	22357.75	22817.37	22798.24	22786.15
α	2.5	2.5*	2.5*	2.5*
θ	1.0	1.0	1.0	1.0
T	100	100	100	100*
anisotropy decay function	parabola buffer	parabola buffer	parabola buffer	parabola buffer

Table 2. OLS and spatial error regression results.

The spatial error regression outperformed the OLS regression. Varying the wind speed and direction (set 3) improved over set 2, and using the molecular weight to vary the θ parameter provided further improvement. However, the AIC was optimized with the simplest risk estimation model. Additional examination is needed to ascertain why this happens. Similarly, the simplest spatial decay function, a buffer, performed the best. Lastly, comparing these results against those in Conley (2011) demonstrate that using LandScan data improves over using county-level aggregate populations (min AIC = 22390.78) and the US EPA's RSEI risk-related results (AIC = 22866.82).

We could not examine all possible combinations of α , θ , T , the anisotropy shape, and the decay function, so may not have the optimal parameterization. A fuller search of this parameter space is needed. Continued work will also use the SCIPUFF atmospheric dispersion model to compare atmospheric modeling against spatial interaction approaches.

5. Acknowledgements

This research was supported by the West Virginia Rural Health Research Center (Office of Rural Health Policy, Health Resources and Services Administration, PHS Grant No. 1 U1CRH10664-01-00) and by the first author's appointment to the U.S. Department of Energy Higher Education Research Experiences for Faculty at the Oak Ridge National Laboratory administered by the Oak Ridge Institute for Science and Education.

6. References

- Bhaduri B, Bright E, Coleman P and Urban M, 2007, LandScan USA: A High Resolution Geospatial and Temporal Modeling Approach for Population Distribution and Dynamics. *GeoJournal*, 69:103-117.
 Conley J, 2011, Estimation of exposure to toxic releases using spatial interaction modeling. *International Journal of Health Geographics*, 10:20.

- Croen LA, Shaw GM, Sanbonmatsu L, Selvin S and Buffler PA, 1997, Maternal Residential Proximity to Hazardous Waste Sites and Risk for Selected Congenital Malformations. *Epidemiology*, 8:347-354.
- Cutter SL, Hodgson ME and Dow K, 2001, Subsidized Inequities: The Spatial Patterning of Environmental Risks and Federally Assisted Housing. *Urban Geography*, 22:29-53.
- Fisher JB, Kelly M and Romm J, 2006, Scales of environmental justice: Combining GIS and spatial analysis for air toxics in West Oakland, California. *Health & Place*, 12:701-714.
- Kearney G and Kiros GE, 2009, A spatial evaluation of socio demographics surrounding National Priorities List sites in Florida using a distance-based approach. *International Journal of Health Geographics*, 8:33.
- Mennis J, 2002, Using Geographic Information Systems to Create and Analyze Statistical Surfaces of Population and Risk for Environmental Justice Analysis. *Social Science Quarterly*, 83:281-297.
- Mohai P and Saha R, 2006, Reassessing racial and socioeconomic disparities in environmental justice research. *Demography*, 43:383-399.
- Openshaw S, 1984, *The Modifiable Areal Unit Problem*. Geo Books. Norwich, UK.
- Sen A and Smith TE, 1995, *Gravity Models of Spatial Interaction Behavior*. Springer, New York, NY, USA.
- US EPA, 1997. *Exposure Factors Handbook (Final Report)*. EPA/600/P-95/002F a-c. U.S. Environmental Protection Agency, Washington, DC, USA.
- US EPA. 1989. *Risk Assessment Guidance for Superfund: Volume I - Human Health Evaluation Manual. (Part A)* EPA/540/1-89/002. Office of Emergency and Remedial Response, U.S. Environmental Protection Agency, Washington, D.C., USA.

Defining spatial weights matrices in ecological research

T. A. Nelson¹, C. Robertson²

¹University of Victoria, Department of Geography, Spatial Pattern Analysis and Research (SPAR) Lab, Victoria BC, V8W 3R4, Canada
 Telephone: +1 (250) 472 5620
 Email: trisalyn@uvic.ca

² Wilfrid Laurier University, Department of Geography and Environmental Studies, Waterloo, Ontario, N2L 3C5, Canada
 Telephone: +1 (519) 884 0710 x2160
 Email: crobertson@wlu.ca

Spatial analysis has gained popularity in ecology. However, standard spatial weights matrices can be problematic in ecology where landscape structure can act as a barrier to spatial relatedness. As an example, mountains can create heterogeneity in spatial relationships and may create functionally autonomous populations of a single species. Our goal is to present an approach to semi-automated delineation of ecologically appropriate spatial neighbourhoods. The general approach is to watershed boundaries, derived from a digital elevation model, to constrain the extent of individual spatial neighbourhoods. To demonstrate our methods we present a case study using NOT FINISHED>..

1. Introduction

Spatially minded ecologists have long understood the complex interaction between spatial patterns and process (Legendre and Fortin 1989, Levin 1992). As such, quantifying spatial pattern in biological phenomena has gained popularity in ecological research (Fortin and Dale 2005). For instance, local measures of spatial autocorrelation are used to identify hot spots or locations where abundance of a population is greater than expected (Nelson and Boots 2008). Another example of spatial analysis commonly used in ecology is geographically weighted regression (GWR) (e.g., Wang et al. 2005).

Spatial analysis usually requires the definition of a spatial neighbourhood. Most often neighbourhoods are assigned using definitions based on distance, k-neighbours, or contiguity (Griffith 1996). These standard definitions of spatial neighbourhoods are often inappropriate in ecology where natural physical barriers, such as mountains, are present and create heterogeneity in spatial relationships. For instance, populations of a species may be separated by topographic features creating functionally autonomous sub-populations. However, a neighbourhood defined by distance or contiguity may include individuals from multiple, unrelated sub-populations. As an example, mountain pine beetle populations disperse along valleys and mountains are typically a barrier to dispersion. As such, when quantifying hot spots in data on mountain pine beetle infestations, spatial neighbourhoods should be restricted within valleys or along the side of a mountain and should not reach beyond the mountain peak.

When study areas are small it is reasonably easy to modify spatial neighbourhoods manually. However, as with many applied areas of spatial analysis, data sets are increasingly large and automated approaches are required for appropriate spatial neighbourhood delineation. The goal of this research is to present an approach to semi-automated delineation of ecologically appropriate spatial neighbourhoods. To meet this goal we use terrain data to delineate watersheds and demonstrate how watershed can be used to constrain spatial neighbourhoods. Watersheds are useful because they are often

divided by mountain peaks and can be used to partition terrain that separates species' populations. We apply our approach to a case study on data from an epidemic mountain pine beetle infestation in British Columbia, Canada.

2. Study Area and Data

The study area is a southern region of Vanderhoof Forest District located in central British Columbia. Vanderhoof was chosen for this study due to the variability of topography which ranges from flat in the north to mountainous in the south. As well, as with many locations in British Columbia, the Vanderhoof Forest District has experienced epidemic levels of mountain pine beetle (*Dendroctonus ponderosae* Hopkins) over the last decade, and the infestation continues to have substantial impact on pine (*Pinus* spp.) forests and resource-dependent communities in the area. Periodic population eruptions occur when an abundance of susceptible host trees coincides with climatic conditions amenable for beetle survival (e.g., Safranyik and Carroll 2006). Although epidemic populations are a natural component of forest disturbance, large infestations have substantial impacts and provide unique challenges to forest managers (Safranyik et al. 1974).

Vanderhoof forest District monitors the mountain pine beetle infestation using point-based, global positioning system (GPS) aerial surveys. Aerial surveys of mountain pine beetle infestations use indicators of pine mortality, mainly changes in crown foliage color, to monitor mountain pine beetle activity. During helicopter aerial surveys, clusters of visually infested trees are identified, typically those with yellow and red crowns, indicating mortality one to two year prior to infestation, and a GPS is used to map cluster centers with a point. For each cluster, the number of infested trees is estimated and the infesting insect species recorded. Attributes have been shown to be accurate to ± 10 trees for 92.6 % of points (Nelson et al. 2006). In the current study, there are a total of 25,460 GPS points.

A DEM was used to derive watershed boundaries. The elevation model had 25 m^2 grid cells and was created from 1:20,000 scale Terrain Research Information Management data (Province of British Columbia 1996). The data were interpolated using a linear interpolation process and the DEM is reported to be accurate within 10 m (Province of British Columbia 1996).

3. Methods

We delineated watersheds with standard tools available in ArcGIS. For details on watershed delineation we refer the reader to Chang (2006 chapter 15). The input elevation model was defined for an area that extended 50 km beyond Vanderhoof's boundary to ensure watershed boundaries are defined without edge effects. A threshold of 37,500 was applied to the accumulation layer. Selecting the accumulation threshold value is largely subjective. Larger thresholds will generate smaller watersheds and threshold selection enables users to control the scale of watershed delineation.

Watersheds are defined in both the flat and mountainous terrain throughout the study area. However, neighbourhoods only need to be adjusted in mountain regions, as terrain in low lying areas will not impact the relatedness of species populations. To select mountainous watersheds, each watershed was attributed with maximum elevation. All

watersheds with a maximum elevation ≥ 1800 m were identified as regions where the neighbourhood should be adjusted. The maximum elevation threshold is also subjective, but in this case was based on mountain pine beetle biology (Safranyik and Carroll 2006).

Binary spatial weights matrices were calculated using both a 3km distance bandwidth, as 3km is the maximum distance mountain pine beetle are thought to fly within stand canopies (Safranyik and Carroll 2006). For watersheds with elevations ≥ 1600 m spatial neighbours could only be selected from within the watershed boundary. Also, for comparison, the spatial weights matrix was calculated without consideration of watersheds boundaries.

Using all spatial weights matrices Moran's I_i was calculated. Moran's I_i is a local measure of spatial autocorrelation which is useful for identifying locations of clusters and outliers of extreme values (Anselin 1995). For this research, we focus on identifying clusters of high values that are unexpected based on chance. We term clusters of high values hot spots.

4. Preliminary Results

Preliminary results of hotspots detected from constrained and unconstrained spatial neighbourhoods indicate the location of hot spots varies only slightly when constraining neighbours to within watershed boundaries will change the location of hot spots. The general pattern of hot spots did not change; however, there were some differences in points characterized as belonging to infestation hot spots. When the data were unconstrained 3079 points were identified as being part of an infestation hot spot. Constraining spatial neighbourhoods by watersheds increased the number of hot spots points to 3109. Eighty-six points identified as hot when spatial neighbourhoods were unconstrained became hot spots once the spatial neighbourhoods were constrained by watersheds.

5. Discussion and Conclusion

The location of hot spots did not change substantially when spatial neighbourhoods were unconstrained or constrained by terrain features. This is likely due to the scales of beetle spread relative to terrain features. The valleys tend to be much larger than 3km in width, but 3km is a reasonable scale for modelling mountain pine beetle spread. As such, results indicate spatially adaptive neighbourhoods based on ecological information are most important when the scale of the spatial processes is similar or coarser than the scale of terrain features that influence the spatial process.

When conducting spatial analysis in ecology, spatial neighbourhoods should not cross natural barriers. We present a simple approach for semi-automated generation of ecologically relevant neighbourhoods that only requires a DEM for implementation. In mountainous regions, neighbourhood selection is limited to individuals within the same watershed. As spatial analysis results are highly dependent on the spatial weighting, ecologically informed weights matrices are important for ensuring relevant results. The approach here differs from recently proposed methods for constructing the weights matrix which depend only on the data (e.g., Aldstadt and Getis 2006), in that we are incorporating ecological knowledge and ancillary data to define barriers via watershed delineations. A watershed boundary should only be used to constrain spatial neighbourhoods in mountainous regions or other situations where terrain impacts dispersal or traversability of the landscape by a species. In low lying areas constraining

spatial neighbourhoods by watershed boundaries will cause additional and unnecessary edge effects. In addition to hotspot analysis, ecologically constrained spatial weights matrices have potential use in other spatial methods that require specification of spatial relationships such as simultaneous autoregressive models (Lichstein et al. 2002) and global measures of spatial autocorrelation (Cliff and Ord 1981).

6. References

- Anselin L, 1995, Local indicators of spatial association: LISA. *Geographical analysis*, 27(2): 93-115.
- Aldstadt J and Getis A, 2006, Using AMOEBA to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis*, 38(4): 327-343.
- Chang K, 2006, *Introduction to Geographic Information Systems: Third Edition*. McGraw Hill, New York, NY.
- Cliff AD and Ord JK, 1981, *Spatial processes: models & applications*. Pion Ltd, London, UK.
- Fortin MJ and Dale MR, 2005, *Spatial Analysis: A Guide for Ecologists*. Cambridge University Press, Cambridge, UK .
- Griffith DA, 1996, Some Guidelines for Specifying the Geographic Weights Matrix Contained in Spatial Statistical Models. In: Arlinghaus A (ed), *Practical Handbook of Spatial Statistics*. CRC Press, Boca Raton, USA.
- Legendre P and Fortin MJ, 1989, Spatial pattern and ecological analysis. *Plant Ecology*, 80(2): 107-138.
- Levin SA, 1992, The problem of pattern and scale in ecology. *Ecology*, 73(6):1943-1967.
- Lichstein JW, Simons TR, Shriner SA and Franzreb KE, 2002, Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, 72(3): 445-463.
- Nelson T and Boots B, 2008, Detecting spatially explicit hot spots in landscape-scale ecology. *Ecography*, 31(5):556-566.
- Nelson T, Boots B and Wulder MA, 2006, Large-area mountain pine beetle infestations: Spatial data representation and accuracy. *Forest Chronicle*, 82: 243– 252.
- Province of British Columbia 1996, *Gridded DEM specifications*. Ministry of Sustainable Resource Management, Victoria, BC, Canada.
- Safranyik L and Carroll AL, 2006, The biology and epidemiology of the mountain pine beetle in lodgepole pine forests. In: Safranyik L and Wilson B (eds), *The Mountain Pine Beetle: A Synthesis of its Biology and Management in Lodgepole Pine*. Natural Resources Canada, Victoria BC.
- Safranyik L, Shrimpton DM and Whitney HS, 1974, *Management of lodgepole pine to reduce losses from the mountain pine beetle: Forestry Technical Report No. 1*. Natural Resources Canada, Victoria BC, Canada.
- Wang Q, Ni J and Tenhunen J, 2005, Application of a geographically-weighted regression analysis to estimate net primary production of Chinese forest ecosystems. *Global Ecology and Biogeography*, 14(4):379-393.

Geometric Techniques to Speed up

Geospatial Feature Matching

Constantinos Tsirogiannis¹

¹TU Eindhoven, Postbus 513
5600 MB Eindhoven
Telephone: +31-40-2473201
Fax: +31-40-2436685
Email:ctsirogi@win.tue.nl

1. Introduction

Many up-to-date GIS applications make use of more than one digital representation of the same geospatial entity. Information extracted by multiple representations is compared and combined in order to construct data sets that are of higher quality than each of the initial data sets/representations; a process known as conflation. One of the most important components of conflation is feature matching: given two or more sets of features, each feature from one set is matched with exactly one feature of every other set to convey that they represent the same object in the real world. In most cases, features are represented as 2D or 3D geometric objects such as points, linear segments and polygons. Two features from different sets are likely to be matched if they have a small distance with each other or appear similar under some measure of similarity. It is also possible in practice that a feature of one data set corresponds to more than one features of another data set. From hereon we consider the simplest version of matching, that is one-to-one matchings between two different data sets of geometric objects. We shall refer to these two sets as the red and blue feature sets.

There exist many feature matching algorithms in the GIS literature. All these algorithms comprise of two stages. In the first stage for each red (blue) feature f a *candidate set* $C(f)$ is defined, that is a subset of blue (red) features that can be matched to f . A weight is then assigned between every pair (f, f') where f' is an element of $C(f)$. In the second stage, a matching is computed given the weighted pairs of the previous stage. If the weights represent a distance measure then a minimum weight matching is sought, otherwise if the weights represent a similarity measure a maximum weight matching is sought. Most of the matching algorithms in GIS use a greedy approach to compute the matching in this stage. Thus it is not guaranteed that the computed matching is the optimal solution.

Li and Goodchild (2010) recently presented a non-greedy algorithm for matching features between two different sets. Their algorithm first computes the distance for every red-blue pair of features. Then a matching is computed by formulating a Linear Program (Ferguson 1955) which is solved by standard LP software. They show experimentally that their method matches correctly a higher percentage of features than the standard greedy approaches. However, the execution time of their algorithm is large even for medium size data sets.

This is not surprising if we consider the size of the candidate sets that are constructed by their algorithm. If each of the two feature data sets consists of n elements then the total size of all computed candidate sets is n^2 . Thus for $n=1000$ the total size of candidate sets is one million and the resulting Linear Program consists of approximately three million numerical values. Solving an LP instance of this size takes considerable time, depending also on the software used. Li and Goodchild therefore address the need of using a method to reduce the number of computed distances among feature pairs and thus speed up their approach.

Indeed, it is not reasonable to consider so large candidate sets. In practice, for say a red feature f only few blue elements will have small distance from f and are thus likely to be matched with f . Hence we can use a *candidate selection technique* to avoid computing the weights for most possible pairs of features. Such techniques were implicitly used in matching algorithms before (Kim et al. 2010) yet those approaches either are of an ad hoc nature as they depend on specific representations (e.g. only for raster data) or involve non-geometric parameters (semantics).

In the current work we list a collection of candidate selection techniques that rely on concepts from the field of Computational Geometry (de Berg et al. 2008). We have used the Computational Geometry Algorithms Library (CGAL 2010) to conduct matching experiments on geospatial data sets using these techniques. We show that in practice the candidate sets can be kept very small without affecting substantially the output quality of the matching algorithm.

2. Proposed Candidate Selection Techniques

We now describe the techniques that we use in our experiments. The techniques are defined here for point sets yet they can be extended for other geometric objects too. In each of these techniques we first construct a data structure over one of the two input object sets, say the red set. For each blue object p we execute a query in this data structure. The output of this query is the candidate set of p . We construct candidate sets only for the blue objects which we consider sufficient for computing the final matching. The candidate selection techniques are the following:

- I) Natural neighbour selection (NNS): A Voronoi diagram is constructed over the red points. The queried point p is inserted in this diagram and its Voronoi cell $vc(p)$ is computed. The red points whose cells occupied the area of $vc(p)$ before inserting p constitute the candidate set of p .
- II) k -nearest neighbour selection (kNNS): The candidate set of p consists of the k red points that have the smallest Euclidean distance from p .
- III) Orthogonal range searching selection (ORSS): The candidate set of p contains those red points that fall inside an axis-parallel square of dimension w centered around p .

Each of the above queries can be executed fast using known geometric data structures (de Berg et al. 2008). Considering small candidate sets does not guarantee the computation of the optimal matching or, in theory, even the computation of a matching

that includes all input features. Yet, as we show later, in practice the presented techniques allow almost always for computing an optimal or near-optimal solution.

3. Experiments

We conducted matching experiments using the techniques described above. Given two 2D point sets and using each time one of the techniques we construct the candidate sets of the blue features. We then compute the Euclidean distance between each blue point and each red point in its feature set and then apply the algorithm of Li and Goodchild to compute a minimum weight matching. For method II) we considered $k=5$ and for method III) we considered a window of area equal to 1% of the total terrain area. For the implementation of the LP solver and for the data structures that support the technique queries we used CGAL.

As an input data set we used 2D points extracted by a USGS (2010) raster terrain model. The extracted points are the 2D projections of local minima of the terrain. We added noise to the point coordinates to generate a second point set and thus a matching problem instance. The size of each input point set in the experiment instances were of the range {10,20,...,300}. For each experiment instance and each technique we measured the total size of the constructed candidate sets, the running time of the algorithm and the percentage of features that were matched correctly. As a reference, we conducted the experiments also for the case where each candidate set contains all red points (ALL). The results of these experiments are summarized in fig. 1-3.

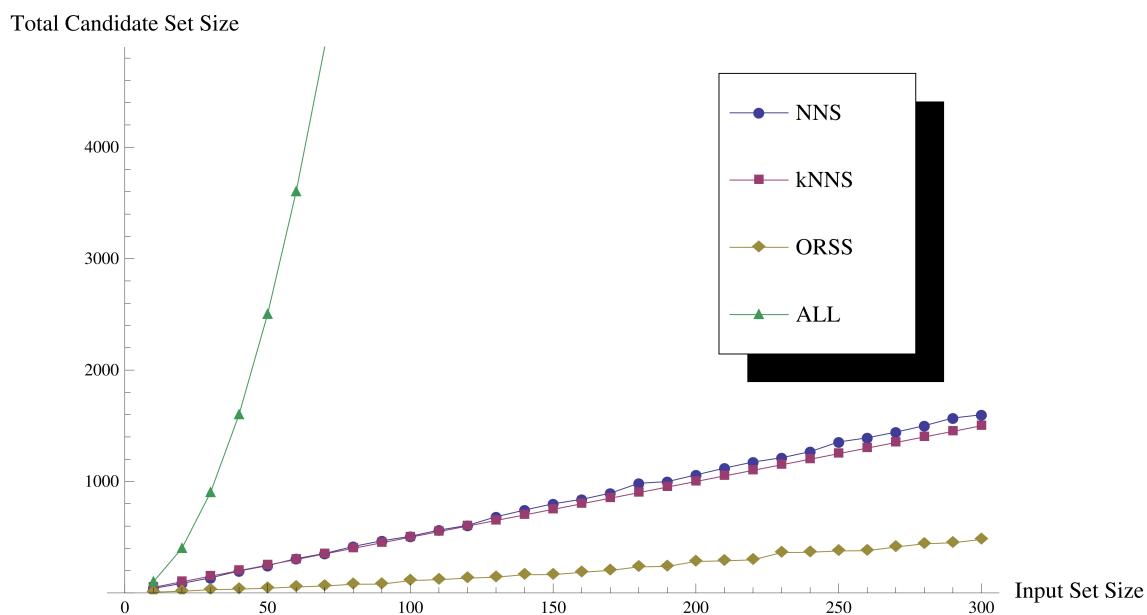


Figure 1. Total size of candidate sets.

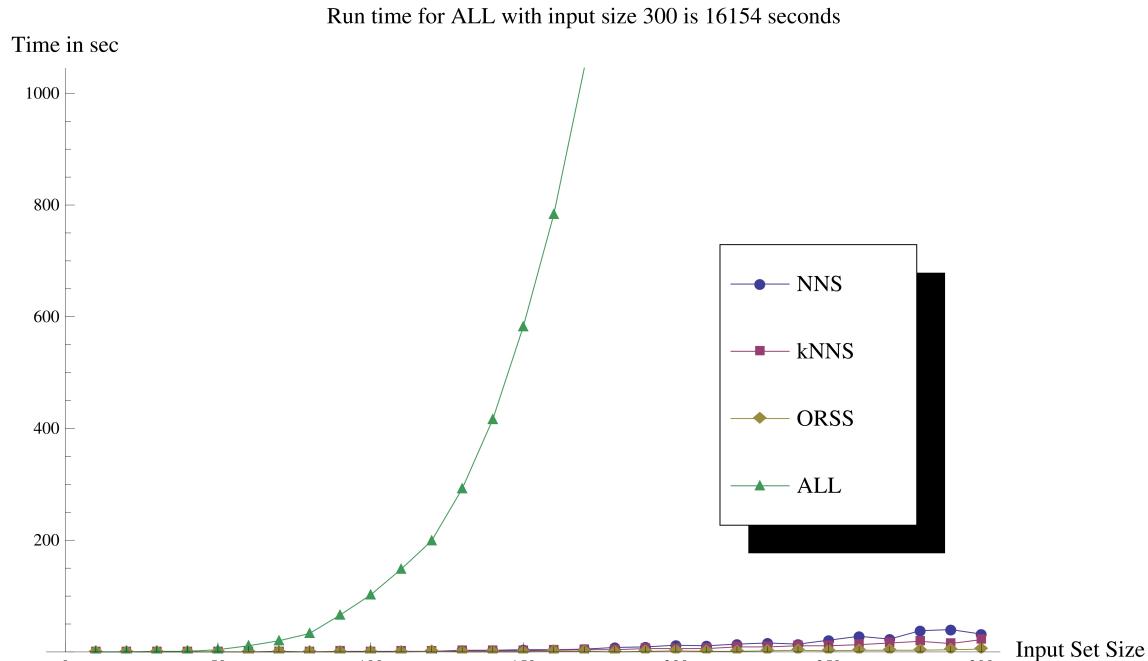


Figure 2. Running time of the algorithm.

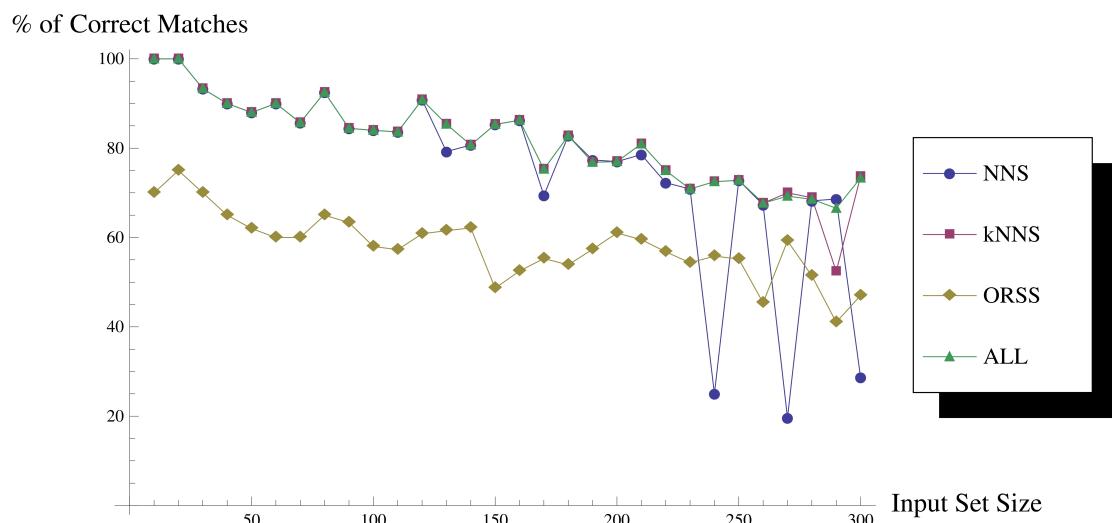


Figure 3. Percentage of correctly matched points over all points.

As indicated by the measurements, the algorithm runs incredibly faster when using one of the geometric selection techniques. The geometric technique that leads to a higher quality matching is kNNS. In fact kNNS leads to matchings which are as good as the ones resulting from ALL. The NNS technique performs well but there appear also

instances where the matching quality is low, around 25%. The ORSS method does not perform well for the chosen window size. For a larger window the performance improves but in any case the total candidate set size seems to grow superlinearly as the input point density increases.

4. Future Work

It will be interesting to evaluate how the proposed selection techniques perform for different kinds of geospatial data and also to examine algorithms for matching features of more than two sets.

5. References

- De Berg M, Otfried C, van Kreveld M and Overmars M, 2008, *Computational Geometry: Algorithms and Applications*, 3rd Edition, Springer Verlag, Berlin, Germany.
- CGAL, 2010, *Computational Geometry Algorithms Library*, <http://www.cgal.org>
- Ferguson R, 1955, Linear Programming. *American Machinist*, 121-186.
- Kim J, Yu K., Heo J. and Lee W, 2010, A new method for matching objects in two different geospatial datasets based on the geographic context, *Journal of Computers and Geosciences*, 36(9):1115-1122.
- Li L and Goodchild M, 2010, Optimized Feature Matching in Conflation, *Proceedings of sixth international conference on Geographic Information Science*.
- USGS, 2010, *US Geological Survey web server*, <http://dds.cr.usgs.gov/pub/data/DEM/250/>.

Author Index

Author	Session	Author	Session
Adnan, Muhammad	1A	Delavar, Mahmoud Reza	3B
Afonso, Fábio	3A	Delmelle, Eric	4
Al-Ahmadi, Khalid	4	Derungs, Curdin	3B
Alavipanah, Seyed Kazem	4	Downs, Joni	3A, 7B
Al-Zahrani, A.	4	Ellul, Claire	2A, 5B
Andris, Clio	7B	Elsner, James	9B
Antunes , António P.	1B	Emmonds, Andy	2A
Arifin, Md. Shamsul	8A	Ferreira, Joseph	7B
Attard, Maria	3A	Fisher, Pete	4
Augustijn-Beckrs, Ellen-Wien	2A	Foley, Peter	4
Bajanda, Therese	3A	Fotheringham, A. Stewart	7A
Barbosa, Fernanda	3A	Fritz, Steffen	5B
Benenson, Itzhak	8B, 9A	Gao, Wenxiu	4, 5B
Bhaduri, Budhendra	4	Goli, Ali	4
Birkin, Mark	1A, 2A	Gong, Jianya	5B
Blank-Baron, Eilon	9A	Goodarzimehr, Saeed	4
Bolbol, Adel	6A	Goovaerts, Pierre	2B
Brimicombe, Allan	8A	Grêt-Regamey, Adrienne	7A
Brunsdon, Chris	7A, 8A	Grindal, Scott	9A
Cagdas, Gulen	4	Guan, Xuefeng	5B
Caha, Jan	3B	Hadley, Stan	4
Charlton, Martin	7A	Hagen-Zanker, Alex	2B
Chaudhuri, Gargi	8B	Haklay, Muki	2A, 5B
Cheng, Tao	2A, 5A, 6A	Harland, Kirk	1A
Cheshire, James	1A	Harris, Paul	7A
Chetcuti Zammit, Luana	3A	Haworth, James	5A, 6A
Chow, Andy	5A	Hebblewhite, Mark	9A
Chrisman, Nicholas	3B	Heppenstall, Alison	1A, 4
Chukwusa, Emeka	8A	Heydecker, Benjamin	5A
Cladera, Josep R.	1B	Hiraga, Yuko	5A
Clarke, Keith	1B	Horner, Mark W.	3A, 3B, 7B
Clarke-Lauer, M.	1B	Huisman, Otto	8A
Comber, Alexis	4, 7A, 8A	Ingram, Ben	2B
Conley , Jamison	9B	Jacob, Ricky	7B
Corcoran, Padraig	6B, 7B	Jafarbeglou, Mansour	4
Crespo, Ricardo	7A	Janoska, Zbyněk	4
Crooks, Andrew	4, 9A	Jiang, Bin	5A
Cui, Xiahui	4	Jin, Ying	2B
Dančák, Martin	4	Kerry, Ruth	2B

Author	Session	Author	Session
Khakbaz, Bahere	4	Pinto, Nuno Norte	1B
Khmag, Abdulhakim	4	Purves, Ross	3B
Khmag, Abdullhakim	4	Qin, Cheng-Zhi	6B
Khosravi, Farzam	4	Qiu, Weili	6B
Kolyaie, Samira	2B	Radburn, Robert	7A
Koukoletsos, Thomas	5B	Rey, Sergio	3A
Kyoung Kim, Hoe	4	Rieser, Verena	1B
Laffan, Shawn	6B	Roan, Tyng-Rong	2A
Li, Yang	8A	Robertson, Collin	3B, 9B
Lin, Jingyi	5A	Robinson, Derek T.	1B
Liu, Cheng	4	Rodrigues, Armanda	3A
Liu, Xintao	5A	Rosenthal, Amit	8B
Longley, Paul	1A	Rounsevell, Mark	1B
Lu, Binbin	7A	Salazar, Josue	4
Lu, Yan-Jun	6B	Samany, Najmeh	3B
Mack, Elizabeth	3A	Santos-Hernandez, Jennifer	4
Malek, Mohammad Reza	3B	Sariyildiz, Sevil	4
Malizia, Nicholas	3A	Sarmasti, Nader	4
Malleson, Nick	2A	Scerri, Kenneth	3A
Manley, Ed	2A	Scerri, Mark	3A
Marceau, Danielle	9A	Scheitlin, Kelsey	9B
Marek, Lukáš	4	Schill, Christian	5B
Martens, Karel	8B	Schryver, Jack	4
Matinfar, Hamidreza	4	See, Linda	4, 5B
Mccallum, Ian	5B	Semeniuk, Christina	9A
Medina, Richard	4	Shawe-Taylor, John	6A
Mesev, Victor	9B	Singleton, Alex	1A
Mooney, Peter	6B, 7B	Smit, Izak	2B
Murray-Rust, Dave	1B	Smith, Dianna	1A
Musiani, Marco	9A	Sokmenoglu, Ahu	4
Nara, Atsushi	7B	Stepinski, Tomasz	4
Nelson, Trisalyn	9B	Stewart, Robert	9B
Neuhaus, Fabian	4	Takizawa, Atsushi	6A
Nutaro, James	4	Tan, Xiaojin	8B
Obersteiner, Michael	5B	Torrens, Paul	7B
Osaragi, Toshihiro	5A	Tsirogiannis, Constantinos	9B
Osei, Frank	2A	Tuček, Pavel	3B, 4
Pászto, Vít	4	Turner, Andrew	6A
Pechanec, Vilem	3B	Turton, Ian	6A
Perger, Christoph	5B	Useya, Juliana	2A
Preston, Benjamin	4	Vallejo, Marta	9A

Author	Session
Waizman, Gennady	9A
Waldvogel, Bettina	3B
Wandl-Vogt, Eveline	4
Wang, Jiaqiu	5A
Warsop, Thomas	6B
White, Denis	4
Winstanley, Adam	6B, 7B
Wise, Sarah	4, 9A
Wu, Huayi	4, 5B
Xie, Yichun	8B
Yaghooti, Marjan	2B
Yang, Wenbai	7A
Ying, Fangli	6B
Yue, Peng	5B
Zhu, A-Xing	6B
Zhu, Xinyan	5B
Zurita-Milla, Raul	2A, 8A