UCL

Proceedings of the International Symposium on Spatial-Temporal Analysis and Data Mining

University College London, UK
18th – 19th July 2011

Ordnance Survey®

esri UK

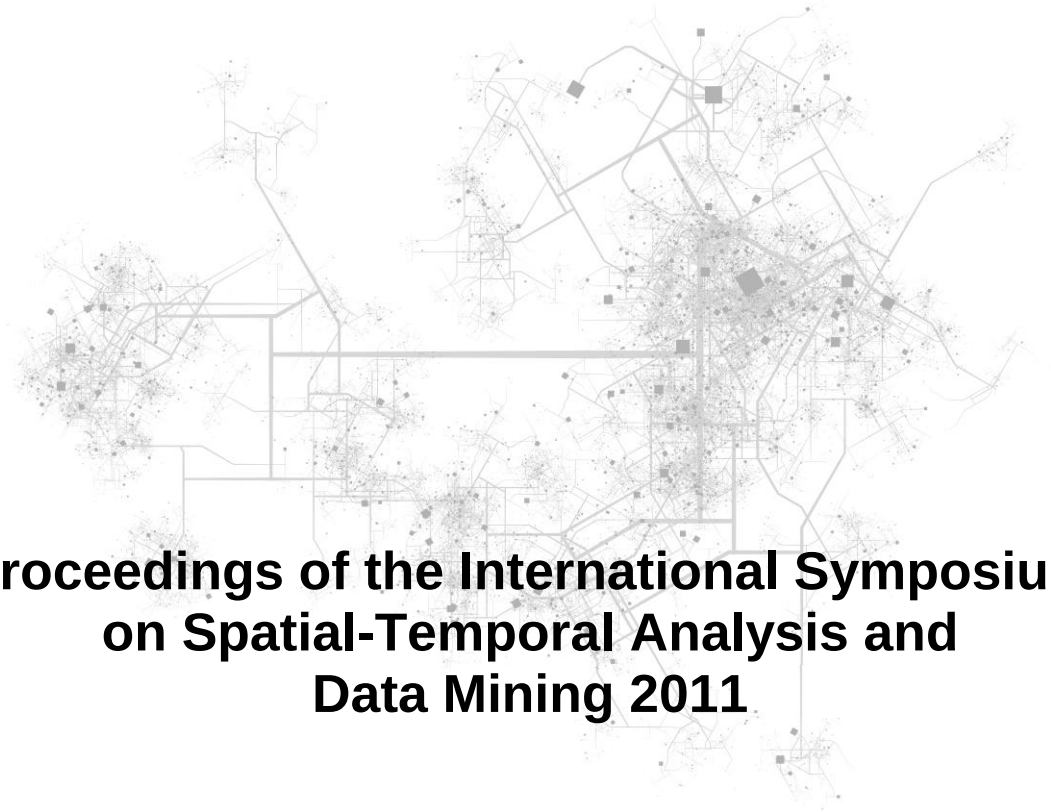
WILEY-BLACKWELL

Pion

Taylor & Francis
Taylor & Francis Group

isprs
a century of information from imagery
WG II/3

STANDARD
Spatio-Temporal Analysis of Network Data and Route Dynamics
http://standard.cege.ucl.ac.uk



Proceedings of the International Symposium on Spatial-Temporal Analysis and Data Mining 2011

University College London
18th – 19th July 2011

Editors: Tao Cheng
Paul Longley
Claire Ellul
Andy Chow

WG II/3

WG II/3

Proceedings of the International Symposium on Spatial-Temporal Analysis and Data Mining 2011

University College London
18th – 19th July 2011

July 2011, University College London

Programme Committee

Chris Brunsdon	University of Leicester, UK
Tao Cheng	University College London, UK
Andy Chow	University College London, UK
Christophe Claramunt	Naval Academy Research Institute France, France
Andrew Frank	University of Twentie, Austria
Diansheng Guo	University of South Carolina, USA
Mike Goodchild	University of California Santa Barbara, USA
Robert Haining	University of Cambridge, UK
Bo Huang	The Chinese University of Hong Kong, Hong Kong
Andrew Hudson-Smith	University College London, UK
Bin Jiang	University of Gävle,Sweden
Mikhail Kanevski	University of Lausanne, Switzerland
Brian Lees	Australian National University, Australia
Yee Leung	The Chinese University of Hong Kong, Hong Kong
Yaolin Liu	Wuhan University, China
Antonion Marvuglia	Public Research Centre, Luxembourg
Jeremy Mennie	Temple University, USA
Tomoki Nakaya	Ritsumeikan University, Japan
Tao Pei	Chinese Academy of Sciences, China
Alexei Pozdnoukhov	National University of Ireland (Maynooth), Ireland
Abdülvahit Torun	University of Minnesota, USA
Shuliang Wang	Wuhan University, China
Jo Wood	City University, UK
Xiaobai Yao	University of Georgia, USA
May Yuan	University of Oklahoma, USA

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

Under the auspices of ISPRS Commission II WG3, the motivation for the International Symposium on Spatial-Temporal Analysis and Data Mining is to exchange the latest ideas about the deployment of advanced technologies to methods of spatial analysis, spatial-temporal data modelling and data mining. It has been jointly organised by the Department of Civil, Environmental and Geomatic Engineering, and the Department of Geography, University College London on July 18th – 19th, 2011. Over 60 scholars from 21 countries and regions have registered for this symposium – from Australia, Austria, Canada, China, Finland, Germany, India, Italy, Ireland, Israel, Japan, Malta, Portugal, The Netherlands, Spain, Slovenia, Sweden, Switzerland, Turkey, the UK and the USA.

These proceedings assemble together the 24 papers accepted for presentation at the Symposium. The meeting is organised into seven sessions, demonstrating the breadth of the subject. It includes four sessions on methods and algorithms, including "*Space-Time Regression*", "*Space-Time Clustering*", "*Space-Time Prediction*", and "*Fuzzy Approaches*". Two sessions contribute to applications: "*Mobility & Location*" and "*Health*". Finally, one session focuses upon digital infrastructure - "*Service & Management of Space-Time Data*".

Four world-leading scholars kindly offered to give keynotes in the Symposium. They cover '*What is Special about Mining Spatial and Spatio-Temporal Datasets?*' (Shashi Shekhar, University of Minnesota, USA); '*Graphical and Analytical Techniques in Space-Time Analysis*' (Chris Brunsdon, University of Liverpool, UK) and '*Why Topology and Scaling Matter in Geospatial Analysis*' (Bin Jiang, University of Gävle, Sweden). Another keynote session - '*Visualising Space-Time Dynamics: Graphs and Maps, Plots and Clocks*' is contributed by Mike Batty, CASA, UCL, as a joint event with the 11th International Conference on Geocomputation, which runs immediately after the Symposium at same location. A joint reception held after this keynote session is intended to foster closer dialogue between participants to this Symposium and the Geocomputation Conference .

We are grateful to all the members of the Programme Committee for laying the foundations to a successful Symposium. I am indebted to the reviewers of the papers contained in this volume. The Symposium has also benefitted from sponsorship from the Ordnance Survey (GB), Esri UK, John Wiley & Sons, Taylor & Francis, Pion, and the STANDARD Project. Special thanks go to the STANDARD team members: Berk Anbaroglu, Adel Bolbol, James Haworth, Ed Manley, Ioannis Tsapakis, Garavig Tanaksaranond, Artemis Skarlatidou and Jiaqiu Wang, for all of their hard work for this event. The help from Lee Philips, Richard Sharp, and 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

A STUDY ON PARAMETER CALIBRATION OF STARIMA MODEL

J. Wang, T Cheng, B. G. Heydecker, A.H.F Chow

Dept. of Civil, Environmental, and Geomatic Engineering, University College London, UK,
w.jiaqiu@ucl.ac.uk

Commission II, WG II/3

KEY WORDS: STARIMA, ARIMA, Parameter Calibration, Hannan-Rissanen Algorithm

ABSTRACT:

Space-Time Autoregressive Integrated Moving Average (ST-ARIMA) model family, one of most popular models in space-time integration modelling field, has widely used in space-time series analysis and modelling. Although it is common knowledge that the small sample leads to biased estimation, not much details are readily available to tell the essential length of the series in order to derive an unbiased or an accepted model. This research attempts to investigate how the accuracy of parameter estimation is changed with the series length by comparing two parameter calibration algorithms. This can help us to understand the computation and modelling procedures of space-time series and give us useful guidance when modelling space-time series using STARIMA.

1. INTRODUCTION

Geographic Information Science (GIS) have traditionally focused on geospatial and not temporal referencing of data. Recently GIS research has focused on spatio-temporal analysis capable of providing the foundation for a temporal GIS (Griffth, 2010). Space-Time Autoregressive Integrated Moving Average (ST-ARIMA) model is a useful tool to model space-time autocorrelation. It is a spatial extension of Autoregressive Integrated Moving Average (ARIMA). In reality, ST-ARIMA and ARIMA can apply the same calibration algorithm (i.e. Hannan-Rissanen algorithm) to estimate model parameters due to the fact that ST-ARIMA model origins from ARIMA model. Although parameter estimation of ARIMA model has been discussed in previous studies, parameter estimators are biased when the samples are small (Kim, 2003). How the accuracy of parameter estimation is affected by series length has not been discussed in detail. No readily guideline is available to tell the essential length of the series in order to derive an unbiased or an accepted model.

In the study, an experiment is carried out to explore the relationship between the accuracy of parameter and series length. Hannan-Rissanen (HR) algorithm is applied to parameter estimation of ARIMA model, which is compared with the Durbin-Levinson calibration method used in the software R. It is found that (1) it is possible that parameter estimators of ARIMA model will be unbiased when length of series is enough. (2) Durbin-Levinson algorithm is superior to HR algorithm when sizes of samples are smaller. These conclusions can give us some guidance when we model space-time series using STARIMA.

The remaining parts of the paper are constructed as follows. We first briefly introduce Hannan-Rissanen algorithm in Section 2. Then, an experiment is conducted to validate parameter recovery of Hannan-Rissanen algorithm using simulated time series in Section 3. In Section 4, experiment results are compared with Durbin-Levinson algorithm of time series analysis in free software R in order to further verify the conclusion of Section 3. Finally, conclusion is drawn to summarize our major findings.

2. HANNAN-RISSANEN (HR) ALGORITHM

Parameter estimation of STARIMA model is to minimize the following sum of squared error function.

$$S(\hat{\beta}) = \sum_{t=1}^{T}\left(\hat{z}(t) - \sum_{k=1}^{p}\sum_{h=0}^{m_k}\hat{\boldsymbol{\varphi}}_{kh}\boldsymbol{W}^{(h)}\boldsymbol{z}(t-k) + \sum_{l=1}^{q}\sum_{h=0}^{m_k}\hat{\boldsymbol{\theta}}_{lh}\boldsymbol{W}^{(h)}\boldsymbol{\varepsilon}(t-l)\right)^2 \quad (1)$$

where $z(t)$ is the observation vector at time t; T is the number of observations in time; k is temporal lag in autoregressive term of STARIMA model, and l is the temporal lag in moving average term of STARIMA model; h is spatial order; w is spatial weighted matrix at spatial order h; $\boldsymbol{\varepsilon}(t)$ is the random error vector at time, and $\hat{\beta} = (\hat{\boldsymbol{\varphi}}_1, \cdots \hat{\boldsymbol{\varphi}}_p, \hat{\boldsymbol{\theta}}_1, \cdots, \hat{\boldsymbol{\theta}}_q)$. Here if spatial lag m_k is equal to 0, Equation (1) becomes an ARIMA model without spatial influence. Thus, we can view ARIMA as special case of STARIMA model. Minimization of Equation (1) is a nonlinear optimization problem with iterations because random noise term (or moving average term) cannot be estimated by linear least square algorithm. Indeed, it is important to furnish an appropriate starting point to $\boldsymbol{\varepsilon}(t)$ because it is found that the optimization process either converges to a local optimum or does not converge at all due to inappropriate initial value of $\boldsymbol{\varepsilon}(t)$.

Thus, to avoid reaching a local optimum and to reduce the number of iterations, a pre-estimates process in HR algorithm is implemented to estimate random noise term (or moving average term) in Equation (1). Then, Equation (1) can be calibrated using linear least square algorithm. HR algorithm has three iterative steps

Firstly, a high order STAR model is fitted to the data using the time Yule-Walker equation. Then, we have the following approximate model.

$$\hat{z}^*(t) = \sum_{k=1}^{n}\sum_{h=0}^{m_k} \hat{\eta}_{kh} \boldsymbol{W}^{(h)} z^*(t-k) + \boldsymbol{\varepsilon}^*(t), \quad (2)$$

where $\{\hat{\eta}_{kh} | k = 1,\cdots,n, h = 0,\cdots,m_k\}$ are the Yule-Walker estimates. The estimated random noise vectors in Equation (2) can be computed as

$$\hat{\boldsymbol{\varepsilon}}(t) \equiv \boldsymbol{\varepsilon}^*(t) = \hat{z}^*(t) - \sum_{k=1}^{n}\sum_{h=0}^{m_k} \hat{\eta}_{kh} \boldsymbol{W}^{(h)} z^*(t-k), \quad (3)$$

Then, once the estimated random noise vector $\hat{\boldsymbol{\varepsilon}}(t), t = m+1,\cdots,T$ have been found from equation (3), pre-estimation of the model parameters, $\hat{\alpha}_i = (\hat{\varphi}', \hat{\upsilon}')$ are determined by least square linear regression of $\mathbf{z}_i(t)$

$$\begin{bmatrix} \mathbf{z}(t-1), \mathbf{z}(t-2), \cdots, \mathbf{z}(t-p), \\ \hat{\varepsilon}(t-1), \hat{\varepsilon}(t-2), \cdots, \hat{\varepsilon}(t-q), \\ t = m+1, \cdots, T \end{bmatrix},$$

where $m = max\{(n+p+1), (n+q+1)\}$. By minimizing the sun of square errors

$$S(\hat{\alpha}) = \sum_{t=m+1}^{T} \left(z(t) - \sum_{k=1}^{n}\sum_{h=0}^{m_k} \hat{\phi}_{kh} \boldsymbol{W}^{(h)} z(t-k) + \sum_{l=1}^{q}\sum_{h=0}^{m_k} \hat{\upsilon}_{lh} \boldsymbol{W}^{(h)} \varepsilon(t-l) \right)^2 \quad (4)$$

We can estimate using following equation

$$\hat{\alpha} = (X'X)^{-1} X'Z$$

where $\hat{\alpha} = [\hat{\phi}_{10}, \hat{\phi}_{11}, \cdots, \hat{\phi}_{kh}, \hat{\upsilon}_{10}, \hat{\upsilon}_{11}, \cdots, \hat{\upsilon}_{lh}]$,

$Z = [z(m+1), z(m+2), \cdots, z(T)]'$,

Finally, using the pre-estimates $\hat{\alpha}$ as the initial point for the optimization of Equation (1), the final least square estimates of the model, $\hat{\boldsymbol{\varphi}}_{kh}$ and $\hat{\boldsymbol{\theta}}_{lh}$, can be obtained by finding the best $\hat{\beta} = [\hat{\boldsymbol{\varphi}}_{10}, \hat{\boldsymbol{\varphi}}_{11}, \cdots \hat{\boldsymbol{\varphi}}_{pm}, \hat{\boldsymbol{\theta}}_{10}, \hat{\boldsymbol{\theta}}_{11}, \cdots, \hat{\boldsymbol{\theta}}_{qm}]$ to minimize Equation (1).

3. VALIDATION OF PARAMETER RECOVERY WITH HANNAN-RISSANEN ALGORITHM

To test the parameter estimation of HR algorithm, an experiment is designed to look whether HR algorithm can recover parameters of STARIMA model. For simplicity, here we suppose Equation (1) have no spatial influence, we therefore can make use of single time series to validate HR algorithm. The procedure of validation can be divided into three steps: 1) Set up an arbitrary state equation with known parameters (phi, theta); 2) Generate random time series with the state equations; 3) Recover the ARIMA model parameters with HR algorithm (HR-ARIMA).

3.1 Set up state equation with parameters (phi, theta)

Parameters of phi ($\hat{\boldsymbol{\varphi}}$) and theta ($\hat{\boldsymbol{\theta}}$), are arbitrary selected with the range phi from 0 to 3, theta from 0 to 3 as well. Table 1 summarizes the lists of parameters, which constitute the various combination of state equation.

Table 1. Summary of parameters (phi, theta) of state equation

(p,q)	*phi*			*theta*		
	1	**2**	**3**	**1**	**2**	**3**
(0, 1)	-	-	-	0.556	-	-
(1, 1)	-0.165	-	-	0.556	-	-
(0, 2)	-	-	-	0.556	-0.096	-
(1, 2)	-0.165	-	-	0.556	-0.096	-
(2, 1)	-0.165	-0.111	-	0.556		-
(2, 2)	-0.165	-0.111	-	0.556	-0.096	-
(0, 3)	-	-	-	0.556	-0.096	-0.033
(1,3)	-0.165	-	-	0.556	-0.096	-0.033
(2,3)	-0.165	-0.111	-	0.556	-0.096	-0.033
(3,3)	-0.165	-0.111	-0.116	0.556	-	-
(3,2)	-0.165	-0.111	-0.116	0.556	-0.096	-
(3,3)	-0.165	-0.111	-0.116	0.556	-0.096	-0.033

3.2 Generate simulated time series

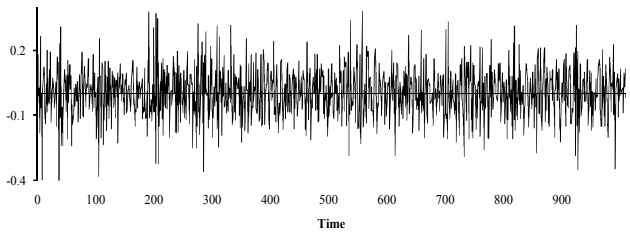


Figure 1. Profiles of series 1,000

Simulated time series with length of 1000, 10000, 100000, 500,000, and 1,000,000) are produced using state equations consisting of parameters of Table 1. All simulation series have the same statistic characteristics such as mean values (mean=0) and standard deviation (std = 0.1), but they have different length, which vary from 1,000 to 1,000,000 so as that we can explore the relationship between recovered parameters and series lengths. Figure 1 shows the profiles of simulated time series 1,000. ACF graph of series 1,000 is displayed in Figure 2, showing the time series is a stationary series.

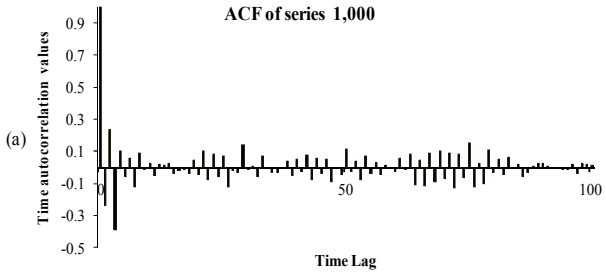


Figure 2. ACF graph of series 1,000

3.3 Parameter recovery using ARIMA with HR algorithm (HR-ARIMA)

Parameters are calibrated using HR algorithm for time series with different length (from 1,000 - 1,000,000). Taking ARIMA (3, 3) model as example, Table 2 summarizes the recovered

parameters of HR-ARIMA (3, 3) model. Figure 3 displays the parameter recovery changes of HR-ARIMA (3, 3) model as series length increase, where dot line is reference line indicating raw value of parameter *phi* or *theta*.

Table 2. Summary of recovered parameters with HR-ARIMA(3, 3) model

Parameters / Length	*phi* 1	*phi* 2	*phi* 3
	-0.165	**-0.111**	**-0.116**
1,000	-0.3029	-0.0896	-0.1280
10,000	-0.1414	-0.1034	-0.1294
100,000	-0.1669	-0.1078	-0.1215
500,000	-0.1634	-0.1079	-0.1135
1,000,000	-0.1652	-0.1115	-0.1156
Parameters / Length	***theta***		
	1	**2**	**3**
	0.556	**-0.096**	**-0.033**
1,000	0.5074	-0.1024	-0.1426
10,000	0.5597	-0.1018	-0.0325
100,000	0.5539	-0.0919	-0.0366
500,000	0.5591	-0.0951	-0.0318
1,000,000	0.5559	-0.0959	-0.0311

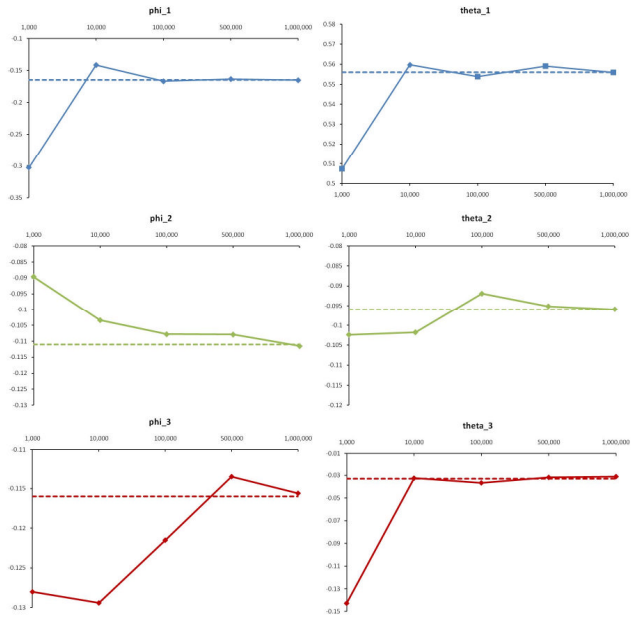


Figure 3. Parameter recovery curves. X-axis represents sequence length, and Y-axis represents the parameter values recovered. Dot line represents raw value of parameter.

As seen in Table 2, accuracy of parameter recovery improves as series length increases. From the Figure 3, we found in surprise that parameter curve is infinitely close to reference line when sequence length reaches around one million. Other HR-ARIMA (p, q) models in Table 1 also demonstrate the same findings with HR-ARIMA (3, 3) model. This indicates that longer time series is necessary in order to get unbiased estimation when estimating ARIMA model.

4. COMPARISONS WITH INNOVATIVE ALGORITHM

In order to further verify the conclusion of Section 3, R time series software package is applied to validate the results of Section 3. R employs Durbin-Levinson algorithms to calibrate parameters of ARIMA model (McLeod et al, 2007). The summary of parameters calibrated by R software and a comparison with HR algorithm is shown in Table 1. Then, the same datasets with Section 3 is applied to calibrate model parameters. Table 3 summarized results of parameter recovery of Hannan-Rissanen and Innovative algorithms.

Table 3. Comparisons of recovered parameters with different calibration algorithms

Length	Parameters / Algorithms	*phi* 1	*phi* 2	*phi* 3
		-0.165	**-0.111**	**-0.116**
1,000	**HR**	-0.3029	-0.0896	-0.1280
	Durbin-Levinson	-0.2485	-0.1202	-0.1270
10,000	**HR**	-0.1414	-0.1034	-0.1294
	Durbin-Levinson	-0.1492	-0.1027	-0.1322
100,000	**HR**	-0.1669	-0.1078	-0.1215
	Durbin-Levinson	-0.1674	-0.1085	-0.1205
500,000	**HR**	-0.1634	-0.1079	-0.1135
	Durbin-Levinson	-0.1622	-0.1057	-0.1137
1000,000	**HR**	-0.1652	-0.1115	-0.1156
	Durbin-Levinson	-0.1656	-0.1092	-0.1163
Length	**Parameters / Algorithms**	***theta***		
		1	**2**	**3**
		0.556	**0.556**	**0.556**
1,000	**HR**	0.5074	0.5074	0.5074
	Durbin-Levinson	0.5451	0.5451	0.5451
10,000	**HR**	0.5597	0.5597	0.5597
	Durbin-Levinson	0.5563	0.5563	0.5563
100,000	**HR**	0.5539	0.5539	0.5539
	Durbin-Levinson	0.5551	0.5551	0.5551
500,000	**HR**	0.5591	0.5591	0.5591
	Durbin-Levinson	0.5588	0.5588	0.5588
1000,000	**HR**	0.5559	0.5559	0.5559
	Durbin-Levinson	0.5567	0.5567	0.5567

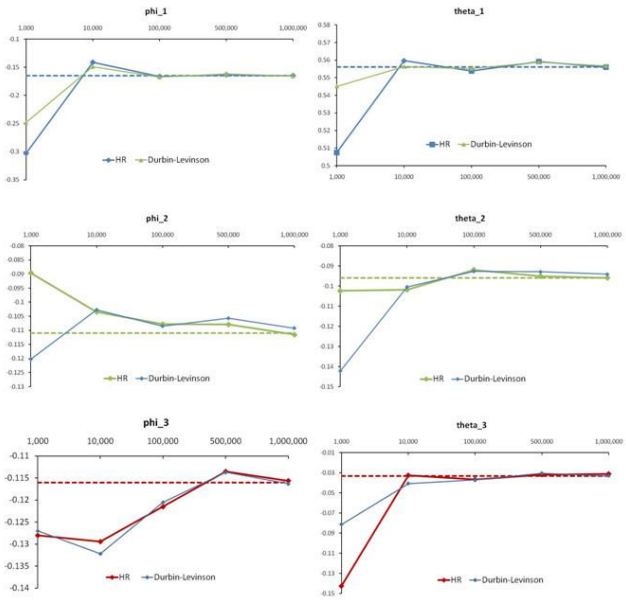


Figure 4. Parameter recovery curves of different calibration algorithms with same model ARIMA (3, 3).

Figure 4 shows the changes of parameter recovery as increase of sequence length. As seen in Figure 4, two calibration algorithms tend to resemble with the increase of sequence length. Besides, from Figure 4, we can see that Durbin-Levinson algorithm is superior to HR algorithm when sizes of series are smaller. This suggests that when samples of time series are shorter Durbin-Levinson calibration algorithm can be adopted.

5. CONCLUSIONS

In the paper, we discuss Hannan-Rissanen (HR) calibration algorithm and an experiment is conducted to check the correctness of the algorithm using simulated time series. Then, innovative calibration algorithm is employed to compare the results of parameter recovery. It is found to our surprising that (1) parameters of ARIMA model can be recovered when length of series is enough. In the study, the length reaches around one million. This indicates that longer time series is necessary in order to get unbiased estimation; (2) Durbin-Levinson algorithm is superior to HR algorithm when sizes of samples are smaller. These conclusions can give us some guidance when we model space-time series using STARIMA.

REFERENCES

Box, G. E. P., G. M. Jenkins, and G. Reinsel. 1994. *Time series analysis, forecasting and control, 3rd edition*, Englewood Cliffs, Prentice Hall.

Brockwell, B.J. and Davis R.A. 1897. Time Series: Theory and Methods, Springer-Verlag.

Durbin, J. 1960. The fitting of time-Series models. *International Statistical Institute*, 28(3), 233-244.

Griffith, D. A. 2010. Modeling spatio-temporal relationships: retrospect and prospect. *Journal of Geographical Systems*, 12(2), 111-123.

Hannan, E. J, Rissankn, J. 1982. Recursive estimation of mixed autoregressive-moving average order, *Biometrika*, 69, 81-94.

McLeod, AI., Yu, H., Krougly, Z. 2007. Algorithms for linear time series analysis: with R package, *Journal of statistical software*, 23(5), 1–26.

Kim, H.J. 2003. Forecasting autoregressive time series with bias-corrected parameter estimators. *International Journal of Forecasting*, 19, 493-502.

ACKNOWLEDGEMENTS

This research is sponsored by UK EPSRC (EP/G023212/1) and data provided by Transport for London (TfL) is highly appreciated.

Spatio-temporal regression models for deforestation in the Brazilian Amazon

Giovana M. de Espindola[1], Edzer Pebesma[2,3], Gilberto Câmara[1]
[1] INPE, Brazil [2] Institute for Geoinformatics, University of Muenster
[3] 52°North GmbH, Muenster

1 Introduction

As one of the largest tropical forests in the world, the Brazilian Amazon is an area where deforestation affects environmental themes such as biodiversity and greenhouse gas emission with global proportions. After a long period of increase, deforestation in the Brazilian Amazon has sharply decreased over the past five years. Following Aguiar et al. (2007), in this study we try to explain the spatio-temporal changes of deforestation in the Brazilian Amazon by relating yearly data from 2002-2007 to a number of explanatory variables. We do so by considering the yearly fraction of deforestation for 25 km × 25 km cells, and by using spatial multiple regression models that incorporate autoregressive components in space and in time, and predictors that vary over space and time as well. The goal is to understand the changes in deforestation, and ultimately to understand the effect of control actions and to obtain process knowledge needed for land change models that are developed to evaluate future actions.

2 Methods

Regression modelling approximates a dependent variable with n observations $y = (y_1, ..., y_n)'$ to a set of p independent variables $x_j = (x_{1j}, ..., x_{nj})'$ by a linear function,

$$y = \sum_{j=1}^{p} \beta_j x_j + e = X\beta + e$$

with X the design matrix, having x_{ij} on row i and column j. The regression coefficient vector β is typically estimated by minimizing the residual sum of squares, $e'e$.

Spatial linear regression models can be built by including a spatial autocorrelation effect, for a single observation y_i written as

$$y_i = \lambda \frac{1}{m} \sum_{k=1}^{m} y_k + \sum_{j=1}^{p} \beta_{ij} x_j + e_j$$

where m neighbours are addressed, and λ expresses the strength of the autoregressive effect of the neighbours. For the full set of observations this can be written as

$$y = \lambda W y + X\beta + e$$

variable	meaning
`Change pr. areas`	change in protected areas (grid cell fraction) between t and $t-1$
`Change in cattle`	change in cattle (normalized to range from 0 to 1), between t and $t-1$
`Change soy bean`	change in soy bean area (grid cell fraction) between t and $t-1$
`Change sugar cane`	change in sugar cane area (grid cell fraction) between t and $t-1$
`Time lagged defor`	deforestation at time $t-1$ (autoregressive time effect)

Table 1: Explanatory variables used in the spatio-temporal regression model

where W is the sparse, row-standardized matrix that for each of the observations indicates whether it is a neighbour. In this study we used the queen neighbours, meaning the 8 cells adjacent to each grid cell, or less in case of boundary cells.

For a spatio-temporal regression model, where we will denote $y_{[t]} = (y_{1,t}, ..., y_{n,t})$ as the observation in grid cell i and time step $t \in \{1, ..., m\}$, as a first step one can in addition to a spatial autoregressive effect take a temporally lagged observation $y_{[t-1]}$ into account, as in

$$y_{[t]} = \lambda W y + \gamma y_{[t-1]} + X\beta + e, \quad t = 2, ..., m \tag{1}$$

Regressions were carried out with the R function `spautolm` in R package `spdep` (Bivand et al., 2008). This function provides maximum likelihood estimation of β and λ, but does not simultaneously estimate λ, γ *and* β using maximum likelihood. One solution to this would be to define neighbours in space *and* time, combined with a weighting factor that defines how neighbouring in space compares to neighbouring in time, in terms of weights, would be a minimum requirement for this to make sense. The solution chosen here was to add the temporal factor to the fixed effects $X\beta$, effectively leading to an more least squares oriented solution.

3 Results

Maps of yearly deforestation for the period 2002-2007 are shown in figure 1. The explanatory variables addressed are, for each grid cell and time step t defined in table 1:

The summary of a spatio-temporal autoregressive regression model is given below:

```
Residuals:
        Min          1Q      Median          3Q         Max
-0.08238071 -0.00079415 -0.00050981 -0.00038770  0.18402406
```

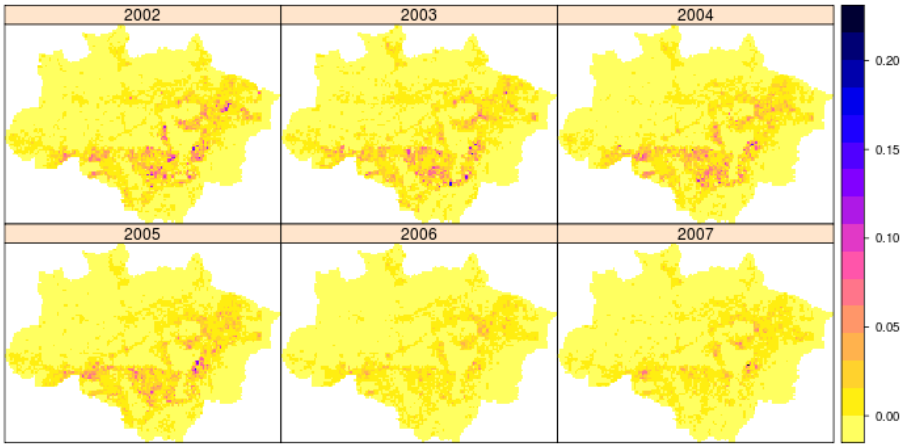


Figure 1: Yearly deforestation rates in the Brazilian Amazon, per year, as fractions of 25 km × 25 km cells, over the period 2002-2007

```
Coefficients:
                    Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)       0.00184834  0.00010835  17.0597 < 2.2e-16
Change pr. areas -0.00115911  0.00047140  -2.4589   0.01394
Change in cattle  0.00564735  0.00110023   5.1329 2.853e-07
Change soy bean   0.00684987  0.00750858   0.9123   0.36163
Change sugar cane 0.02473398  0.05564156   0.4445   0.65666
Time lagged defor 0.41590968  0.00381642 108.9790 < 2.2e-16

Lambda: 0.72431 LR test value: 13493 p-value: < 2.22e-16

Log likelihood: 156019.8
ML residual variance (sigma squared): 3.6955e-05, (sigma: 0.006079)
Number of observations: 42900
Number of parameters estimated: 8
AIC: -312020
```

From these results it can be seen that the change in protected areas and change in cattle are significant at the $\alpha = 0.05$ level, in addition to the spatial and temporal autoregressive terms. Change in soy bean and change in sugar cane were not found significant.

4 Discussion and conclusions

This paper gave a first step into the direction of explaining changes in deforestation for 25 km × 25 km grid cells covering the Brazilian Amazon by changes in protected areas, changes in amount of cattle, changes in soy bean and sugar cane plantation coverage. The regression model evaluated here considers yearly deforestation as it depends on the very limited set for which spatially distributed time series were available. In addition, it was only evaluated to which extent the change in deforestation depended on the *changes* in each of these independent variables, i.e. the variables as such were not included directly as predictor. As a consequence, a number of effects found significant may result from confounding effects. No pure time series (e.g. market prices) or purely spatial factors (e.g. climate) were included. Improved understanding of the governing processes may be obtained by evaluating a wider range of regression models.

The regression model entertained here (1) was held deliberately simple, and these first results should be interpreted with caution and some reservations. Improvement of these first results might be obtained when (i) transformation of the dependent and/or independent variables improve the linearity of the relationships, (ii) other grid cell sizes are used than the current 25 km × 25 km cells used here, (iii) more than one time lagged autoregressive terms are used (iv) a full maximum likelihood estimation procedure is used.

Previous results have shown that protected areas are significant in preventing deforestation in high-pressure areas, and the creation of those areas have been increased as a control policy applied by the Brazilian government.

On the other hand, a debate is growing about the extent of the deforestation as a result of the expansion of cattle (pasture) and soy industry. Most recent analyses suggest that deforestation is driven by the expansion of cattle ranching, rather than soy bean. Soy bean and sugar cane seem to be replacing deforested areas previously under pasture.

References

Aguiar, A. P. D., Camara, G., and Escada, M. I. S. (2007). Spatial statistical analysis of land-use determinants in the Brazilian Amazonia: exploring intra-regional heterogeneity. Ecological Modelling, Volume 209 (Issues 2-4), Pages 169-188

Bivand, R. S., Pebesma, E. J., and Gomez-Rubio, V. (2008). Applied Spatial Data Analysis with R. Springer, NY.

EVALUATION OF AREA-BASED CRIME PREVENTION PROGRAMS USING GEOSPATIAL DATA MINING

Ch. Kreis [a,b] *, M. Kanevski [b], A. Kuhn [a]

[a] Institute of Criminology and Penal Law, University of Lausanne, 1015 Lausanne, Switzerland - (christian.kreis,andre.kuhn)@unil.ch
[b] Institute of Geomatics and Analysis of Risk, University of Lausanne, 1015 Lausanne, Switzerland - mikhail.kanevski@unil.ch

KEY WORDS: Impact evaluation, crime prevention, data mining, self-organizing maps, GIS

ABSTRACT:

The current research is a test of the application of geospatial data mining algorithms to evaluate area-based crime prevention programs in Switzerland. Both unsupervised and supervised learning methods are used to detect spatio-temporal patterns in high-dimensional data on crime, socio-economic and demographic structure, and the built environment in order to enhance the validity of a non-experimental impact evaluation of community policing programs in the five major cities. The visualization of the results for better communication with practitioners and intelligence-based decision making forms an important part of the study.

1. INTRODUCTION

One of the dominant themes in criminology in recent years has been the recurrent demand that practices in the field be „evidence-based", meaning that criminal justice policies should be subjected to rigorous evaluation (Sherman et al., 2002). The methodological standards of evaluation research have been cogently defined already during the 1960s and the 1970s and have been reaffirmed more recently with particular reference to crime prevention (Cook and Campbell, 1979; Farrington, 2003). This body of knowledge posits a clear hierarchy of the methodological quality of different research designs, with the randomized controlled trial (RCT) held as the "gold standard" of scientific evaluation.

In the fields of crime prevention and policing, however, RCT designs have seldom been implemented, because field experiments are perceived as politically risky, or ethically questionable, or both. For area-based crime prevention programs, which target specific places or entire jurisdictions, finding a sufficient number of matching control areas can be difficult and statistical power correspondingly low. The consequent dearth of methodologically sound evaluations of policing strategies has prompted repeated calls for alternatives to the randomized experiment paradigm.

2. METHODOLOGY FOR PROGRAM EVALUATION

The current study employs geospatial data mining techniques as a means to enhance the validity of observational research designs of area-based crime prevention programs. Exploratory spatial data analysis or unsupervised learning algorithms serve as analytical and modeling tools for dimensionality reduction and clustering of the high-dimensional input data in order to classify areas with (dis-)similar features in attribute and geographic space (Kanevski et al., 2009). This initial step reflects earlier research on the use of geodemographic profiling for resource allocation for neighborhood policing (Ashby and Longley, 2005), and can be thought of as analogous to finding matching treatment and control groups for evaluation research. By contrast, supervised data mining algorithms seek structures in the data while taking a priori knowledge on outcomes or classifications into account. For the purposes of evaluation, supervised algorithms may be used to cluster neighborhoods into categories of different types in order to model the effect of program-related measures and contextual variables on the variance of a specific indicator of treatment outcome (e.g. fear of crime or satisfaction with the police). As a final step, geo-visualization techniques can be used to display the results of both the unsupervised and supervised learning procedures as geographic maps in a GIS. Such maps not only help to indicate areas where a policy has been particularly effective or may be implemented successfully, but also serve to communicate the results to both policymakers and practitioners.

The current study uses a comparative evaluation of community policing across Switzerland's five biggest cities to test the above theoretical argument. Community policing has long stressed crime prevention and the contextual role of the police, emphasizing greater interaction with local communities to solve persistent crime and disorder problems (Greene, 2000). The police departments under study began to embrace this new paradigm from the mid-1990s onwards but in Switzerland the strategy has never been subject to an empirical evaluation.

3. DATA AND METHODS

The data analyzed to evaluate the impact of community policing come from three main sources: (a) area crime rates are derived from official annual police crime records; (b) data on the demographic and socio-economic composition of the resident population, the housing structure, and local business and commercial activity come from the federal government's regular population census and business census counts, respectively; and (c) data on the indicators of community policing impact, i.e. the neighborhood levels of fear of crime and subjective

* Corresponding author.

victimization risk, perceptions of disorder, and popular satisfaction with the police are based on the samples from the Swiss Crime Survey, a large-scale victimization survey that has polled residents of the five cities repeatedly between 1998 and 2005. All categories of data are measured at the level of census tracts or postal ZIP code districts within the five cities, which are the finest level of spatial granularity for which crime and survey data are available.

During the exploratory phase, self-organizing maps are being used in combination with hierarchical agglomerative clustering of the resulting Kohonen map in order to develop a typology of urban neighborhoods (Vesanto and Alhoniemi, 2000). The objective is to create a neighborhood classification system based on the demographic, socio-economic, business and housing structure data that minimizes intra-neighborhood and maximizes inter-neighborhood variance in area crime patterns and the victimization survey responses that serve as indicators of community policing impact. In the supervised learning phase, these indicators of treatment impact are imposed one-by-one on the demographic, socio-economic, and environmental data during training so that the resulting neighborhood typology explains a maximum of the variance of a given outcome variable. The rationale of this procedure is to minimize the within-cluster variance in the contextual variables that affect the outcome variables and may thus confound inferences about program effect in an evaluation design without proper control group(s). This allows for a more accurate assessment of the impact of community policing strategies across different neighborhood types within the five urban areas.

4. RESULTS

The spatio-temporal patterns of the indicators of fear of crime, disorder, and public satisfaction with the police exhibited great variance over the long time-span evaluated. The three indicators measuring fear pointed out a decline in fear levels in the city centers and a simultaneous rise in the surrounding suburbs, underlining the need for a community policing impact evaluation to control for shifting neighborhood characteristics.

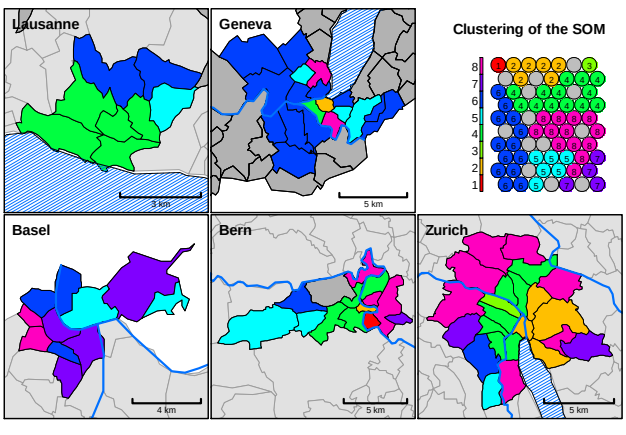


Figure 1. Map of neighbourhood typology across five major Swiss urban areas

The neighborhood typology resulting from unsupervised learning significantly reduced the variance of survey variables within clusters, especially perceptions of disorder and emotive measures of fear, despite the inherent heterogeneity at the current level of analysis. Supervised learning, by contrast, led to better results with regard to fear measures of more cerebral judgments about risk of victimization or behavioral reactions to crime, suggesting the latter should be used to evaluate specific crime prevention or fear reduction strategies.

5. DISCUSSION

The current study proposed a methodology, based on geospatial data mining, to evaluate community policing across five cities over a time span of several years. Sustaining an experimental research design over such a long period of time would be a tall order, yet the police literature often alludes to the "glacial" pace of police strategic reforms, the impact of which may take years to materialize (Greene, 2000). Where extraneous influences cannot be ruled out through randomization, geo-computational methods allow evaluators to track the spatio-temporal patterns of contextual variables that rival treatment as the plausible explanation for the observed variance in outcome variables. These methods thus hold the potential to enhance the validity of non-experimental research designs and so to narrow the gap between theory and practice in criminological evaluation research.

REFERENCES

Ashby D I, Longley P A, 2005. Geocomputation, Geodemographics and Resource Allocation for Local Policing. *Transactions in GIS*, 9(1), pp. 53-72

Cook T D, Campbell D T, 1979. *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Rand McNally College Publishing Company, Chicago.

Farrington D P, 2003. Methodological Quality Standards for Evaluation Research. *The ANNALS of the American Academy of Political and Social Science*, 587(1), pp. 49-68

Greene J R, 2000. Community policing in America: Changing the nature, structure, and function of the police. In: *Policies, Processes, and Decisions of the Criminal Justice System*, U.S. Department of Justice, National Institute of Justice, Washington, DC, pp 299-370.

Kanevski M, Pozdnoukhov A, Timonin V, 2009. *Machine Learning for Spatial Environmental Data: Theory, Applications and Software*. EPFL Press, Lausanne.

Sherman L W, Farrington D P, Welsh B C, MacKenzie D L, 2002. *Evidence-Based Crime Prevention*, Routledge, London.

Vesanto J, Alhoniemi E, 2000. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks* 11(3), pp. 586-600.

ACKNOWLEDGEMENTS

This research is supported by funding from the Swiss National Science Foundation (Award No. 100015-122463; Award No. 100017-132675). The opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect those of the aforementioned agency.

WHERE AND WHEN DOES THE TRAFFIC CONGESTION BEGIN AND END? A SPATIO-TEMPORAL CLUSTERING APPROACH TO DETECT CONGESTION

B. Anbaroglu [a,b] T. Cheng [a]

[a] Dept. of Civil, Environmental and Geomatic Engineering, University College London, UK
[b] Dept. of Geodesy and Photogrammetry Engineering, Hacettepe University, Ankara, Turkey
{b.anbaroglu; tao.cheng}@ucl.ac.uk

Commission II, WG II/3

KEY WORDS: Spatio-temporal clustering, similarity function, traffic congestion

ABSTRACT:

Detecting the extent of traffic congestion is an important research problem in traffic engineering. Although this problem has been extensively researched, none of the existing approaches deal with the problem from the spatio-temporal data mining aspect. This paper shows that this problem can be handled via doing spatio-temporal clustering of the road network.

1. INTRODUCTION

Detecting the extent of traffic congestion is an important research problem as it can be used in various problems like estimating the cost of traffic congestion or operational management. Some of the solutions on detecting traffic congestion treat it as a temporal phenomenon (Xiaolei et al., 2009), although it is commonly accepted that traffic congestion is a spatio-temporal phenomenon. On the other hand, the researches that benefit from both spatial and temporal dimensions require high preprocessing effort in terms of labelling recurrent and non-recurrent congestion to generate a model of congestion (Jin et al., 2006). However, the necessary information to label the non-recurrent congestion might not be readily available.

This paper shows that traffic congestion can be detected via a spatio-temporal data mining method; spatio-temporal clustering (STC). STC uses both the spatial and temporal aspects of the traffic congestion while doing this without the need of labelled traffic data, thus increasing its practical applicability.

This research builds upon the previously proposed STC algorithm (Cheng and Anbaroglu, 2010) in which STC has been used to detect traffic patterns on the road network. The main contribution of this paper is to show STC can be used to detect traffic congestion. To do this, we assumed that a congestion event starts and ends with a drastic speed change. A similarity function that can capture the drastic changes in the time-series data (e.g. daily speed variation) is proposed. These drastic changes would indicate the beginning and end of a congestion event (from now on, we will refer to each congestion event as "a congestion").

This paper is organized as follows. The next section presents the spatio-temporal clustering algorithm, which is followed by a case study in Section 3. The conclusions and future research directions are discussed in the last section.

2. DETECTING TRAFFIC CONGESTION VIA STC

STC has been developed in our previous research (Cheng and Anbaroglu, 2010) to find the traffic patterns of road networks based upon a similarity function which identifies clusters of links behaving similarly.

This similarity function can be defined according to the research aim (i.e. which patterns are of interest?) and STCs would take the relevant meaning accordingly. In this research we aim to detect congestion events by assuming that the congestion starts and ends with drastic speed change. This can be achieved by arranging the similarity function in a way that it outputs a positive similarity (i.e. '1' in equation 1) when the speed values change drastically.

For this purpose, we define the similarity function of the spatio-temporal clustering algorithm as in equation 1, where *percLimit* is a user defined percentage value between 0 and 100:

$$\text{simFunc}(p_t, p_{t+1}, x, y) = \begin{cases} 1, & \frac{|p_{t+1} - p_t| * 100}{p_t} \geq percLimit \\ & and \\ & \frac{|x-y| * 100}{x} \geq percLimit \\ 0, & otherwise \end{cases} \quad (1)$$

The p_t and p_{t+1} denote the speed values at link p at times t and $t+1$ consecutively, where $t \in (1, 2, .., T)$ and T is the total number of observations. The x and y can take any of the three possible assignments which would change the interpretation of the similarity function. The first is q_t and q_{t+1}, which yields to the comparison of the two adjacent links (p and q) at same times. The second is p_{t+1} and p_{t+2}, which yields to the comparison of the link p with itself at consecutive time periods. The third is q_{t+1} and q_{t+2}, which yields to the comparison of adjacent links at consecutive times.

Once all the similarities are found at all of these three search directions, those which are continuous in space and time are merged into a spatio-temporal cluster (Cheng and Anbaroglu, 2010).

If the change in percentage is greater than the defined *percLimit* in both of the consecutive speed observations (i.e. p_t, p_{t+1} and x, y) then a positive similarity is output.

3. CASE STUDY

The aim of the case study is to find the beginning and end of a congestion on a real life data set from Blackwall Tunnel region of London which is shown in figure 1a. Because of the space limitations, here we report the results obtained from southbound links (i.e. links 720, 578, 579, and 580 as shown schematically in figure 1b) on 3 June 2010.

The average speed data is collected at 5 minute intervals. Thus, for one day of data, *T* is equal to 288 (in one hour we'll have 12 observations and 12×24 = 288).

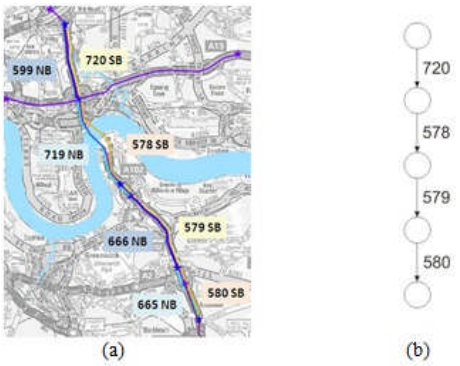


Figure 1. Case Study Area (a) and the Schematic Representation of the Links of the Experiments (b)

The *percLimit* in equation 1 is chosen as 20 after conducting empirical observations.

The STCs found on 3 June are illustrated as rectangles having labels from *a* to *x* in figure 2. After the visual interpretation of the STCs, they can take one of the three possible meanings.

Firstly, STCs represent the beginning and end of congestion, which are actually the sought STCs. For example, STCs represented by *s* and *v* denote the beginning and end of the traffic congestion that happened at links 579 and 580 between 16:30 and 18:15. Secondly, they can represent the congestion as a whole because these congestions have a shorter duration than the previous ones and the STCs that denote the beginning and end of the congestion is merged into one spatio-temporal cluster. For example, the spatio-temporal cluster represented by *b*, describes the congestion that happened at links 578 and 720 between 05:00 and 05:15. Thirdly, they represent the erroneous patterns as noise. For example, the spatio-temporal cluster represented by *n* and *t* are noise, since the sharp change in speed happened initially as an increase which indicates that the traffic was relieved at those times. Also, STCs that are observed during the congestion are not of interest. For example, the STCs having the label *r* and *u* occurred during the congestion happening between *s* and *v*, so they don't provide any further information in terms of understanding the congestion. Thus, we treat this kind of spatio-temporal cluster as noise as well. This differentiation of STCs of figure 2 is shown in table 1 as follows:

Meaning	STCs
Beginning and End of Congestion	f and k, l and m, q and w, s and v p and ?
Congestion	a , b, c, d, e , i, o, x
Noise	h , j , g , n , r , t , u

Table 1. The detected STCs and their meanings

It can be seen that the overall performance of the STC to infer about the beginning and end of a congestion is acceptable by capturing 4 pairs of STCs indicating the beginning and end of a congestion out of 5 possible (the congestion starting with *p* does not have a spatio-temporal cluster for its end). Also, it is observed that some STCs could be used directly to infer about the congestions which last shorter, and 8 such congestions were detected as STCs. There are 7 STCs that are found as noise.

4. CONCLUSIONS AND FUTURE RESEARCH

In this research we proposed a solution to detect congestion by using STC. The similarity function of the STC algorithm is designed in the way that it outputs a positive similarity if there is a drastic change in speed values, thus it can be used to capture the beginning and end of congestion.

Although the method can be used to detect the congestion events, it has several limitations as well. Firstly, it produces some erroneous STCs. This can be prevented by using more complex similarity function. Secondly; the meaning of the STCs should be interpreted visually which prevents the method's appliance on a wider network, thus its scalability. Although this problem seems to be inevitable due to the inadequacy of the validation data (i.e. the labelled data indicating the beginning and end of the congestions), we will be investigating ways of assessing the ways to automatically infer the meaning of the STCs. Finally; due to the assumption that congestion starts and ends with drastic speed changes, the slowly building congestion events will be undetected leading to false negatives. False negatives also occur when there the drastic speed drop between p_t and p_{t+1} was not continued between p_{t+1} and p_{t+2} as seen at the link 720 at times 01:30 and 03:15.

Our future research will focus to cover the aforementioned limitations. Firstly, we will be working to derive a better similarity function that will capture the congestion events while having lower false positives. Secondly, we will be investigating ways to remove the need for visual interpretation.

REFERENCES

Cheng, T. and Anbaroglu, B., 2010. Spatio-Temporal Clustering of Road Network Data. In F. Wang et al., eds. *Artificial Intelligence and Computational Intelligence*. LNCS, pp. 116-123.

Jin, Y., et al., 2006. Spatial-Temporal Data Mining in Traffic Incident Detection. In: *Workshop on Spatial Data Mining SIAM DM*.

Xiaolei Li et al., 2009. Temporal Outlier Detection in Vehicle Traffic Data. In *IEEE 25th International Conference on Data Engineering*. pp. 1319-1322.

ACKNOWLEDGEMENTS

This research is jointly supported by UK EPSRC (EP/G023212/1). First author thanks to Hacettepe University and Higher Education Council of Turkey for the PhD scholarship they provide. We also would like to thank to Transport for London for providing the data and insight about it. Finally, we greatly appreciate the reviewer's suggestive comments.

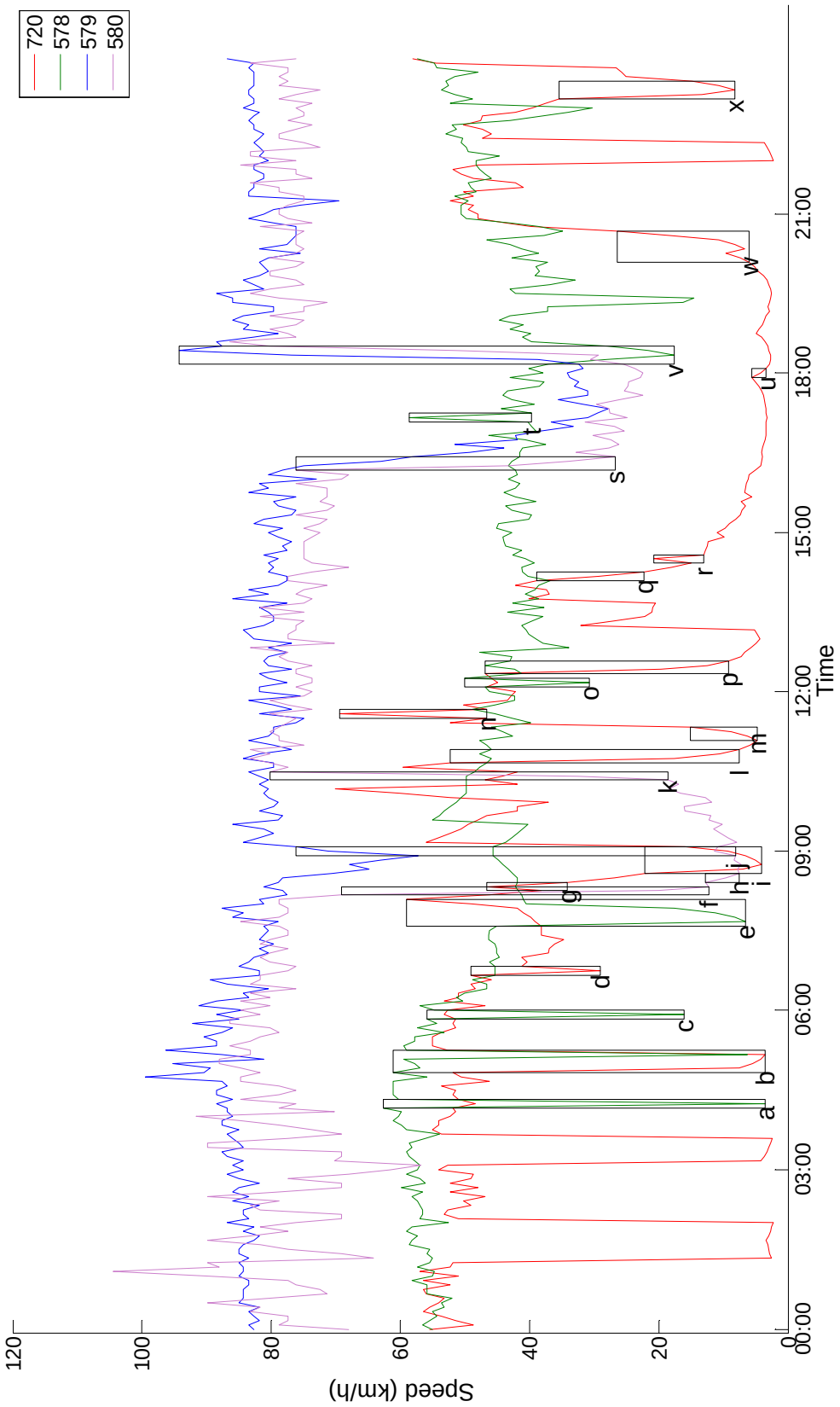


Figure 2. STCs found to detect traffic congestion

FILTERING SPATIAL POINT PROCESSES BASED ON KTH NEAREST TRANSFORMATION

Tao Pei [a, *],

[a] State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, 11A, Datun Road Anwai, Beijing 100101, China - (peit)@lreis.ac.cn

KEY WORDS: Poisson Process, Clustering Pattern, MCMC, Seismicity

ABSTRACT:
To determine arbitrarily-shaped clusters in a point data in an automatic way, we proposed a theory for filtering spatial point processes. In this theory, a set of point data can be seen as the mixture of Poisson processes (i.e. a set of events, each of which has the same likelihood of occurring in its support domain). The point data set can thus be separated into different Poisson processes according to their densities. In this paper, we use the *k*th nearest distance (the distance between an event and its *k*th nearest neighbour) to measure the local density of an event. To separate the point data into distinct Poisson processes of different densities, we constructed an objective function of the *k*th nearest distance, in which different Poisson processes are modelled as a probability density function (pdf) with the transformation of the *k*th nearest distance. EM algorithm and MCMC algorithm are employed to estimate the parameters of each pdf in the objective function. After the parameters are determined each Poisson process can be extracted. The method is applied to two seismic case studies and the results shows that the seismic clusters were correctly identified.

1. INTRODUCTION

A point process is defined as a series of events which occur in a restrained area, such as the seismicity, crime venues, disease cases, traffic accidents. The pattern of point process data are divided into three types: completely spatial randomness (CSR), clustering and regularity (Cressie, 1991). Because the clusters or hotspots in the clustering pattern may represent some anomalous phenomena, it is very important to identify the clusters and their support domains (the territories of the clusters). There are three types of methods for identifying clusters from the clustering patterns. The first is the celled-based method, such as STING (Wang et al., 1997), CLIQUE (Agrawal et al., 1998) and MAFIA (Nagesh et al., 1999). Although these methods can identify arbitrarily-shaped clusters, they need to calculate the local densities in cells and the clustering results will be significantly affected by the choice of resolutions of cells. The second is based on the *k*th nearest distance (the distance between an event and its *k*th nearest neighbour), such as DBSCAN (Ester et al. 1996), OPTICS (Ankerst et al. 1999). Although this type of methods does not need to calculate the local densities in cells, the parameters to separate the clusters and background are usually determined subjectively. If a data set contains different types of clusters (namely, there are many clusters of different densities in the data), they may produce errors or even fail to separate them correctly. The third is the model-based cluster method, such as the MCLUST method (Fraley & Raftery, 2003). The method can estimate the parameters of the classification model in an objectively way, but MCLUST can only identify Gaussian-shaped clusters. In order to determine the arbitrarily-shaped clusters in an automatic way, we proposed a theory for filtering point processes based on the *k*th nearest distance transformation.

2. METHODOLOGY

This theory is based on the hypothesis that the clustering point pattern is the mixture of unknown number of homogenous spatial point processes (i.e. spatial Poisson processes, we call it Poisson process hereafter) whose support domains are mutually exclusive and collectively exhaustive. In this context, the point processes with high densities can be thought of as the potential hotspots or clusters and those of low densities can be treated as background. As a result, the hotspots can be identified by filtering the point patterns according to the difference of densities between point processes.

The filtering process is achieved by the following steps. First, the *k*th nearest distance of each event in the point pattern is calculated. The Probability Density Function (pdf) of the *k*th nearest distance is:

$$f_{X_k}(x;k,\lambda) = \frac{e^{-\lambda\pi x^2} 2(\lambda\pi)^k x^{2k-1}}{(k-1)!} \qquad (1)$$

where X_k is the *k*th nearest distance, λ is the density of the Poisson process. Second, the pdf of the *k*th nearest distance for the pattern is modelled as the mixture function of unknown number of pdfs of the *k*th nearest distance of different Poisson processes. The objective function can be written as:

$$X_k \sim \sum_{i=1}^{m} w_i f(x;k,\lambda_i) = \sum_{i=1}^{m} w_i \frac{e^{-\lambda_i\pi x^2} 2(\lambda_i\pi)^k x^{2k-1}}{(k-1)!} \qquad (2)$$

where w_i is the proportion of the *i*th Poisson process with $\sum_{i=1}^{m} w_i = 1$, *m* is the number of Poisson processes and λ_i is the intensity of the *i*th Poisson process. Third, the parameters (i.e. λ_i and w_i) of the mixture function then can be estimated by

* Corresponding author. Tao Pei.

using optimization algorithm, such as Expectation-Maximization (EM) and the Monte Carlo Markov Chain (MCMC) (Pei et al. 2006; Pei et al. 2009). The Bayesian classification function is thus built based on the determined parameters. Finally, the clustering point patterns are decomposed into distinct Poisson processes. The clusters with high densities are determined by filtering out the background.

3. RESULTS

Two seismic case studies are then used to validate the theory. In the seismic data, the clusters are thought of as the foreshocks or the aftershocks of strong earthquakes or seismic sequences relating to some active tectonics (e.g. the fault or the boundary between plates) while the events with low density are thought of as the background earthquakes. In the first case study, we used the seismic data of eastern China to validate the method. Seismic clusters were identified by using our method, which were verified to be the clustering earthquakes in the Tanlu fault region and its adjacent area (Figure 1). In the second case study, we used the strong earthquakes of China. Result shows that the strong earthquakes were separated into five point processes of different densities. The point processes with high densities are the strong earthquake clusters. Their support domains were confirmed to be the most intensive seismic areas in China.

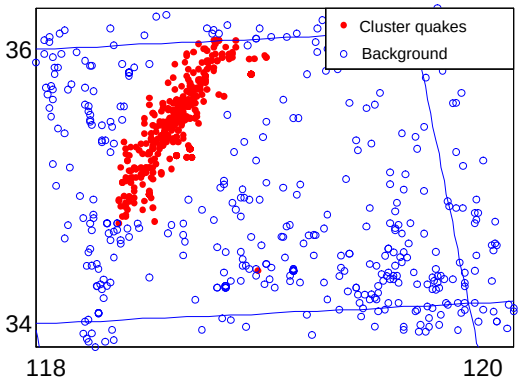


Figure 1: The clustering earthquakes of eastern China determined by the point process filtering method.

4. CONCLUSIONS

The contribution of our theory to point processes is similar to that of the Fourier transformation to the elementary functions. In the Fourier transform, any elementary function can be seen as the summation of base functions (e.g. the sinusoidal function and cosine function). The elements of different frequencies can be extracted based on the frequency spectrum of the function. In our theory, similar to the Fourier transformation, the clustering point patterns are seen as the mixture of a finite number of Poisson processes. Point data can be decomposed into distinct Poisson processes of different densities based on the *k*th nearest transformation. We can identify the clusters by filtering the point process data. The theory can also be extended in three directions. First, the method can be extended to handle multi-dimension points. Second, it can be extended to deal with spatio-temporal data if we use the windowed nearest distance (Pei et al. 2010). Third, in addition to clusters, we can delineate the support domains of the clusters so long as we replace the *k*th nearest distances of events with those of geometric points (the geometric point is a location on which there is no event) when modelling the mixture function (Pei et al. 2007).

REFERENCES

Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P. 1998. Automatic subspace clustering of high dimensional data for data mining applications, Proceedings of 1998 ACM-SIGMOD International Conference on Management of Data, New York: ACM Press, pp. 94–105.

Ankerst, M., Breunig, M.M., Kriegel, H.-P., and Sander, J. 1999. OPTICS: Ordering Points to Identify the Clustering Structure. Proceedings of ACM-SIGMOD'99 International Conference on Management Data, Philadelphia, USA, pp.46-60.

Cressie, N.A.C. 1991. Statistics for Spatial Data (1st edn). New York: John Wiley & Sons, Inc.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. Portland, OR, pp.226–231.

Fraley, C., Raftery, A. E. 2003. Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. Journal of Classification, 20, pp.263-286.

Nagesh, H., Goil, S., & Choudhary, A. (1999). Mafia: efficient and scalable subspace clustering for very large data sets, Technical report TR #9906-010, Northwestern University, 1999, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.36.8684 (accessed in Jan. 2009).

Pei, T., Zhu, A.X., Zhou, C.H., Li, B.L. and Qin, C.Z. 2006. A new approach to the nearest-neighbour method to discover cluster features in overlaid spatial point processes. International Journal of Geographical Information Science, 20, pp.153-168.

Pei T, Zhu AX., Zhou CH, Li BL, Qin CZ. 2007. Delineation of support domain of feature in the presence of noise. Computers and Geosciences, 33(5), pp.952-965.

Pei, T., et al., 2009. DECODE: A new method for discovering clusters of different densities in spatial data. Data Mining and Knowledge Discovery, 18, pp. 337–369.

Pei T et al., 2010, Windowed nearest neighbour method for mining spatio-temporal clusters in the presence of noise, International Journal of Geographical Information Science, 24(6), pp. 925 – 948.

Wang, W., Yang, J., Muntz, R. 1997. STING: A statistical information grid approach to spatial data mining. Proceeding of the 23rd International Conference on Very Large Data Bases (VLDB'97), Athens, Greece, August, pp. 186-195.

ACKNOWLEDGEMENTS

This study was funded through supports from a grant from the CAS (KZCX2-YW-QN303), a grant from the IGNSRR (Project Number: 200905004).

PATTERNS MINING IN SPATIAL-TEMPORAL SEQUENCES: THE CASE OF FOREST FIRES IN TICINO (SWITZERLAND)

C. D. Vega Orozco[1,*], M. Kanevski[1], M. Tonini[1], M. Conedera[2]

[1] Institute of Geomatics and Risk Analysis (IGAR), University of Lausanne, CH-1015 Lausanne, Switzerland - (CarmenDelia.VegaOrozco, Mikhail.Kanevski, Marj.Tonini)@unil.ch
[2] WSL Swiss Federal Research Institute, Insubric Ecosystems Research Group, Via Belsoggiorno 22, CH-6500 Bellinzona, Switzerland – marco.conedera@wsl.ch

KEY WORDS: Point patterns, cluster analysis, scan statistics, forest fires, spatio-temporal data analysis, hot spots.

ABSTRACT:

Forests cover about 30% of the total planet land area and play an essential role in the dynamic of life. Their distribution and composition are perturbed by *fires*: irregular events displaying a complex spatio-temporal distribution (Sousa, 1984). Worldwide statistics reveal a growing trend of these fires producing remarkable impacts on ecosystems, human settlements and environment. Currently, forest fire sequences can be modelled as a stochastic point process where events are characterized by their spatial locations (*X,Y* coordinates), occurrence in time, size of burnt area, ignition cause, slope, altitude, distribution of vegetation, etc. Cluster analysis of environmental point process is a fundamental approach to detect distribution of patterns in space and time which, in the case of fires sequences, can assist fire-managers to identify problematic areas, to implement preventative measures and to better distribute resources against fire. This paper aims detection of spatio-temporal clusters in forest fire sequences to identify forest fire hot spots regarding their ignition causes by means of the *space-time scan statistics permutation model*; a scan statistical methodology that, through the use of a gradually scanning a window across time and/or space, detects a local excess of events in a specific zone and tests whether the excess have rather occurred by chance (statistical significance). The case study is based on a geo-referenced forest-fire database collected by the Forest Service in canton Ticino (Switzerland) from 1969 to 2008, organized and stored in a relational database (Pezzatti *et al*., 2010). Results revealed that forest fires events in canton Ticino are mostly clustered in the southern region where most of the population is settled and with a predominantly cause of arson and private activities. On the other hand, two lightning-induced fires clusters were also identified in the north of the canton characterized by a mountainous region.

1. INTRODUCTION

Swiss Alpine forest is disturbed by numerous natural hazards among which *fires* play an important role in forest ecosystem structure; however it also threats socio-economic, ecological and environmental systems. Forest fires count for about 90% of the surface area burnt on a national level.

Canton of Ticino is located in the Southern Alps and is the most *fire-prone* region of Switzerland. Since the 1960s, the number of forest fires has significantly increased due to factors such as climate change and the accumulation of combustible material as a result of socioeconomic shifts.

The study of the spatio-temporal pattern of forest fire sequences imparts knowledge about the characteristics and structure of clusters which can assist fire managers and policy-makers to manner adequate distribution of fire-fighting resources.

From this point of view, this paper aims to detect spatio-temporal clusters of forest fire occurrence in canton Ticino, Switzerland, during the period from 1969 to 2008 in order to identify forest fire hot spots and characterize their origins of ignition. The geo-referenced dataset, provided by the Forest Service of canton Ticino, is characterized by 2,401 fire ignition points (*X,Y* coordinates) and counts with additional information such as occurrence in time, size of burnt area, ignition-cause, slope, altitude, etc.

Analyses were carried out applying a scan statistical method: *space-time scan statistics permutation model*; which allows detection and location of spatio-temporal clusters simultaneously and evaluate their statistical significance. Analyses of fire-ignition causes of outcome clusters were performed in order to identify the ignition-nature of the clusters and characterize the incidence of arson actions in fire occurrence in the canton territory.

2. METHODOLOGY

2.1 Study case

Canton of Ticino is located in the southern part of Switzerland. It has a total area of 2'812 km^2 and it is the most fire-prone region of this country. See Figure 1. Topographically, the canton is split in two parts, the northern part, more mountainous, called *Sopraceneri* and the southern part, less mountainous, called *Sottoceneri*, characterized by a hilly region where most of the population is settled.

The region of Ticino is commonly characterized by a warm-temperate and rainy climate, which differs from other Swiss regions. Altitudes range from 197 to 3,402 m.a.s.l. The mean annual precipitation depends on the geographical position due to the differences in the relief, ranging from 800 to 2600 mm. The mean annual temperature ranges from 3 to 12°C (Conedera, 2009).

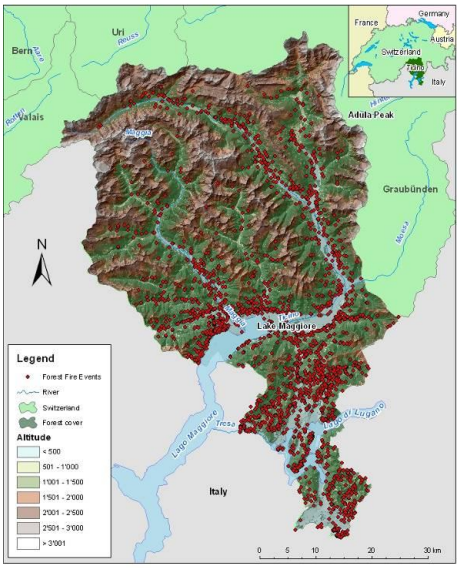


Figure 1. Canton Ticino and Forest Fire Events distribution.

At low elevations cultivation of the sweet chestnut (Castanea sativa) represents the dominant tree in the canton followed by the oak and other broadleaved species. At medium elevations (900-1400 m.a.s.l.) the forest mostly consists of Fagus sylvatica followed by coniferous forest (Picea abies). At higher elevations, it is found European larch (Larix decidua). (Conedera, 2007).

2.2 Data

The real case study is based on the 2'401 forest-fire event data of Canton Ticino (Switzerland) for a period from January 1969 to August 2008. This data has been collected by the local Forest Service of the canton for most forest fires in its territory since 1900. But it was only after 1969 when the dataset was enlarged to include drawings of the burned area on maps.

In 1990, the existing data was organized in a relational database including the location of the ignition points and the fire perimeters in a GIS platform by the Swiss Federal Institute for Forest, Snow and Landscape Research.

This wildfire database includes the following variables: geographical location (*X*,*Y* coordinates), fire-ignition date and time, fire-end data and time, fire duration, fire-ignition cause, size of burnt area, altitude and slope of the ignition point, etc.

2.3 Space-Time Scan Statistics

Scan statistic is a collection of methods widely applied to analyze and detect a local excess of events (clusters) in both time and/or space. It seeks to identify whether an observed cluster has occurred by chance assuming that the events are distributed independently and uniformly over space and/or time.

The first in introduce this method was Naus (1965), and more recently, Kulldorff *et al*. (1998, 2005) has developed spatio-temporal extensions.

Space-time scan statistics method uses cylinder windows to scan the study region, where the circular base represents the geographical extension (space parameter) and the height represents the time period of potential clusters, time parameter (see Figure 2). Windows sizes can increase from zero up to a maximum value defined by the user.

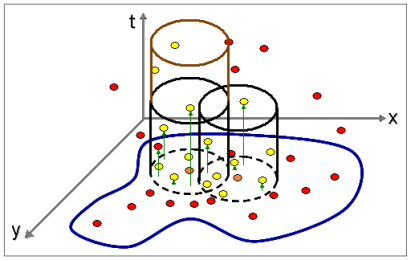


Figure 2. Diagram of the space-time scan statistics

The number of cases in a cluster is compared to what would have been expected if the spatial and temporal locations of all cases have no interactions (no dependence). In order to consider an area as a cluster, it must have a higher proportion of its cases in a specific period compared to the remaining geographical areas in that time period.

The probability that a certain area contains a cluster is estimated by the likelihood function which is computed for all space-time cylinders, and the one with the maximum value constitutes the most likely cluster (the cluster least likely to have occurred by chance).

The statistical significance of the candidate clusters is performed by Monte Carlo hypothesis simulations. In this papers this tested was done with 999 replications. All possible most likely clusters in the real dataset are ranked and compared them with the rank of the maximum likelihood from the random datasets. A threshold value is set to establish the level of the statistical significance, for example, if this level is defined at 5%, that is, a *p*-value of 0.05, a cluster from the real dataset is considered statistically significant when its rank is higher than the 95% of the Monte Carlo simulations for the random datasets (Kulldorff *et al*., 2005).

Space-Time Permutation Model

A major limitation for using some models in scan statistics to assess forest fire analyses is the lack of a control population data represented in this case by the total forest population risking to be burnt, to compute the expected number of cases inside each scanning window.

To avoid this problem, a *space-time permutation model* is used requiring only case data which corresponds to the forest fire events (Tonini *et al*., 2009). Therefore, the expected number of cases (control population) is evaluated on the base of the observed cases under the assumption that there is no space-time interaction.

Let C be the total number of observed cases and c_{zd} the observed number of cases within a zone z in a day d. The expected number of cases μ_A for a space-time cylinder A is estimated by the sum of μ_{zd} within cylinder A, (Kulldorff *et al*. 2005, Tuia *et al*., 2008):

$$\mu_A = \sum_{z,d \in A} \mu_{zd}$$
$$\mu_{zd} = \frac{1}{C}\left(\sum_z c_{zd}\right)\left(\sum_d c_{zd}\right) \quad (1)$$

where μ_A = expected number of cases for cylinder A
μ_{zd} = expected number of cases per zone z and day d
C = number of observed cases

Let c_A be the observed number of cases in cylinder A. Assuming this variable follows a hypergeometric distribution and that C is large compared to $\Sigma_{z \in A} C_{zd}$ and $\Sigma_{d \in A} C_{zd}$, then c_A can be considered as Poisson-distributed with mean μ_A. Thus, the evidence that cylinder A contains a cluster is measure by the Poisson Generalized Likelihood Ratio (GRL) (Kulldorff *et al*. 2005, Tuia *et al*., 2008):

$$GLR = \left(\frac{c_A}{\mu_A}\right)^{c_A}\left(\frac{C-c_A}{C-\mu_A}\right)^{(C-c_A)} \quad (2)$$

where μ_A = expected number of cases for cylinder A
c_A = observed number of cases in cylinder A
C = number of observed cases

This ratio is calculated and maximized for every possible cylinder and Monte Carlo simulation is performed to test the statistical significance of detected clusters.

2.4 Results and discussion

In the present study computations were performed with *SaTScan*™ software, developed by Kulldorff (2009).

Two simulations were performed:

1. First simulation was done by dividing the entire period (1969-2008) into three sub-periods (1969-1978, 1979-1990 and 1991-2008) in order to detect clusters in more homogeneous fire regimes conditions using the entire dataset which includes all type of causes. These sub-periods were defined according to the most significant fire-fighting measures, implemented by the cantonal authorities, which have influenced forest fire regimes during the analyzed period, such as the major fire brigades in 1978 and two preventive legal acts, 1989 and 1991 aiming the prohibition of fire in open spaces (Conedera, 2009).

2. And second simulation was performed using only lightning induced fires in the entire period (1969-2008), considering that these type of fires are not affected by the legal and action measures implemented by cantonal authorities.

The parameters for the space-time permutation model used to execute simulations were set as follows:

- Spatial window: 3 km maximum radius to prevent detection of clusters in large unconnected forest areas.

- Time window: for the first simulation, the time aggregation was set from 1 month up to 25% of the length of each sub-period; while for lightning induced fires simulation, time aggregation was set from 15 days up to 1 year.

- For the statistic significance, Monte Carlo simulations were set at 999, giving a smallest *p*-value of 0.001, and cluster detection was accepted at 5% level of confidence (*p*-value ≤ 0.05).

Results of the space-time permutation model for the three sub-periods using the entire dataset, presented in Table 1 and Figure 3, showed two statistically significant clusters for the first sub-period (1969-1978) with a radius of ~3km and 40m, respectively; while for the other 2 sub-periods (1979-1990 and 1991-2008), it detected 4 clusters in each sub-period with radius ranging from 0m (second cluster in 1991-2008 period) to 2,510m (in the same period).

Period	Cluster	Radius (m)	Time frame	Observed /Expected cases	P-value
1969 - 1978	1	2987.5	01-02-1973 28-02-1973	7/0.40	0.007
	2	40.0	01-11-1978 30-11-1978	3/0.03	0.050
1979 - 1990	1	2212.1	01-12-1986 31-01-1987	10/0.69	0.001
	2	2213.3	01-01-1984 31-01-1985	21/5.08	0.001
	3	1638.7	01-01-1981 31-01-1981	8/0.67	0.006
	4	1387.9	01-04-1981 30-06-1981	4/0.08	0.034
1991 - 2008	1	1439.6	01-10-1997 31-10-1997	7/0.13	0.001
	2	0.0	01-02-2001 28-02-2001	4/0.03	0.002
	3	2510.8	01-05-1997 31-08-1997	6/0.27	0.012
	4	998.5	01-03-1992 31-03-1992	4/0.07	0.018

Table 1. Space-time permutation model results for the three sub-periods

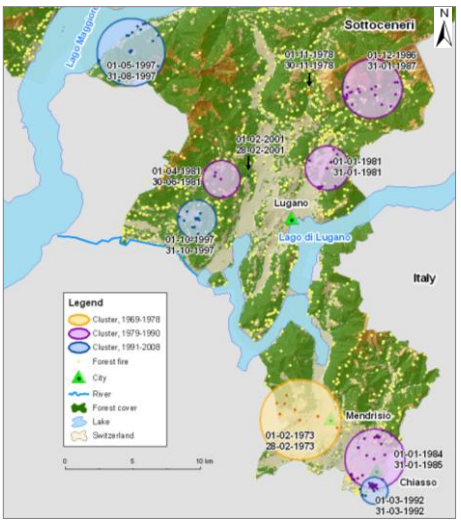


Figure 3. Forest fire clusters for the three sub-periods

The detected forest fire clusters are mostly localized in the hilly forest of the southern part of canton Ticino, *Sottoceneri*, where the highest population density is settle. These clusters are mainly characterized of anthropogenic origins. From the

seasonal point of view, six of the ten detected clusters occurred during the winter season (December – April), while two of the ten were developed in summer season (May – November).

For lightning-induced forest fires using the entire studied period (1969-2008), space-time permutation model outcomes are presented in Table 2 and Figure 4. They reported two clusters, one in 1989 and second in 1997, which are not considered statistically significant according to parameters. However, they were considered because of the interest to characterize their spatio-temporal structure which is different from those fires of anthropogenic origins. In this context, lightning-induce forest fire clusters were detected in the mountainous coniferous forests in the north-eastern part of the study area, *Sopraceneri*; where higher altitudes and slopes are found and less development is settle. This region covers two-thirds of the canton territory. These types of fires are reported in the geo-database only during the summer period (considered from May to November), therefore clusters are only detected in this period.

Period	Cluster	Radius (m)	Time frame	Observed /Expected cases	P-value
1969 - 1978	1	1375.0	30-06-1989 12-08-1989	3/0.051	0.007
	2	200.0	17-08-1997 30-09-1997	3/0.069	0.032

Table 2. Space-time permutation model results for lightning induced fires

Figure 4. Forest fire clusters for lightning induced fires

In order to identify hot spots due to some specific conditions, for instance particular characteristics of clusters concerning fire-ignition causes, the distribution of the fire-ignition sources in each single outcome cluster was compared with respect to the whole area during the same frame of time. Analyses are displayed in Tables 4 and 5.

Studying Table 4 in the third sub-period, 1991-2008, the first two clusters put in evidence a high occurrence of arson-induced fires; in the first cluster, October 1st 1997 to October 31st 1997, all fires are caused by arson actions, and in the second cluster the incidence of arson is 8.6 times more than the expected average in the period of February 1st 2001 to February 28th 2001.

Period	Cluster	Arson			Private		
		% clus	% perio	ratio	% clus	% perio	ratio
1969 - 1978	1	0.0	10.0	0.0	85.7	60.0	**1.4**
	2	0.0	5.0	0.0	66.7	47.5	**1.4**
1979 - 1990	1	60.0	25.8	**2.3**	30.0	30.1	**1.0**
	2	33.3	17.7	**1.9**	52.4	35.1	**1.5**
	3	37.5	46.0	0.8	0.0	25.3	0.0
	4	0.0	5.6	0.0	0.0	46.3	0.0
1991 - 2008	1	100.	0.0	-	0.0	25.0	-
	2	100.	11.6	**8.6**	0.0	44.2	0.00
	3	0.0	7.7	0.0	33.3	21.4	**1.6**
	4	0.0	17.7	0.0	50.	31.7	**1.6**

Table 4. Relative frequency of arson and private causes in the clusters (%clus= % in the cluster, %perio= % in the period)

Period	Cluster	Unknown			Lightning		
		% clus	% perio	ratio	% clus	% perio	ratio
1969 - 1978	1	14.3	30.0	0.5	0.0	0.0	-
	2	33.3	32.5	**1.0**	0.0	0.0	-
1979 - 1990	1	10.0	33.1	0.3	0.0	0.0	-
	2	14.3	30.5	0.5	0.0	5.3	-
	3	62.5	20.7	**3.0**	0.0	0.0	-
	4	75.0	32.8	**2.3**	25.0	4.5	**5.5**
1991 - 2008	1	0.0	37.5	0.0	0.0	0.0	-
	2	0.0	34.9	0.0	0.0	0.0	-
	3	16.7	15.8	**1.0**	0.0	40.2	0.0
	4	0.0	39.2	0.0	0.0	0.0	-

Table 5. Relative frequency of unknown and lightning causes in the clusters (%clus= % in the cluster, %perio= % in the period)

In both tables (4 and 5), clusters from sub-period 1979-1990 exhibit, with some little exceptions, mainly frequencies of arson, private and unknown causes: in the two first clusters higher incidences are characterized by arson actions; in the third cluster, unknown causes presents the highest incidence regarding other causes; and for the fourth cluster, the maximum occurrence is given by lightning-induced fires.

3. CONCLUSIONS

A cluster analysis method for forest fire events in canton Ticino was carried out. Statistical significant spatio-temporal clusters at local scales were detected in the canton in a period from 1969 to 2008.

This cluster analysis was performed applying a scan statistics technique: the *space-time permutation model*; which has revealed to be a rather useful method for environmental data analysis because it does not require a control population. As it was highlighted before, this population-at-risk, for the case of forest fires, is quite difficult to quantify due to the complexity of the factors and components that intervene in the availability of burning material such as the biomass. Thus, space-time scan statistic permutation model calculates the expected number of cases on the base of the observed events under the assumption that the spatial and temporal locations are independent of each other.

These analyses had uncovered valuable information of the structure characterizing forest fire clusters in canton Ticino. Spatial and temporal components such as the frequency incidence of arson and private actions on the ignition of fires and the geographical differences between human and natural-induced fires were clearly identified. Therefore, these analyses provide effective information that can be employed in fire-fighting planning taking into account differences on fire behaviour regarding their ignition origins, topography, proximity to human settlements and meteorological conditions of fire-prone periods.

Space-time permutation model results where combined with GIS techniques in order to map and display the detected clusters. This approach is very functional in the identification and mapping of fire-prone zones to support decision-making.

The future research will deal with the detection and study of anisotropic clusters and multivariate extensions of the Scan Statistics.

REFERENCES

Conedera, M., 2009. Implementing fire history and fire ecology in fire risk assessment: the study case of canton Ticino (southern Switzerland). University Karlsruhe (TH), Switzerland.

Conedera, M., Vassere, S., Neff, C., Meurer, M. and Krebs, P., 2007. Using toponymy to Reconstruct Past Land Use: a Case study of "Brüsáda" (burn) in Southern Switzerland. *Journal of Historical Geography*, 33(4), pp. 729-748.

Kulldorff, M., 1997. A Spatial Scan Statistic. *Communications in Statistics: Theory and methods*, 26(6), pp. 1481-1496.

Kulldorff, M., 2009. SaTScan™ User Guide for version 8.0. Technical Documentation. http://www.satscan.org/techdoc.html (accessed 5 Feb. 2010)

Kulldorff, M., Athas, W.F., Feuer, E.J., Miller, B.A. and Key, C.R., 1998. Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos. *American Journal of Public Health*, 88(9), pp. 1377-1380.

Kulldorff, M., Heffeman, R., Hartman, J., Assunção, R. and Mostashari, F., 2005. A Space-Time Permutation Scan Statistic for disease outbreak detection. *PLoS Medicine*, 2, pp. 216-224.

Naus, J., 1965. Clustering of random points in two dimensions. *Biometrika*, 52(1/2), pp. 263-267.

Pezzatti, G.B., Reinhard, M. and Conedera, M., 2010. Swissfire: die neue schweizerische waldbranddatenbank. *Schweiz. Z. Forstwes*, 161(12), pp. 465-469.

Sousa, W.P., 1984. The role of disturbance in natural communities. *Annual Review of Ecology and Systematics*, 15, pp. 353-391.

Tonini, M., Tuia, D. and Ratle, F., 2009. Detection of clusters using space-time scan statistics. *International Journal of Wildland Fires*, 18(7), pp. 830-836.

Tuia, D., Ratle, F., Lasaponara, R., Telesca, L. and Kanevski, M., 2008. Scan statistics analysis for forest fire clusters. *Communications in Nonlinear Science and Numerical Simulation*, 13(8), pp. 1689-1694.

EXTRACTING CLUSTERED URBAN MOBILITY AND ACTIVITIES FROM GEOREFERENCED MOBILE PHONE DATASETS

Y. Yuan*, [1], M. Raubal[1, 2]

[1] Dept. of Geography, University of California, Santa Barbara, USA, 93106 – (yuan, raubal)@geog.ucsb.edu
[2] Institute of Cartography and Geoinformation, ETH Zurich, Switzerland - mraubal@ethz.ch

Commission VI, WG VI/4

KEY WORDS: Human Mobility, Information and Communication Technologies (ICTs), Urban Clustering

ABSTRACT:

This paper focuses on extracting and understanding the clustering of human mobility based on the tracking information in mobile phone records. Two categories of clustered mobility (dynamic clustering and points of interest clustering) are explored in the sample dataset. The results provide references for understanding human mobility and updating urban policies in the age of instant access.

1. INTRODUCTION

Modeling human mobility and activities in urban systems has been a continuing research topic in several areas such as transportation planning and behavior modeling. However, most of the previous research is based on data acquired from travel diaries and questionnaires, which is a widely adopted data collection method when studying individual behavior (Yamamoto *et al.* 1999). Due to the limited number of people covered by travel diaries, these datasets fail to provide comprehensive evidence when studying the characteristics of the whole urban system, such as identifying clustering of urban mobility. Meanwhile, the development of information and communication technologies (ICTs) created a wide range of new spatio-temporal data sources (e.g., georeferenced mobile phone records). These datasets opened the way to a new paradigm in urban planning, i.e., Real-time cities (Ratti *et al.* 2007), as well as facilitating the study on behavior analysis and spatio-temporal data mining (Miller 2009). Undoubtedly, these technologies are a major step forward in identifying and characterizing dynamic hotspots and clustering in urban systems.

In this research, we focus on extracting clustered human mobility and activities based on a mobile phone dataset from northeast China. The research is conducted from two perspectives: 1) dynamic clustering (such as the hourly mobility patterns) and 2) Points of interest (POIs) clustering (such as the home locations of residents). The second aspect is considered as an indirect result derived from the first one. There have been several studies on modelling urban dynamic patterns from mobile connection datasets (e.g., the real time Rome project at the MIT SENSEable Lab[1]), but our research focuses on extracting the implications of various clustering patterns, as well as relating these patterns to the distribution of urban infrastructures. These results would be very useful in updating environmental, urban and transportation policies. Moreover, the results can be used as informants of human activity including long-term choices such as where to live and short-term choices such as daily activity scheduling. In addition, it is highly helpful for policy makers to understand the characteristics of individual mobility with wide-spread ICT usage, as well as updating environmental and transportation policies.

2. DATASET

In this research we use a dataset from city A[&], which is a major commercial and transportation center in northeast China. Figure 1 shows a basic road map of the city.

Figure 1. The basic road map of City A

The dataset covers over one million people and includes mobile phone connection records for a time span of 9 days. It includes the time, duration, and approximate location of mobile phone connections, as well as the age and gender attributes of the users. Table 1 provides several sample records. The phone number, longitudes and latitudes are not shown for reasons of privacy.

Phone #	**1350*********
Longitude	**126.*******
Latitude	**45.*******
Time	**14:36:24**
Duration	**12mins**
Receiver phone #	**1340********

Table 1. Sample records from the dataset of city A

* Corresponding author
& The name of the city is not shown as required by the data provider
1 http://senseable.mit.edu/realtimerome/

For each user, the location of the nearest mobile phone tower is recorded both when the user makes and receives a phone call, resulting in a positional data accuracy of 300m-500m. Note that the locations are recorded only when a phone call connection has been established.

3. METHODOLOGIES AND PRELIMINARY RESULTS

3.1 Identifying dynamic clustering

In Figure 2, we locate the phone connections that occurred during three time periods (T_1: 8am-9am; T_2: 2pm-3pm; T_3: 7pm-8pm) on two separate days: one weekday (07/23/2007) and one weekend day (07/21/2007), then we conduct a kernel density analysis to model the distribution of mobility density.

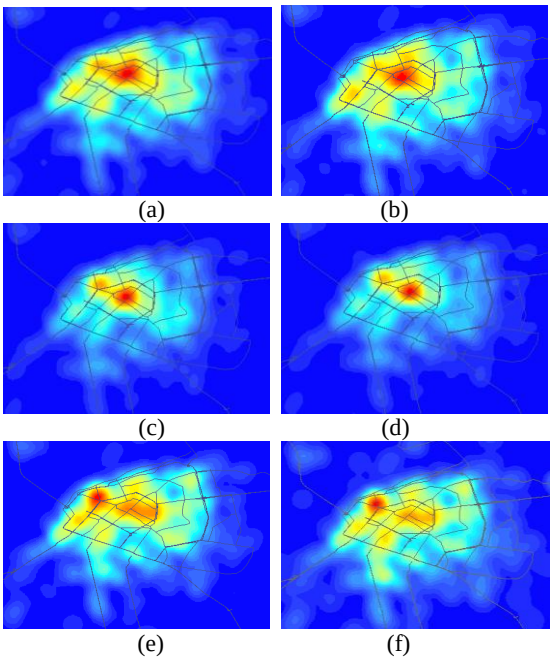


Figure 2. Changing clustering of urban mobility in city A. (a) T_1, weekday; (b) T_1, weekend day; (c) T_2, weekday; (d) T_2, weekend day; (e) T_3, weekday; (f) T_3, weekend day

As indicated in Figure 2, the hotspots of urban mobility change for different time periods. In both T_1 and T_2, the densest clustering appears in the city center, whereas in T_3, the pattern is more spread over the city except for a small clustering in the northwestern part. This may indicate that the cluster in the Northwest represents resident home locations and we will further confirm this hypothesis in the later part of Section 3. Moreover, Figure 2 shows a high similarity of mobility patterns between weekday and weekend day for all three time spans.

Some clustering patterns are also different for various population groups (e.g., age, gender). Figure 3 shows the mobility hotspots for two age groups during 2-3pm on a weekday: teenagers (age 12-17) and seniors (age>60). As indicated in Figure 3, the clustering of teenagers appears both in the center and in the Northwest of the city, whereas the density pattern of seniors is more widely distributed. Due to the different demands for living resources in various population groups, such analysis can provide helpful references for updating urban infrastructures for different population groups (e.g., high schools, hospitals, etc.). For example, there appears to be a cluster of seniors in the north part of the city, the center of which is very close to a large park in City A.

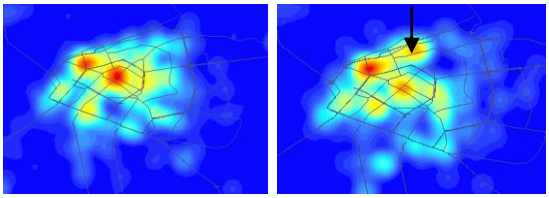


Figure 3. The clustering of (a) teenagers and (b) seniors

3.2 Identifying POIs clustering

In this research, POIs refer to regularly visited locations (e.g., home locations) of mobile phone users derived from their mobility patterns. Such analysis offers valuable input for enriching the personal profiles of users and studying urban areas according to their functions. Here we applied the methodologies described in Phithakkitnukoon *et al.* (2010) to identify the *stops* in trajectories: The trajectory of a certain individual is identified as a sequence of chronological locations:

$$R = \{(p_1, t_1) \rightarrow (p_2, t_2) \rightarrow .... \rightarrow (p_n, t_n)\}$$

Where the p_i refer to spatial locations and the t_i refer to time points. Then the trajectories are regrouped into sub-trajectories based on the restriction that any two consecutive points within a sub-trajectory are located within the cell of the same mobile phone tower. If the time duration of a sub-trajectory is longer than the temporal threshold ΔT, the sub-trajectory is identified as a *stop* for the particular user. Once the *stops* have been extracted, the home location of each user is estimated as the most frequent *stop* during the night hours and the work location is the most frequent stop during day hours on weekdays. Figure 4 demonstrates the distribution of home and work locations of users in City A. These POIs can also be combined with our previous research on user trajectory patterns to further examine the determinants of an individual's activity space (Yuan and Raubal 2010).

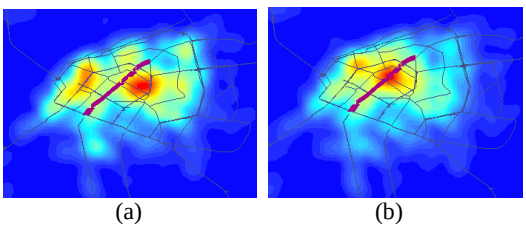


Figure 4. Clustering of (a) home locations and (b) work locations

As shown in Figure 4, both home and work locations are clustered in the city center, however, there are slight differences between the locations of hotspots in Figure 4a and Figure 4b. The highlighted street in Figure 4 is one of the main streets and it runs across the whole city. The home locations are mostly concentrated on the southeast side of the main street, whereas the work locations are evenly distributed on both sides of the street. Additionally, the home locations show two clustering centers in the study area (one in the middle, the other on the western side), indicating that city A has multiple active sub-areas that function as residential districts.

4. CONCLUSIONS

The pervasive usage of mobile technologies highly facilitates the modeling of urban mobility from different perspectives (e.g., mobility flow between sub-areas, transportation mobility density, etc). In this research, we demonstrate that mobile information is highly effective in characterizing the dynamic clustering of urban mobility. Furthermore, we extracted the POIs based on user trajectories and discussed the distribution of work and home locations across the whole city. Although this research focuses more on the implications of clustering results rather than the methodologies of cluster detection, in a next step we will generate a dynamic cluster detection and cluster significance model to simulate the daily rhythm of human mobility in urban systems. Future research will also focus on correlating the mobility patterns with the distribution of various types of urban infrastructures. Other continuing research includes the modeling and predicting of spatio-temporal trends of urban activities based on existing mobility patterns, as well as characterizing the regularity of individual space-time paths.

REFERENCES

Miller, H., 2009. Geographic data mining and knowledge discovery: An overview. In: H. J. Miller and J. Han, Editors, *Geographic Data Mining and Knowledge Discovery (Second Edition)*, CRC Press, London, pp.3-32.

Phithakkitnukoon, S., T. Horanont, G. Di Lorenzo, R. Shibasaki and C. Ratti, 2010. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In: A. A. Salah, T. Gevers, N. Sebe and A. Vinciarelli, Editors, *HBU 2010*, LNCS, Springer, Heidelberg, pp. 14-25.

Ratti, C., A. Sevtsuk, S. Huang and R. Pailer, 2005. Mobile Landscapes: Graz in Real Time. In *the 3rd Symposium on LBS & TeleCartography*. Vienna, Austria.

Yamamoto, T. and R. Kitamura, 1999. An analysis of time allocation to in-home and out-of-home discretionary activities across working days and non-working days. *Transportation*, 26, pp. 211–230.

Yuan, Y. and M. Raubal, 2010. On Correlation between Mobile Phone Usage and Travel Behavior – A Case Study of Harbin, China (Extended Abstract). In: *Geographic Information Science - 6th International Conference, GIScience 2010*, Zurich, Switzerland.

WAYFINDING AND VISITOR TRACKING IN MUSEUMS: ACCURACY ASSESSMENTS OF HYBRID POSITIONING SERVICES

T. Rains[a], J. Barros[b]

[a] Birkbeck, University of London, Malet Street, London WC1E 7HX, +44 (0) 7894 587720 tim.rains@cbuchanan.co.uk
[b] Birkbeck, University of London, Malet Street, London WC1E 7HX, +44 (0) 20 7079 0644 Fax: +44 (0) 20 7631 6498 j.barros@bbk.ac.uk

KEY WORDS: Wayfinding, Visitor Tracking, Museums, Hybrid Positioning, Wireless Positioning

ABSTRACT:

This study investigates whether indoor positioning systems on smartphones could benefit both visitors to museums, and the museums themselves, through serving as a wayfinding and visitor tracking aid (for museum planning purposes). As with all positioning systems, the accuracy of the systems needs to be understood when assessing the fitness for purpose of the system. This study therefore investigates four services (Google, Navizon, Skyhook and Wigle) offering positioning systems that use a combination of wireless network reference databases, GPS and cell towers. This is achieved by examining the accuracy of the positions that they report around the Victoria and Albert Museum in London. The results find that all services investigated offer too much positional error to be suitable for such applications. However the findings provide a useful benchmark for others wishing to harness similar services who may have different positional accuracy requirements.

1. INTRODUCTION

Paper maps have traditionally provided assistance to those trying to find their way around complex buildings such as museums. However, it is not uncommon for people to have problems with reading maps and spatial orientation, particularly when faced with an absence of known landmarks. Paper maps also offer limited feedback to museums on the exhibits visitors view or the routes they choose to take through the space, confining visitor research to traditional approaches such as face to face interviews or manual paper-based tracking by intrusively following visitors.

In recent years, relatively new technologies, such as GPS, have revolutionised the way in which we navigate and collect information on journeys. In museums, handheld guides have made the transition from early radio broadcasts to multimedia guides operating on PDAs and Smartphones. These two aspects are now starting to converge, such as in the American Natural History Museum Explorer application (AMNH, 2010) and the Museum of London's Street Museum (MoL, 2010).

Yet such developments are faced with challenges. Whilst outdoor positioning can easily be detected by the GPS system, indoor positioning can not make use of the same technology. In order to solve the problem of indoor positioning, different technologies have been proposed and tested in both research and museum environments through a variety of applications. Implementing such solutions on widespread scales can be testing, due to reasons such as high costs, insufficient positional accuracy, and the need for bespoke hardware and software.

One of the most well known methods involves harnessing the power of existing wireless infrastructure present in urban areas (see e.g. Hermasdorf, 2006), which can be then used as a base network to facilitate positioning indoors. This technology, known as hybrid positioning (when packaged with cell and GPS methods), is provided by several companies (and a community) and is readily available to developers of smartphone applications. But despite some publicity claims of 10-20 meter accuracy, it is uncertain whether the method is sufficiently accurate as there have been few published studies investigating these claims. This is a point highlighted by Zandbergen (2009), who partially answers this by investigating the accuracy of iPhone positions. Others (e.g. Curran et al, 2009) comprehensively investigate indoor positioning, but have more of a focus on comparing across alternative methods rather than comparing similar services by multiple providers.

This study attempts to fill this gap through investigating four key providers of wide-scale hybrid positioning services (Google, Navizon, Skyhook and Wigle). Tests of the positional accuracy reported are conducted, from which we establish the suitability for the purpose of wayfinding and visitor tracking around museums as part of a case study at the Victoria and Albert Museum in London, UK.

2. INDOOR POSITIONING TECHNOLOGY

A number of approaches have been proposed to overcome the problem of indoor positioning and it can be argued that this is a research domain in its own right. Approaches have included cellular network strengths, WiFi signals, Bluetooth, Infrared, Ultrasound and other radio frequencies, RFID tags, shape and image recognition, and more recently, augmented reality and TV signal based technologies (Zandbergen 2009; Benksy 2008; Rosum 2010).

Within the museum environment, Radio Frequency Identification tags (RFID) have been a commonly used technology (see Baldwin et al (2009) and Ishikawa et al 2009). Alternative approaches include Image Recognition (Chan et al 2005), Bluetooth (Bay et al 2005), Wireless Network Fingerprinting (Tsai et al 2010), audio beacons

(Landau et al 2005) and a combination of multiple methods as outlined by Savidis et al (2008). Whilst these technologies have been implemented to solve various core problems (such as navigation), another facet is understanding visitor behaviour in order to provide more personalised museum content and plan museum layouts. (See Bohnert & Zuckerman, 2009).

The present study focuses on the use of existing 802.11 wireless access points (APs) to establish positions. This method uses triangulation based on the attenuation of the signal strengths emitted from all APs. In order to provide the position, a reference database is required which holds the positions of the APs. As these are usually unknown, the database is populated through a process called wardriving which involves travelling around and recording the information emitted by APs alongside GPS signals, then using these to reverse calculate the estimated AP location. One of the first studies that implemented this approach was the Place Lab project, which consisted of a database of APs stored in a server that mobile devices accessed for a position to be provided (LeMarca et al 2005).

The eruption of smartphones, each fitted with sensors such as digital compasses, wireless, and GPS, has brought about the need to investigate whether these devices, coupled with wireless and hybrid positioning systems, can be deemed suitable for navigation and tracking in such environments.

3. THE CASE OF THE V&A MUSEUM

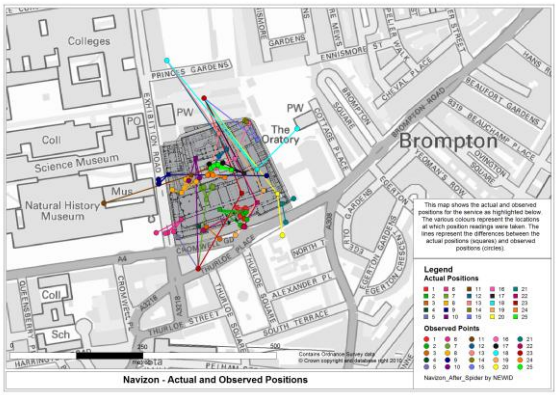
The V&A museum was chosen as a suitable location after tests in museums across London due to the extent of the existing wireless networks found within the building, eliminating any hardware requirements. Mapping of the wireless landscape around the building and within its courtyard (where GPS signals were available) was first carried out, with the intention of refining and updating the databases in question. This was completed to test the services in a situation where the services could report higher accuracies – the better the reference positions, the better the estimated positions should be. Problems found during this stage are examined within the full paper.

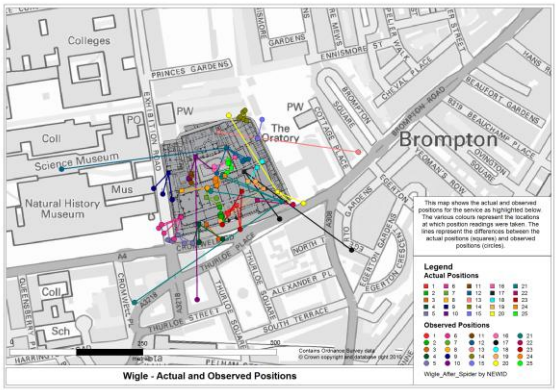
Within the museum, a series of known benchmark locations were established by using CAD plans for the building. In the sense of wayfinding in museums, we are interested in whether or not the positioning systems were able to provide positions that were accurate to room and exhibit level. However, to make the results applicable to different purposes, other metrics of error were also determined, such as the Root Mean Square Error (RMSE) and median errors.

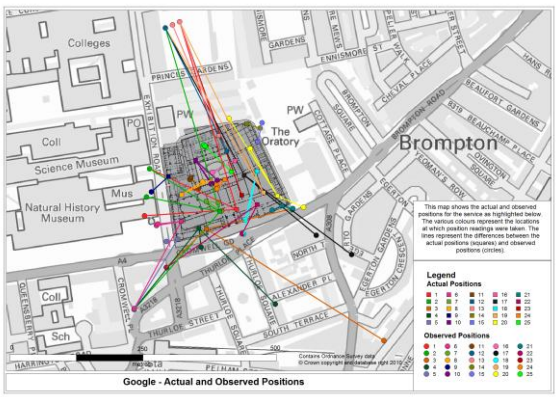
At each benchmark location, observed positions were recorded through the interfaces of each of the applications used for the four companies. These positions were plotted during analysis, and the error was calculated by straight line distance between the observed and actual position. Five readings for each provider were collected at 25 locations within the museum whilst walking a known route through the museum (simulating a visitor tracking process). These recordings are presented in the following section.

4. FINDINGS

The tests carried out show severe limitations with this method of positioning, as the results of all providers have shown unsuitable levels of positional error for both wayfinding and visitor tracking. The error at room level and by distance of each service investigated as part of the study is too great to be classed as fit for purpose. The actual and observed positions are highlighted below in Figure 1.







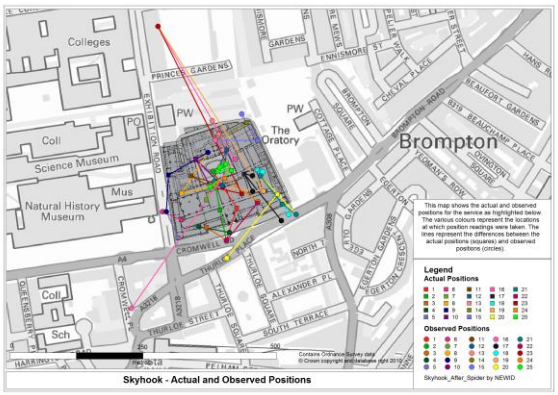


Figure 1. Comparison between Observed and Actual Positions.

Table 1, below, shows a comparison between services, with overall results for each provider.

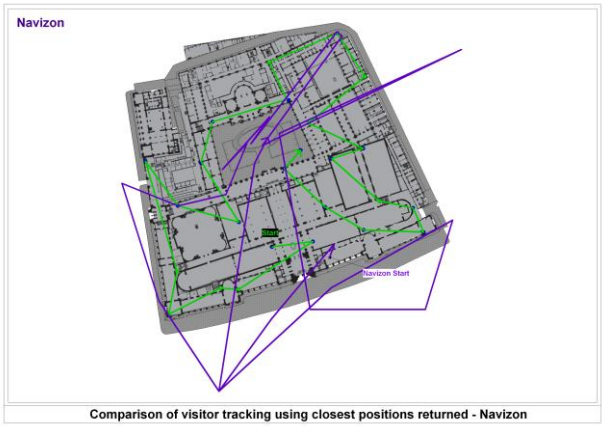
Service	**ID**	**Obs**	**Min Error**	**Max Error**
Navizon	All	110	1.15	262.95
Wigle	All	125	1.65	334.05
Google	All	125	5.11	391.18
Skyhook	All	125	8.23	364.24
Service	**RMSE**	**Mean Error**	**Median Error**	**68th Percentile**
Navizon	94.47	70.34	39.98	69.99
Wigle	95.72	62.86	37.45	48.63
Google	167.48	122.9	59.64	167
Skyhook	98.59	69.68	57.18	66.77
Service	**95th Percentile**	**Number Correct to room level**	**Number <20m**	**Number <50m**
Navizon	204.34	7	10	66
Wigle	263.02	8	29	90
Google	356.4	3	6	52
Skyhook	259.01	11	13	47

Table 1. Summary of results

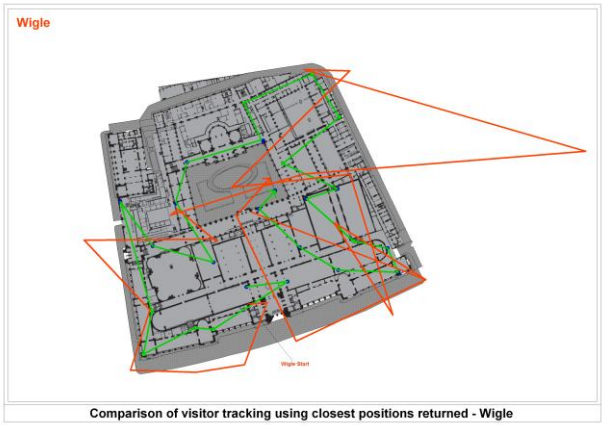
From the results in Table 1, it is clear that Wigle offers the lowest mean and median errors, along with the most positions less that 20 and 50 metres by a significant margin. Navizon showed the smallest RMSE. Google was the least successful, followed by Skyhook, although the latter did find the most positions correct to room level (most of which came when the service switched to GPS in the courtyard).

Navizon, Skyhook and Wigle, displayed RMSEs that were significantly larger than the 68th percentile, suggesting a non-normal distribution. For these, the median is a better metric of error than the RMSE. Google's results displayed a normal distribution, for which the RMSE is a suitable metric (Zandbergen 2008).

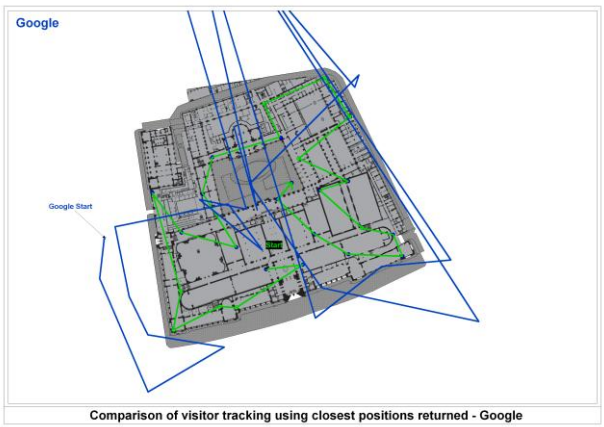
Tracking of visitors is clearly limited by these reported errors. Figure 2 shows comparisons of the actual track through the museum (shown in green) alongside the track created by joining together the closest positions found by each service at each location. Note that Google often returned vastly inaccurate points as the closest positional fixes (thought to occur when using cell towers rather than the wireless database as the reference network), and that Skyhook very frequently returned exactly the same positions for a variety of sample locations, leading to the patterns shown. Wigle shows the track that most closely resembles reality, though none of the positions found come close to being suitable for visitor tracking purposes, instead taking us on a journey through interior and even exterior walls with careless abandon.



Comparison of visitor tracking using closest positions returned - Navizon



Comparison of visitor tracking using closest positions returned - Wigle



Comparison of visitor tracking using closest positions returned - Google

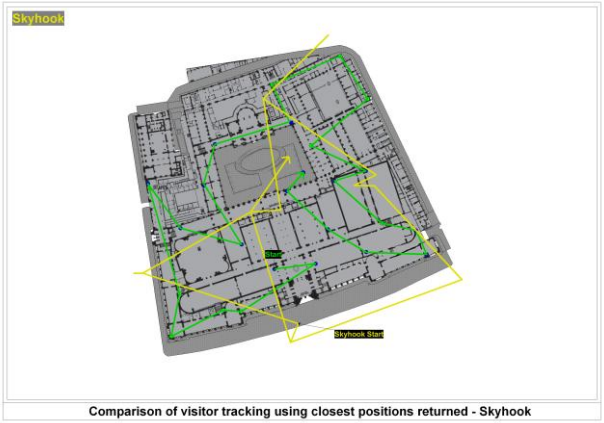


Figure 2. Comparisons between actual track and reported track through museum

The extent of these errors can be explained by a number of factors. Firstly, the method of estimating AP positions through wardriving contains limitations when creating reference databases to apply positioning algorithms from (Kim et al, 2006). Secondly, multipath errors mean that the distances calculated in triangulation can vary significantly, even by small events such as another visitor walking through the signal. A third point relates the architecture of the services. As they offer a position based on a single scan, which is sent/returned to/from the server, they offer no ability to compare this to the recent history of the device's movements. This does not aid the tracking process.

The wireless network within the V&A itself is also a factor that must be considered. As it is not set up for positioning purposes, but rather for traditional internet/network connection purposes, the network is largely concentrated in one side of the museum and the density affects the accuracy of the position estimates within other parts of the building.

Despite the failings in terms of the purposes investigated here, such positional error may be acceptable for other uses requiring less resolution. The full study also provides evidence that intense wardriving can refine the position estimates provided by some services due to an increase in the volume and accuracy of the AP positions within the reference database.

5. CONCLUSIONS

Of the positioning providers investigated in this study, Wigle's service was shown to be the most accurate and Google's service the least. The results obtained by testing the providers on the case study show that hybrid positioning systems are not suitable for use in museum tour-guide, wayfinding or visitor tracking applications. This is due to the insufficient positional accuracy in the horizontal plane, and the lack altogether of a position in the vertical dimension (of importance within multi-floored spaces).

Whilst acknowledging that each location, museum or otherwise, will yield different results (due to the nature of the wireless landscape), the study was carried out in a space deemed to have better than average volumes of APs (in comparison to other London museums). This means that the positional accuracies reported are likely to be more refined than indoor spaces with less extensive wireless networks. Using wireless APs for positioning in this manner will clearly produce positional estimates of varying accuracies in different situations – less or more APs, a better or poorer reference database, for example – but the results found here can be viewed as a guide for the types of results to be expected in other spaces with wireless networks of similar extents.

The study provides a valid comparison of commonly available wireless positioning services on consumer mobile devices that is of benefit for those investigating possibilities of indoor positioning who wish to develop applications using such technologies. Whilst the positional accuracies found are not suitable for purposes as outlined here, they may be for others and the results presented here serve as a guide for other locations and applications.

REFERENCES

References from Journals:
Ishikawa, T, Murasawa, K, Okabe, A, 2009, 'Wayfinding and art viewing by users of a mobile system and a guidebook', In *Journal of Location Based Services*, Vol. 3(4), pp 277 – 293.

Landau, S, Wiener, W, Nagshineh, K and Giust, E, 2005, 'Creating Accessible Science Museums With User-Activated Environmental Audio Beacons (Ping!'), In *Assistive Technology 2005*, vol. 17, pp 133–143.

Tsai, C-Y, Chou, S-Y, & Lin, S-W, 2010, 'Location-aware tour guide systems in museums' in *Scientific Research and Essays*, vol. 5(8).

Zandbergen, P, 2008, 'Positional Accuracy of Spatial Data: Non-Normal Distributions and a Critique of the National Standard for Spatial Data Accuracy', *Transactions in GIS*, vol. 12(s1), p. 103-130.

Zandbergen, P, 2009, 'Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning', *Transactions in GIS*, vol. 13(s1), p. 5-26.

References from Other Sources:
Bensky A 2008 *Wireless Positioning: Technologies and Applications*, Artech House, Boston, MA.

Baldwin, T. & Kuriakose, K, 2009, 'Cheap, Accurate RFID Tracking of Museum Visitors for Personalized Content Delivery' In *Museums and the Web 2009: Proceedings*.

Bay, H, Fasel, B, & Van Gool, L, 2005, 'Interactive museum guide', in *The Seventh International Conference on Ubiquitous Computing UBICOMP, Workshop on Smart Environments and Their Applications to Cultural Heritage*, September 2005

Chan, L, Hsuz, Y, Hung, Y, & Hsu, J 2005, *Orientation-Aware Handhelds for Panorama-Based Museum Guiding System.*

Fabian Bohnert and Ingrid Zukerman: "A Computer-Supported Methodology for Recording and Visualising Visitor Behaviour in Museums". In *Adjunct Proceedings of the First and 17th International Conference on User Modeling, Adaptation, and Personalization (UMAP-09)*, Trento, Italy, 2009.

Hermersdorf, M, 2006 'Indoor positioning with a WLAN access point list on a mobile device' *WSW06 at SenSys06*, Boulder, Colorado, USA.

Kim M, Fielding J, and Kotz, D, 2006, 'Risks of using AP locations discovered through war driving', In *Proceedings of the Fourth International Conference on Pervasive Computing*, Dublin, Ireland: pp 67–82.

LeMarca, A, Hightower, J, Smith, I & Consolvo, S, 2005, 'Self-Mapping in 802.11 Locations Systems', In *Proceedings of Ubicomp 2005*, Tokyo, Japan, Sept 2005

Savidis, A, Zidianakis, M, Kazepis, N, Dubulakis, S, Gramenos, D & Stephanidis, C, 2008. 'An Integrated

Platform for the Management of Mobile Location-aware Information Systems', In *Proceedings of Pervasive 2008*, Sydney, Australia, pp. 128–145.

References from websites:
AMNH, 2010, *Explorer, the new way to find your way*, viewed 07/09/10, http://www.amnh.org/apps/explorer.php

Museum of London, 2010, *You Are Here App*, viewed 07/09/10, http://www.museumoflondon.org.uk/MuseumOfLondon/Resources/app/you-are-here-app/index.html

Rosum, 2008, *Rosum TV Technology*, viewed 20/08/10, http://www.rosum.com/rosum_technology.html

AN ONTOLOGICAL ROUTE DETERMINATION SERVICE

O. Akcay

ITU, Civil Engineering Faculty, 34469 Maslak Istanbul, Turkey - akcayoz@itu.edu.tr

KEY WORDS: Semantic web, Ontology, Geospatial data, Route, Short path

ABSTRACT:

During last two decades, supply of data sources has been increased with development of network and internet technology. Many GIS applications need to use geospatial data that has been supplied from various sources. The main problems are interoperability between different data models and to make data machine understandable. Neither Web Feature Services (WFS) nor relational spatial data bases are able to overcome these problems. Some semantic web approaches have been applied to obtain geospatial services that are able to provide interoperability and possibility of inference for geospatial data. This paper defines a method which provides an ontological short path algorithm for geospatial data. Therefore semantic data model enables an interoperable path finding tool without using traditional computing algorithms such as Dijkstra's Algorithm. Ontological short path algorithm has been compared with traditional Dijkstra's Algorithm. Consequently, the similar results have been obtained with Dijkstra's Algorithm, though ontological short path algorithm has not used coordinate information explicitly.

1. INTRODUCTION

During last two decades, supply of data sources has been increased with development of network and internet technology. Many GIS applications need to use geospatial data that has been supplied from various sources. The main problems are interoperability between different data models and to make data machine understandable. Neither Web Feature Services (WFS) nor relational spatial data bases are able to overcome these problems. Some semantic web approaches have been applied to obtain geospatial services that are able to provide interoperability and possibility of inference for geospatial data.

Semantic Web needs Ontology Web Languages providing knowledge representation on the World Wide Web. Ontologies and their languages are important in order to make web resources machine understandable also known as the Semantic Web [1]. OWL DL (Ontology Web Language with Description Logics) is a one type of OWL produced by the W3C Web Ontology Working Group. OWL DL provides maximum expressiveness without losing computational completeness [2]. OWL DL an extension of Description Logics is equivalent to DL *SHOIN(***D***)* logic [3]. In *SHOIN*(**D**), *S* stands for the DL *ALC* [4] also known as propositional modal logic $K_{(m)}$ extended with transitive roles, *H* stands for role hierarchies, *O* stands for nominals, *I* stands for inverse roles, *N* stands for unqualified number restrictions and **(D)** stands for datatypes.

In geosciences, semantic data has been considered as key notion in order to provide interoperable systems. Therefore many efforts which obtain ontological geospatial data from non-semantic one, has been made by geospatial data experts. [5] explained an ontological based discovery of geospatial data. Setting up semantic structure is necessary for not only data mining but also geoprocessing services. Geoprocessing services should have some properties and standards according to [6]. [7] and [8] discussed about semantic geospatial services and their chains. To provide ontological approach for traditional geospatial data, [9] proposed an rdf-based ontological interface for transportation data. WFS queries were mapped to rdf format. [10] explained ontological tuples to obtain semantic data. A match algorithm was applied between knowledge tuple and query tuple so as to find the most appropriate data.

Many short path algorithms which have different computational features have been developed [11]. Their computational performance of an algorithm varies depending on the network input. Spaito-temporal algorithms are also an important field for short path routing [12]. However, short path algorithms ignore ontological approaches for artificial intelligent systems.

This paper defines a method which provides an ontological short path algorithm for geospatial data. Therefore semantic data model enables an interoperable path finding tool without using traditional computing algorithms such as [13]. Ontological short path algorithm has been compared with Dijkstra's Algorithm. Consequently, the similar results have been obtained with Dijkstra's Algorithm, though ontological short path algorithm has not used coordinate information directly.

2. SEMANTIC GEOSPATIAL RELATIONS

For regions, a logic based RCC8 relation has been explained by [14]. [10] extended regional relations to points and lines. In [15], regional subsumption and neigbourhood property have been used in order to inference closest target places. [16] proposed Attribute Relational Graph (ARG). ARG described direction relations between geographical entities such as north, west and southeast.

In this paper, application area of road network has been divided to 3012 grid regions (Figure 1). The dimension of total area is 1.2kilometers x 17.7kilometers. Seven ontological property and three ontological class has been used in knowledge base (Table 1).

Before running ontological short path algorithm, the aim is to determine regions which are start and end position of the path. Let us assume that user is looking for the shortest path between Building 10 and Building 25.

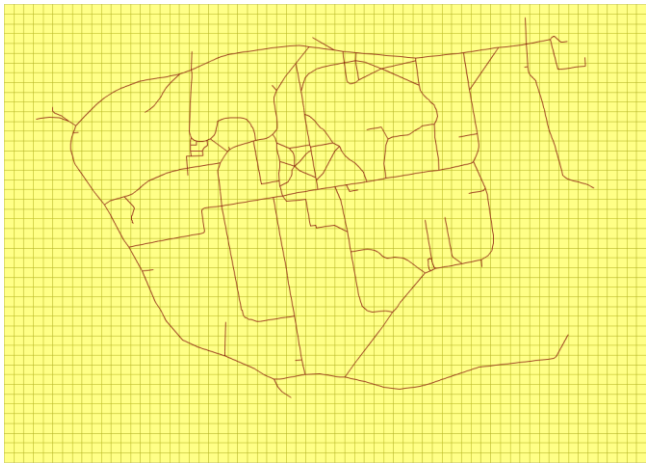


Figure 1. Square regions and road network

In two steps, start region has been obtained ontologically as follows:

OWL Class	Instances of the class	OWL Property	Property between
Building	B1, B2, … B45		
Region	R1, R2, … R3102		
Road	L1, L2, ...L195		
		hasEastNeighbour	Region and Region
		hasNorthNeighbour	Region and Region
		hasSouthNeighbour	Region and Region
		hasWestNeighbour	Region and Region
		hasExitIn	Building and Region
		hasLine	Building and Road
		İncludes	Region and Road

Table 1. OWL classes and properties used in knowledge base.

Step 1: Retrieving roads which connects start building.

Query for the building:
(retrieve (?road-connects-building)
(|http://www.semantic.org/ontologies/11/01/test.owl#B43|
?road-connects-building
|http://www.semantic.org/ontologies/11/01/test.owl#hasLine|))

Answer for the query:
(((road-connects-building
|http://www.semantic.org/ontologies/11/01/test.owl#L196|)))

According to reply of semantic query 1, "L196" is the line connects start building "B43". Figure 2 depicts the semantic relation between the building and the line.

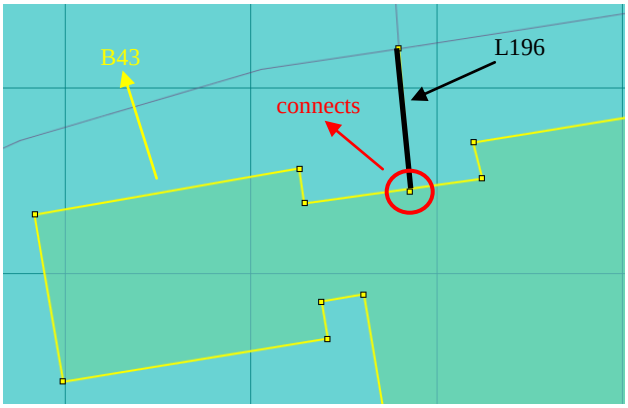


Figure 2. Semantic relation between the building and the line

Step 2: Retrieving grid regions which intersect roads determined in the step 1.

Query for the region:
(retrieve (?region- includes-road)
(?region- includes-road
|http://www.semantic.org/ontologies/11/01/test.owl#L196|
|http://www.semantic.org/ontologies/11/01/test.owl#includes|))

Answer for the query:
(((?region- includes-road
http://www.semantic.org/ontologies/11/01/test.owl#R1933|)))
(((?region- includes-road
http://www.semantic.org/ontologies/11/01/test.owl#R1999|)))

According to reply of semantic query 2, "R1933" and "R1999" are the grid regions include or intersect the line "L196" (see figure 3).

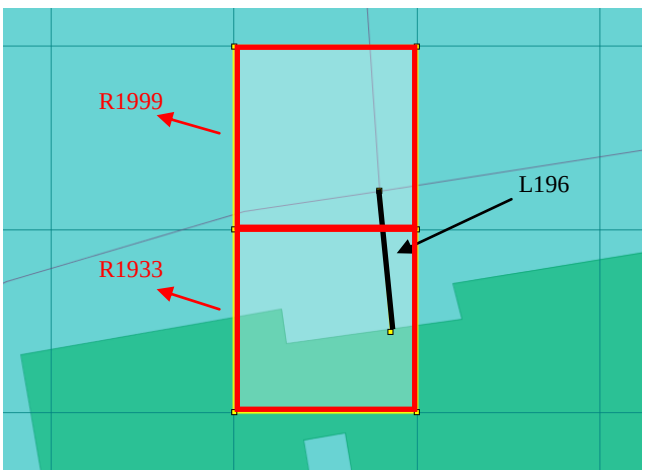


Figure 3. Semantic relation between the road and the region

Therefore start region (R1999) has been determined by two semantic questions. End point of the path can be determined by the same way that start point has been obtained.

3. SEMANTIC PATH FINDING ALGORITHM

Algorithm for the semantic short path algorithm

1 **input** owl file of knowledge base
1 **input** start region, end region
2 **for** all neighbours of start and end
3 retrieve only neighbours share the same road instances
4 iterate step 2
5 **end for** until any regions intersects from departure and arrival
6 **output** define shortpath route in geospatial database

Semantic short path algorithm uses knowledge base file (owl file) as input. As indicated in line two, "for loop" look for neighbors of start region and end region at the same time. The neighbor which is retrieved includes the same road instance with the start region. For example, in figure 4, start region has four neighbors having common road with start region, however south of the start region has been retrieved in the preceding steps as the exit region of the building. The same process is done for end region. Then from two directions (start and end) all neighbors are obtained until one intersection is determined from two directions (see figure 5).

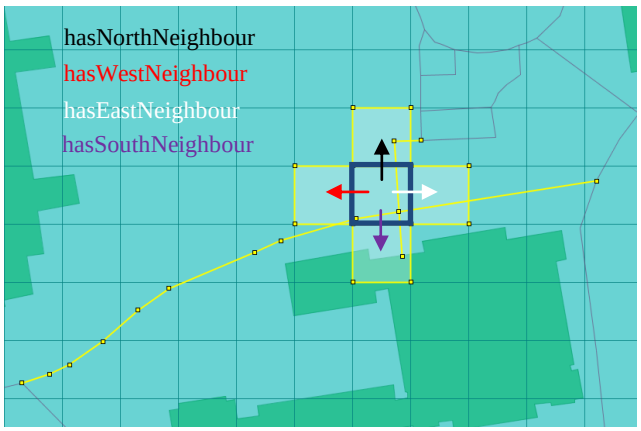


Figure 4. Semantic regional directions for the possible paths

Figure 5. Retrieved short path road

4. IMPLEMENTATION

Semantic short path algorithm needs a knowledge base as an input. To prepare owl input file from a geospatial database, an application schema should be composed according to [17]. In this work, [18] has been used as the application schema. Then, the database which is compatible with application schema encoded to owl file so as to provide input of algorithm. In Java 2SE, semantic short path algorithm has been implemented. RacerPro server has been used as description logic engine and owl inference service. Geoserver has been used as Web Feature Server (WFS) and Web Map Server (WMS). The results have been compared with Dijkstra's Algorithm. Figure 6 depicts both semantic short path and Dijkstra's algorithm results.

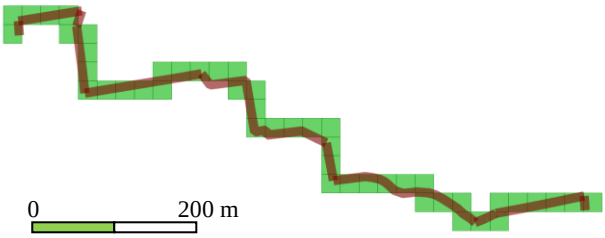


Figure 6. Semantic short path and Dijkstra's algorithm results

5. CONCLUSIONS

In this paper, a semantic short path algorithm has been implemented. The results show that the success of the semantic algorithm depends on the size of the grid regions. The smaller regions produce the more precise short paths. However small grid size naturally raises number of region. High region numbers causes longer computing time.

Another problem which is encountered is to produce ontological instances. A special tool has been developed with Java programming language in order to project spatial data to semantic structure. The tool has been done especially according to the database in this research. To use it again, it should be adapted for different database applications.

REFERENCES

A. Gomez-Perez and O. Corcho, "Ontology Languages for the Semantic Web," IEEE Intelligent Systems, vol. 17, no. 1, pp. 54-60, Jan./Feb. 2002.

M.K. Smith, C. Welty, D.L. McGuinness. OWL Web Ontology Language Guide. http://www.w3.org/TR/owl-guide/ 2004.

I. Horrocks, "Description logics in ontology applications," in KI 2005: Advances in Artificial Intelligence, 28th Annual German Conference on AI, Lecture Notes in Computer Science, (U. Furbach, ed.), p. 16, Springer, vol. 3698, 2005.

F. Baader, D. Calvanese, D. McGuinness, D. Nardi and P. Patel-Schneider, Editors, The Description Logics Handbook: Theory, Implementations, and Applications, Cambridge University Press, Cambridge 2003.

M. Lutz and E. Klien. “Ontology-based retrieval of geographic information,” International Journal of Geographic Information Science, Vol. 20(3):233–260, 2006.

ISO/TC 211, ISO19119:2005, Geographic Information – Services.

Lutz M 2007 Ontology-based descriptions for semantic discovery and composition of geoprocessing services. GeoInformatica 11*:* 1–36

P. Yue , L. Di , W. Yang , G. Yu and P. Zhao "Semantics-based automatic composition of geospatial web service chains", Comput. Geosci., vol. 33, pp. 649 2007.

T. Zhao, C. Zhang, M. Wei, Z. Peng, “Ontology-Based Geospatial Data Query and Integration”, Journal of Geographical Information Science, vol. 5266, 2008, pp. 370-392

C. Zhang, T. Zhao, W. Li, J. P. Osleeb. Towards logic-based geospatial feature discovery and integration using web feature service and geospatial semantic web. International Journal of Geographical Information Science, Volume 24, Issue 6 June 2010 , pages 903 – 923.

F. Benjamin Zhan , Charles E. Noon, Shortest Path Algorithms: An Evaluation Using Real Road Networks, Transportation Science, v.32 n.1, p.65-73, January 1998.

B. George, S. Kim, and S. Shekhar. Spatio-temporal network databases and routing algorithms: A summary of results. In *SSTD*, 2007.

E.W. Dijkstra, A note on two problems in connection with graphs, Numerische Matematik. 1 (1959) 269-271.

Randell, DA, Cui, Z and Cohn, AG., A spatial logic based on regions and connection. Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning. 1992.

Akcay O., Altan O., "Spatial relations and inferences for context aware visualization", 05/2010, p. 484-486, Joint International Conference on Theory, Data Handling and Modelling, Hong Kong, Int. Arc. of Photog., Remote Sensing and Spa. Inf. Sci.

W. Liu, H. Gu, C. Peng, D. Cheng, Ontology-based Retrieval of Geographic Information, 2010 18th International Conference on Geoinformatics, Beijing, 18-20 June 2010

ISO/TC 211, ISO19118:2005, Geographic Information – Encoding.

INSPIRE Data Specification on Transport Networks – Guidelines, 2010

ECOTOURISM DEVELOPMENT AND SECURITY RESTRUCTURING: A GI BASED PLANNING FOR PEACEFUL DISSUASION OF ANARCHISM IN FOREST PROVINCES OF INDIA

Abhisek Chakrabarty*

*Lecturer, Dept. of Remote Sensing & GIS, Vidyasagar University - Midnapore, W.B. India.
Principal Investigator, UGC Minor Research Project: 37-651/2009 (SR)
abhisek@vidyasagar.ac.in, destination_abhisek@yahoo.co.in

KEY WORDS: GIS, Ecotourism Planning, Security Restructuring, Anarchism Dissuasion, Sustainable Development

ABSTRACT:

Ajodhya Hill in Purulia district of West Bengal is a treasure house of natural beauties, but hilly terrain and thick forest cover have made many parts of this region inaccessible by road and perpetual agricultural drought over decades had weakened the economy of the area. Taking advantage of this physical and economic handicap, an organized group of social and political activists called Left-Wing Extremists (LWE), perpetrating violence and keeping the people of this region under threat. Non-cooperation of local population being the prime obstacle, government security forces with all its intelligence network and muscle power had not been utterly succeed to stop these activities.

This UGC (University Grant Commission - Govt. of India) sponsored *'Minor Research Project'* plans for economic and social stabilization of this region through the promotion of alternate economic practices for underprivileged forest dwellers and involving them in anti-terrorism operations. Taking into account the severity of physical environment and minimum skill level of tribal people, implementation of *'Ecotourism'* in these forest villages seems to be the best income-generating activity and ecologically permissible too. Due to essentially spatially distributed nature of tourism and terrorism related data and need of various types of spatial and statistical analysis GIS has proven to be a successful means in this study and based on spatial, non-spatial and attribute data overlay ('*Weighted Sum Overlay Analysis'*- ARC-GIS 9.2), ecotourism potential zones were identified. The inputs in the form of arc-coverages were assigned relative weightages according to their influence/importance in ecotourism development. Cadastral level action plan maps have been prepared for ecotourism infrastructure development and sustainable land use practices. Spatial database created on last ten years extremist movements and terrorist attacks to identify the spatial pattern, association and causes of vulnerability of the hot spots. Lastly spatial decisions have been made for allocation and relocation of police out posts, military camps and local participatory groups for fast information transfer and rapid action against any kind of social disorder.

1. INTRODUCTION

In contemporary world extremist movements and terrorism are the biggest threat to the human civilization. In most of the cases the birthplace of these extremist groups are the most backward and inaccessible part of a nation, where physical austerity has impede the development of agriculture, industry and transport network and made the region economically and socially backward. Thus poverty and ignorance among local population sometimes help the extremist groups to prosper more rapidly and effectively. To get the support and involvement of the local people in anti-terrorism operations, the administration first should strive to alleviate poverty of the region by arranging income-generating programs. Economic uplift automatically brings social awareness, and people spontaneously act against any kind of anarchism, which may come across the way of their earnings.

The forest provinces of Indian Plateau are full of natural attractions but the physical environment of this region is not suitable for intensive agriculture. The unskilled tribal people of these regions are also not preferred by the modern industry. Thus considering their low skill level as well as the environmental regulations, implementation of 'ecotourism' in forest villages may be the most suitable income-generating activity. Since 'ecotourism' has been defined by the Ecotourism Society in 1992 as "purposeful travel to natural areas to understand the cultural and natural history of the environment, not altering the integrity of the eco-system, while producing economic opportunities that make the conservation of natural resources financially beneficial to local citizens" (ref[9]). The volume of manpower could be engaged in tourism related activities is one of the highest in the service sector, which can create a wide range of job opportunities for millions of people with minimum skill level.

2. STUDY AREA

Ajodhya Hill (23°05′32″N-23°20′30″N, 85°55′00″E-86°14′20″E) in Purulia District of West Bengal is a part of the *'Jungle Mahals'*, i.e. tropical dry deciduous forest of Chotonagpur Plateau Region and mostly inhabited by tribal population. 176 Mauzas (villages) of Jhalda-I, Jhalda-II, Arsa, Baghmundi and Balarampur Block of Purulia share the 408.53 sq.km hilly tracts of Ajodhya (Fig-1). Undulating topography and dense forest cover of this region are responsible for its surface inaccessibility and thus lacking in medical and educational facilities.

Though the climate of this region is not very harsh (annual average rain fall is 1286 mm, with annual mean temperature 26ºC) but infertile laterite soil (originated from granite-gneiss of oldest precambrian or archean formations) with high evaporation and infiltration losses caused agricultural drought, which accumulated over years and had damaged the economy of the area.

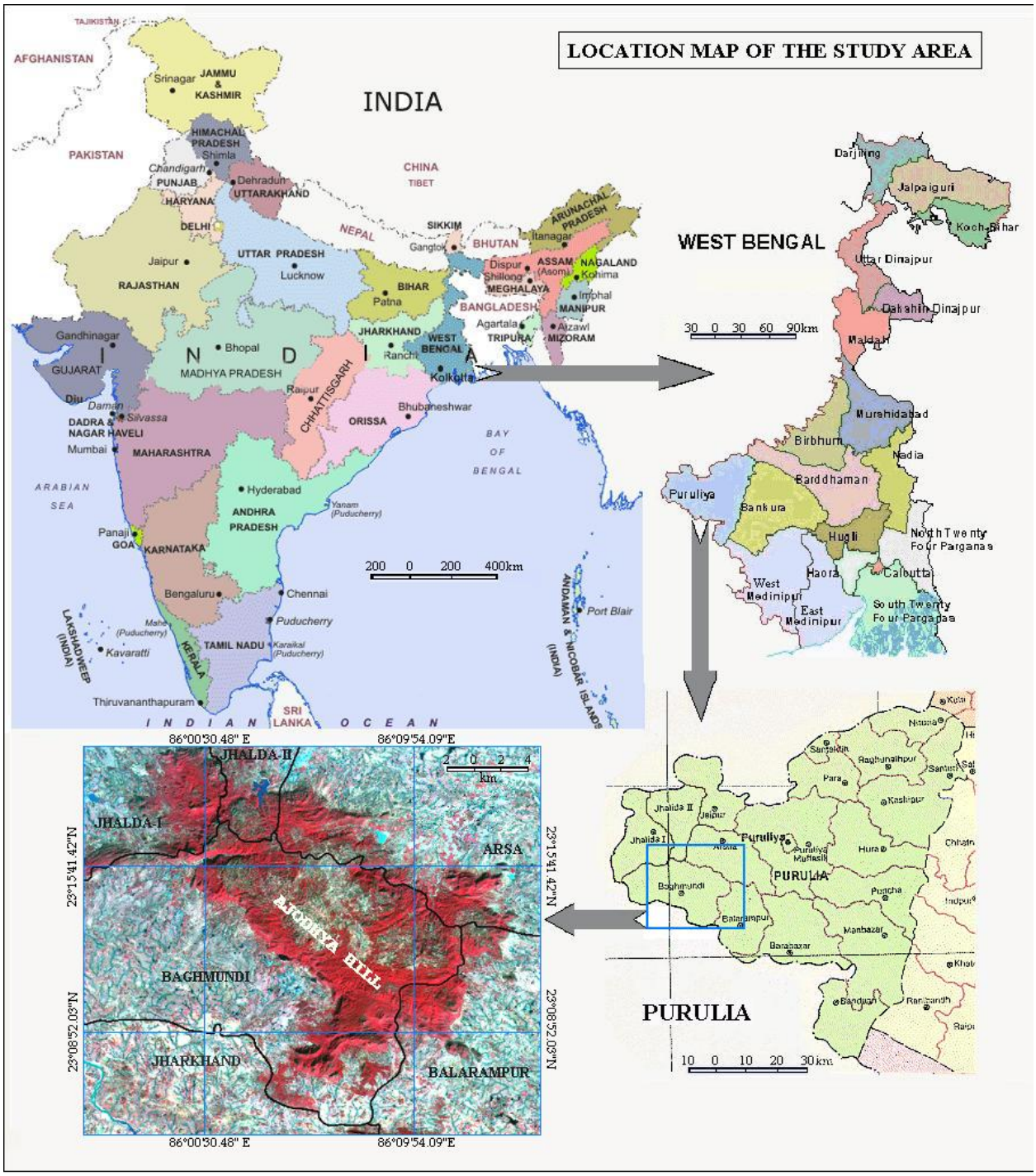


Figure 1. Location map of the study area consisting of 176 Mauzas (villages) of Jhalda-I, Jhalda-II, Arsa, Baghmundi and Balarampur Block of Purulia District, West Bengal, India

In "Integrated Mission for Sustainable Development"-1993 Govt of India identified 152 districts of India as backward district, and Purulia District was one of them (ref[1]). Making use of this physical and economic constrain, an organized group of social and political activists called Left-Wing Extremists (LWE), compelling local people to take part in their anti government insurgency. Left-Wing Extremism (LWE) was described by Chief Minister Buddhadev Bhattacharjee in 2005 as "plagued by the collapse or absence of rural governance" (ref[10]). Aside these socioeconomic hostilities, Ajodhya Hill is blessed with natural marvels. It has a blend of steep mountains, splendid waterfalls, dense forests with her wildlife beauties and huge water bodies (reservoirs). In the year 2006 Ajodhya hills has been declared as "*Conservation Reserve"* at state level by the State Wildlife Board (ref[2]). There is a huge potentiality of development of ecotourism in this 'only hill station' of South Bengal but inadequate infrastructure for transportation and accommodation, and dominance of Left-Wing Extremists (LWE) in Purulia district, have made this region lagging behind other tourist destination of South Bengal.

3. OBJECTIVES

Planning for ecotourism development and security restructuring as the sole objective, the present study is undertaken with the following intents:

i) Study the present physical and socioeconomic condition of Ajodhya Hill along with existing tourist spots, tourism infrastructure and annual tourist flow.
ii) Identification of potential ecotourism sites in the study area based on spatial, non-spatial and attribute data analysis.
iii) Creation of spatial database on extremist movements and terrorist attacks in last ten years to identify the spatial pattern, association and causes of vulnerability of the hot spots and to make an inventory of existing government security infrastructure to assess their positional accuracy and adequacy in combating terrorism.
iv) Planning for direct involvement of the local the people in the tourism sector and spatial decision making on allocation and relocation of police out posts, military camps and self-protection groups to have an administrative control over the entire region.
v) Generation of cadastral level action plan maps for ecotourism infrastructure and security infrastructure development as well as sustainable land use practices.

4. METHODOLOGY

Due to essentially spatially distributed nature of tourism and terrorism related data and need of various types of spatial and statistical analysis, GIS applications have a great relevance in this study.

Keeping in mind the basic thirsts of an ecotourist, *ecotourism potential* (Ep) sites were selected based on nine criteria: high *elevation* (El) and high *relative relief* (Rr), dense to moderate *forest cover* (Vd), proximity to *water bodies* (Wb), waste land or forest-fringes as desired *land use* (Lu) very low *population density* (Pd), *road connectivity* (Rc), *food and lodging facilities* (Fl) and minimum 20 hectares *level ground* (Lg) for ecotourism infrastructure development.

Mathematically, this can be expressed as:

$Ep = f(Rr, EI, Vd, Wb, Lu, Pd, Rc, Fl, Lg)$

These criterions were taken as the parameters to evaluate the areas of high ecotourism potential (Fig-2). For this purpose a '*Weighted Sum Overlay Analysis'* method was adopted (ref[6]). The input in the form of ARC/GIS coverage were assigned relative weightage in accordance to their influence/importance in decision-making based on expert opinion and each class in the individual coverage were ranked from 10 to 1 according to its potential of being or for being developed for ecotourism (Table-1). In this context we can say:

$Ep = \sum WiCVi \quad [With \sum Wi = 100]$

where Ep is ecotourism potential map value, Wi is the significance value of each theme that is theme weight and CVi is the grade value of individual class of a particular theme that is class weight.

Assigning the theme weight, the above equation can be written as:

$Ep = (27*CVRr) + (20*CVWb) + (17*CVLu) + (14*CVEI) + (13* CVPd) + (9*CVRc)$

Sl No.	1		2		3		4		5		6	
Theme Layer	Relative Relief		Distance from Water Body (Water Body Buffer zones)		Land Use		Absolute Relief		Population Density		Distance from Motorable Roads (Road Buffer Zones)	
Theme weight (%)	27		20		17		14		13		9	
Classes With Class weight (1-10)	>120m	10	<1km	10	Open Forest	10	>470m	9	<100/sqkm	9	<1 km	8
	80-120m	8	1-2 km	9	Scrub Land	9	220-470m	6	100-200	8	1-2	7
	40-80m	6	2-3 km	7	Dry Fallow	9	<220m	4	200-300	5	2-3	6
	<40m	4	3-4 km	3	Dense Forest	8			300-600	2	3-4	5
			4-5 km	2	Agricultural fallow	2			>600/sqkm	1	4-5	4
			>5km	1	Others	1					>5 km	3

Table 1. Theme weight and class weight of respective theme layers assigned according to their influence/importance in ecotourism development.

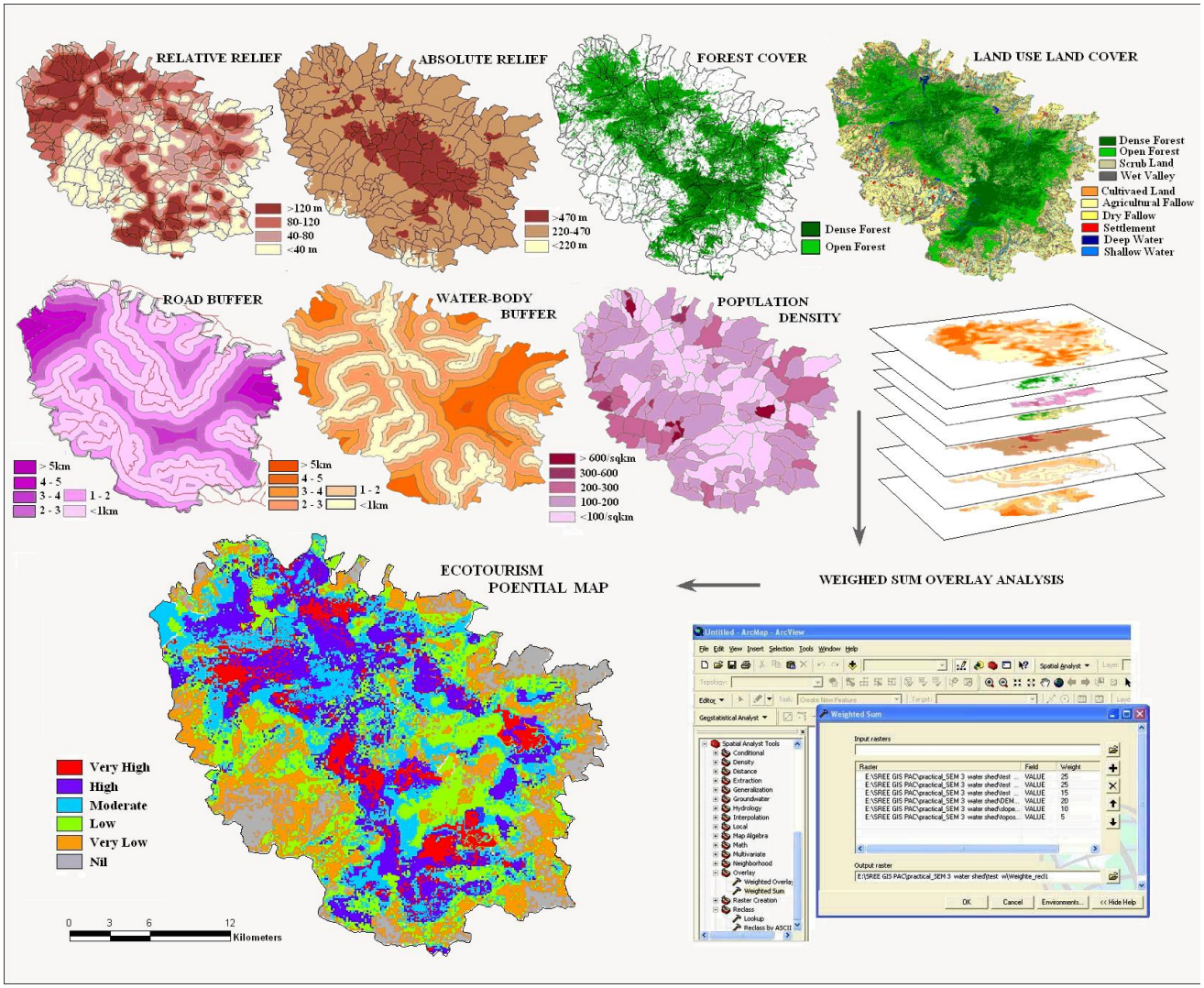


Figure 2. Ecotourism potential zone identification by "Weighted Sum Overlay Analysis" of different thematic layers (e.g. absolute relief, relative relief, forest cover, lad use, land cover etc)

Spatial information on last ten years extremist movements were collected from daily newspapers and their respective websites; crime records from the office of the Superintendent of Police - Purulia; and primary survey in local political offices and forest villages. Data were plotted on the base map as point layer. Probable routes of LWE operation were detected by examining the pattern of occurrences of crime and relating them with the land use / land cover information. Location of police stations, police out posts, paramilitary camps and their jurisdiction area were also plotted on the base map to determine their positional accuracy and adequacy in combating seditious activities. Zones of concentration of extremist incidents were used to predict likely sensitive points/areas (hotspots) and the causes of their vulnerability were analyzed from *'Geographical Profiling of Crime'* (ref[7]) and *'Proximity Analysis'* from security camps, motorable roads, and state border line. Lastly an action plan has been prepared on ecotourism infrastructure development, allocation and relocation of police out posts and paramilitary camps; and creation of self protection groups to support local law enforcement agencies.

5. RESULTS AND DISCUSSION

Annual tourist influx in Ajodhya is much lower than the other tourist spots of South Bengal but the positive aspect was, a steep increase in tourist flow observed each year from 2001 to 2006. Construction of Purulia Pumped Storage Project (PPSP), the engineering marvel, was the main attraction. From 2001 to 2002 tourist flow increased 363%; in 2003 it increased 47%; the increase was low from 2003 to 2004 only 10%; but in 2005 it was 125%; and another 65% more tourists visited the place in the year 2006. Tourist flow was in its peak during 2007. 89026 persons including 634 foreigners visited Ajodhya Hill this year, but due to ever increasing LWE activities from 2008 tourist flow gradually decreased each year and in 2009 it came down to 80245 persons only (ref[3]).

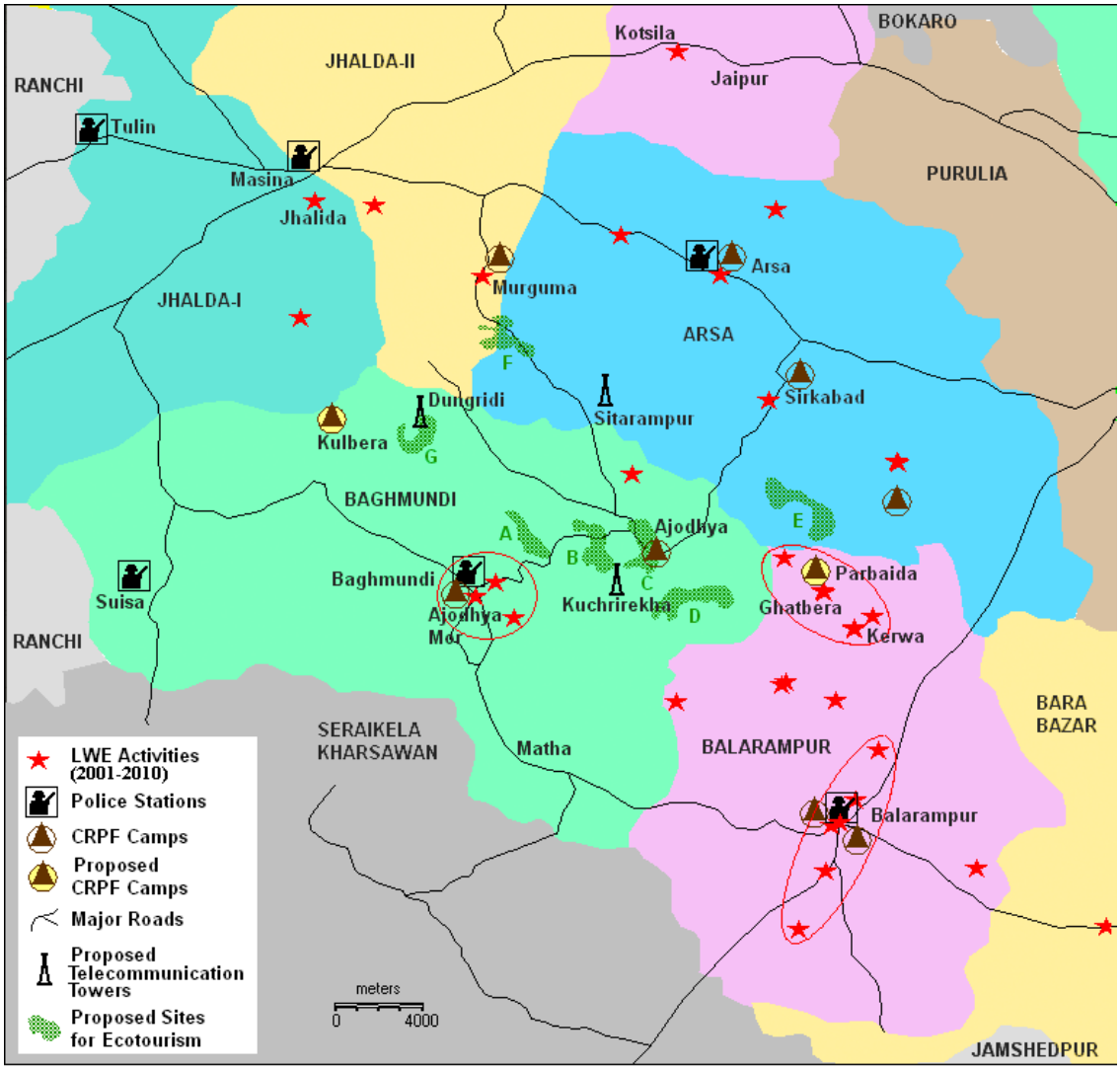


Figure 3. Proposed sites for ecotourism development, LWE activities hotspots and government security infrastructure and around Ajodhya Hill

From '*Weighted Sum Overlay Analysis'* seven (7) spots on the hill are recognized as appropriate for ecotourism development (Fig-3). They are named as Zone-A (Dulgubera - Purna Tanrpania), Zone-B (Bandhghutu - Bidyajara), Zone-C (Ajodhya - Kuchrirekha), Zone-D (Kamarjara -Goberia), Zone-E (Dhanchatani-Lukuchatani), Zone-F (Bamni-Lakshmipur) and Zone-G (Inchakata–Burda). It is noteworthy that out of seven zones, Zone A, B, C, D, G are inside Baghmundi block, Zone-E in the boarder of Balarampur and Arsa Block and Zone-F in the boarder of Arsa and Jhalda Block. Cadastral level action plan maps have been prepared for sustainable land use practices and ecotourism infrastructure development. Ecotourism infrastructure involves tourist cottages/rest houses, green hotels and restaurant, public convenience facilities, elephant watchtower, tourist information centre, conveyance facilities, tourist guide map, public convenience facilities, detail map of the ecotourism destination, tourist circuit map to show its link to other place of tourist interest, do and don't board, medical aid facilities, communication facilities etc. (ref[4]). Strategies related to 'Participatory Forest Management' and 'Community Tourism' are proposed to involve local people in ecotourism planning and management (ref[8]). "*Environmental Approach*" (ref[5]) was adopted for sustainable land use planning. Proposals are given on what are the areas should be afforested immediately, where expansion of settlement and cultivation should be restricted, instead of large-scale cultivation, thrust area will be forestry and forest based economic activities like agro-forestry, hortipasture, floriculture, sericulture, aquaculture, horticulture, animal husbandry etc. The yields will meet the demand of tourists as well as the local people.

Analyzing the spatial pattern of last ten years (2001-2010) extremist movements (i.e. attacking police camps and police vehicles with grenades and landmines and conducting guerrilla warfare inside the forests; taking hostages and assassinating local political leaders in the dark; arson and blasting on railway tracks; abduction for ransom and running extortion racket etc.), mainly three (3) hot spots are identified: i) Balarampur Town, along the Purulia Chandil Road. Here all incidents happened not far away from this main road and we can see a linear pattern of occurrences. ii) Ghatbera-Kerwa zone where seditious activities are clustered within three villages. iii) Ajodhya More, here most of the cases happened along the main road near Bagmundi town (Fig-3). Surprisingly most of the incidents occurred within five km radius from the police camps and in some cases on the police camps. The border line of Jharkhand State is not more than ten km from any part of these regions thus after operation escape is trouble-free. Cross border forest tracks and gullies revealed from high resolution IKONOS data (GCR-1.0m) closely matches with their routes of operations.

At present there are six (6) Police Stations, eight (8) CRPF Camps and one (1) Traffic Out Post / Check Post in and around Ajodhya Hill. But two strategically important points, Kulbera river dam near Zone –G and Parbaida near Zone-E are still unprotected. Therefore two more CRPF cams are required in these points. One Traffic Out Post/Check Post is also necessary at Chak Keryari of Jhalda-I. To bring the entire area under telecommunication network coverage, installation of three new mobile towers (at Kuchrirekha, Dungridih, and Sitarampur village) in 'no-network coverage' areas of Ajodhya Hill have been proposed (Fig-3). It will help in fast information transfer and rapid action of law enforcement agencies. Active role of local administration (e.g. Gram Panchayet) is also solicited in formation of self protection groups among the villagers and involving them in antiterrorism campaigns.

6. CONCLUSIONS

In spite of many attempts taken by the central or state government for social and economic reform of the region, no action plan has become totally successful because of lack of spatial information and finding of the proper solution. But this research will be equipped with highly scientific and contemporary methodologies with reliable satellite data products. Geographical Information System (GIS) has already proven to be a successful means in the field of 'Ecotourism Planning' and GIS application also has relevance in spatial analysis of crime and terrorist movements, because one of the most invaluable tools available for effective crime fighting is spatial information. When the action plan will be materialized the downtrodden aborigines of the region will get the maximum benefit and security of the region both in terms financial and administrative means could be ensured. Thus this will be an unique venture, first of its kind in this region showing the way of economic and social transformation based on advanced technological know how. This model also can be applied in other less reproductive, rugged regions of the world where natural limitations are the root of economic, social and political turmoils.

REFERENCES

1>Anonymous (1996) Integrated Mission For Sustainable Development (Phase-I) Programme In West Bengal - An Overview. State Remote Sensing Centre, Department of Science & Teclmology and NES.Govt. of West Bengal, Calcutta - 700 091. Regional Workshop on Remote Sensing for Sustainable Development (1996) DST & NES Govt. of WB. Kolkata, In Collaboration with Dept of Space Govt. India. pp. 33-46.

2>Anonymous (2006), Ananda Bazar Patrika 10th Jun 2006. 6 Prafulla Sarkar St. Kolkata – 700001. (Editor – Abhik Sarkar). http:// www.anandabazar.com.

3>Anonymous (2010), Annual tourist (persons) inflow register, West Bengal**.** Department of Tourism, Government of West Bengal, 2-Brabourne Road, Kolkata - 700001. http://www.wbtourism.com

4>Banerjee U. K., Kumari Smriti, Paul S. K. and Sudhakar S. (2002) Remote Sensing and GIS based ecotourism planning: A case study for western Midnapore, West Bengal, India. www.gisdevelopment.net

5>De N.K and Jana N.C. (1994) The Land, Multifaceted Appraisal and Management. Sribhumi Publishing Co. Kolkata.

6>Fotheringham Stewar and Rogerson Peter (2002) Spatial analysis and GIS. Taylor & Francis Ltd. UK.

7>Krish Karthik (2003) Application of GIS in crime analysis and geographic profiling.

www.gisdevelopment.net

8>Obadih Bukenya James (1999), GIS for Ecotourism Development Decisions Making in Uganda's national parks of Africa.
http://www.rri.wvu.edu/pdffiles/bukenya2012.pdf

9>Panda Tapan K, Mishra Sitikantha and Parida Bivraj Bhusan (2004), Tourism Management – The Socioeconomic and Ecological Perspective 1st Ed. Orient Longman Pvt. Ltd. Hyderabad.

10>Routray Bibhu Prasad (2008) West Bengal: State Myopia, Maoist Consolidation
http://www.ocnus.net/artman2/publish/Analyses_12/West_Bengal_State_Myopia_Maoist_Consolidation.shtml

TEMPORAL SERVICES FOR SPATIAL DATA AND METADATA IN CIVIL PROTECTION CONTEXT.

Raffaele de Amicis, Giuseppe Conti, Federico Prandi, Alberto De Biasi

Fondazione Graphitech, via alla cascata 56/C, 38123 Trento Italy.
[raffaele.de.amicis,giuseppe.conti,federico.prandi,alberto.de.biasi]@graphitech.it

Commission VI, WG VI/4

KEY WORDS: GeoBrowser, Geospatial web architecture, Spatio-temporal information, Sensors web, Web services

ABSTRACT:

Civil Protection operators and Public Administrations, engaged in urban planning, resource and environmental management, clearly need spatio-temporal processing of GI to support decision-making in the fast pace. Providing access to harmonized data is only a first step towards providing adequate support to environmental management, which requires development of analysis and spatio-temporal data models functionalities. The work presented in this paper shows the first results of project BRISEIDE - BRIdging SErvices, Information and Data for Europe, a Pilot B, funded by the ICT Policy Support Programme (or ICT PSP), which tries to bridge the aforementioned gap, by building on top of existing SDIs to provide users with more complete and adequate data and processing tools capable to handle the time as a dimension. BRISEIDE develops OGC-compliant Web Processing Services (WPS) for spatial analysis and integrates them within existing open source frameworks. Spatio-temporal processing services are exposed via the web and are made available through compatible webGIS applications.

1. MAIN

Civil Protection operators and Public Administrations, engaged in urban planning, resource and environmental management, clearly need spatio-temporal processing of GI to support decision-making in the fast pace. Natural disaster monitoring and mitigation is characterized by frequently updated repositories as well as by provision of dynamic data. For these stakeholders being able to access with shortest timing to the most up-to-dated information is crucial.

Today, many of the elements typical of the concept of Digital Earth, i.e. a multi-resolution, three-dimensional representation of the planet, are not only available but also used daily by hundreds of millions of people worldwide, thanks to innovative ways to organize and present the data based on rapid technological advancements. 3D Geobrowsers can provide higher usability and interactivity that can facilitate exchange and dissemination of spatial information among stakeholders and government agencies. However, providing access to harmonized data is only a first step towards providing adequate support to environmental management, which requires development of analysis and spatio-temporal data models functionalities. The capability to add other geophysical data to the model simulations, such as earthquake locations or images from onsite video cameras is greatly enhanced, also in terms of user-friendliness, by the advent of Virtual Globes (Webley, 2010). Displaying of geophysical data in Virtual Globes is now becoming a common place, while the next evolution will see the possibility to display near-real-time data from multiple data sources (Webley, 2008). The visualization of geophysical data in four-dimensions (x, y, z and t) so far has been a time-consuming process with no common interface for different data sets.

In the perspective of the hugely populated European SDI, as envisioned by INSPIRE, it is increasingly important to be able to filter out geographic resources that not only refer to a given spatial location, but also match both the temporal coverage of interest and the temporal resolution suitable to the application purposes.

This should lead the GI community to reconsider current data models, taking into account time as one of the main variables.
Analysis of spatio-temporal data is especially challenging. It requires tools for representation and manipulation of all three aspects of the data: thematic (values of attributes), temporal and spatial. Due to increased speed, space, and performances of graphical displays, modern computers offer new opportunities to properly visualize the temporal variation of spatially referenced data.

The work presented in this paper shows the first results of project BRISEIDE - BRIdging SErvices, Information and Data for Europe, a Pilot B, funded by the ICT Policy Support Programme (or ICT PSP), which tries to bridge the aforementioned gap, by building on top of existing SDIs to provide users with more complete and adequate data and processing tools capable to handle the time as a dimension. BRISEIDE develops OGC-compliant Web Processing Services (WPS) for spatial analysis and integrates them within existing open source frameworks. Spatio-temporal processing services are exposed via the web and are made available through compatible webGIS applications.

The main objective of the project is to provide a framework for spatio-temporal management of geographic information, by developing geographic services able to consider the time as dimension.

Considering time extension in the geographical data set increases the dimension of the analysis traditional constrained to the space domain. It is possible to consider three issues related to the management of temporal information in spatial domain. The first is the possibility to retrieve information based on temporal information of the dataset; the second involves the capability to manage temporal information in order to generate analysis and process and, thirdly, the last deals with the

capability to simply represent the temporal dimension in a useful manner.

The methodology followed by the project is composed by three main parts, the first is the definition of the data and metadata model in order to manage the temporal information, the second is the deployment of the web services in order to support the proposed data model and third is to consume the services in a client application.

The possibility to retrieve or query datasets based on time properties is a fundamental requirement of the BRISEIDE framework. This involves an extension on the metadata implementation rules which define the four aforementioned categories. However these categories do not allow comprehensive management of the whole spectrum of temporal information. In order to improve the definition of time-related metadata, some authors have proposed a set of six recommendations (Dekkers, 2008). Several other studies have proposed to manage temporal series of data sets, and have discussed the need for metadata extensions in the temporal domain (Bordogna et al., 2009).

The BRISEIDE Metadata profile tries to address this by providing metadata model and applying it to datasets collected within the BRISEIDE SDI. The profile is not oriented to a specific type of dataset and for this reason it is generic and does not contain any specific component. The profile it will be composed by the core metadata for INSPIRE with some additional elements in order to describe the time information contained within the dataset.

The elements defined in the metadata profile allow discovery of dataset by using the temporal dimension. Specifically the mandatory elements identified for BRISEIDE profile are:
Identification.Information.Citation.date represents the date related to the resources coherent with respect to the dates that are included in the INSPIRE Metadata Implementation Rules (creation, revision, publication), IdentificationInfo.resourceMaintenance.maintenanceAndUpdate Frequency represents the frequency update interval,
IdentificationInfo.Extent.temporalElement represents the period covered by the specific resources and it stores the starting and the ending date, if existing.

These three elements let the user identify relevant resources by using the temporal dimension by providing information about the resource data, resource update and resource temporal extent. In particular the combined use of frequency update and temporal extend allows to cover the whole aspects of the temporal properties for the dataset used in civil protection field. For instance, considering a series of speed wind forecast given each day with a temporal horizon of five days and frequency of 1 hour. Using the frequency tag (1 hour) and several temporal elements which start at the day of the starting forecast and end at the day of the last forecast; the user requiring a given date can find all the intervals containing the required date and given the frequency can know the temporal granularity of the resource.

The reason of this simple choice is to find in first in the nature of the project, that is oriented into the integration of existing technologies: in fact extending the Metadata model adding some new elements can bring to some issue in the catalogue standard implementation, and the integration of existing infrastructure may became problematic. Another matter is related to the automatic ingestion of the data and metadata and a complex metadata model with many unusual tags may be challenge to automatically manage. Finally the final user who have to compile the metadata model should be facilitated to fill the information if these are not very complex.

After the definition of the metadata model the second step has been the definition of a data model which allows operating on spatial datasets which include the time dimension. In the majority of the datasets used for the project, the geographic information is represented by a raster image which can be considered as a complex structure describing a set of gridded datasets having the same domain and the same band layout. Typically these are the result of remote sensing observations, imagery data or model runs/executions.

The multi-dimensional raster data model, developed in the project, produces several 3D hypercubes for a predefined number of specific times, separated by a certain time period. Given a specific run time, a data hypercube is in generally speaking intrinsically multi-dimensional. Its spatial dimension can span over Longitude, Latitude, Time (various forecasts time) as well as Elevation. The basic idea is that such a multidimensional object can be built by wrapping instances of the basic 2D implementation of the GridCoverage interface and by associating to the proper value along the additional dimensions with respect to the canonical ones, i.e. Latitude and Longitude (or Easting and Northing). The model descripted will be provided by the framework through standard OGC services, like WMS and WCS, which are able to support additional dimension such as time and elevation.

Another sources of important information related to the project are the sensors as they have a spatial component, usually the sensor position in space, and a temporal component describing the acquisition time. In many civil protection scenarios the information measured by the sensor has a spatial component (for instance the displacement on a monitored landslide). In these circumstances the spatio-temporal data model is already defined as a temporal series and the aim of the work brought forward by BRISEIDE is, by using international standard like OGC Sensor Observation Service (SOS), to provide a spatio-temporal visualization of sensor data. However, in many other cases, this simple visualization is not enough and it is necessary to make use of complex diagrams in order to visualize the results of a given analysis, which can make the understanding more difficult.

GeoVisual Analytics (GVA) is a discipline that combines the benefits of data mining and information visualization within a geospatial context. GVA is capable to provide integrated visualization, by filtering and reasoning solutions to better support operators looking for design decision support (De Amicis et al. 2009). Through GVA tools, users can typical acquire visual cues that can help them formulate a set of viable models. The possibility to unleash the potential of GVA within web-based 3D geo-browsers is a central issue in the usability of Geographic Information (GI) and, in more general terms, to ensure better understanding of geographic-dependent phenomena.

The solution tested will be used to provide a temporal slice in order to navigate along the time dimension and visualize data on a 3D Geobrowser proportionally to the magnitude of the measurement in the specific time or period .The information visualization comprises the possibility to visualize both the data exposed through OGC web services and the measurements recorded by the monitoring system. These involve the capability to create graphs and diagrams based on the historical series.

The result of the project in this first phase is a series of dataset (in particular raster and Sensor data) exposed using standard services WMS WCS and SOS which allow temporal queries by the user. The client performing standard temporal request to the server is able to consume the services allowing to the user to navigate through spatio and temporal dimension.

REFERENCES

Bordogna, G., Bucci, F., Carrara, P., Pagani, M., Pepe, M., Rampini, A. Extending INSPIRE Metadata to imperfect temporal descriptions. *Article under Review for the International Journal of Spatial Data Infrastructures Research, Special Issue GSDI-11, submitted 2009-04-03*

R. De Amicis, G. Conti, B. Simoes, R. Lattuca, N. Tosi, S. Piffer, G. Pellitteri 2009. Geo-Visual analytics for urban design in the context of future internet. In: *International Journal on Interactive Design and Manufacturing (IJIDeM), March 2009.*

Dekkers M. 2008. Temporal Metadata for Discovery - A review of current practice, *Craglia M. (ed.), EUR 23209 EN, JRC Scientific and Technical Report*

Webley, P.W., 2008. Google Earth environment guide: The volcano tracker. Popular Science, July 2008, 66 – 67.

Webley P.W., 2010. Virtual globe visualization of ash aviation encounters, with the special case of the 1989 Redoubt KLM incident.*Computers&Geosciences.doi:10.1016/j.cageo.2010.02.00*

ACKNOWLEDGEMENTS

The project BRISEIDE has received funding from the EC, and it has been co-funded by the CIP-ICT Programme as part of the Competitiveness and innovation Framework Programme (http://ec.europa.eu/ ict_psp). The author is solely responsible this work which does not represent the opinion of the EC. The EC is not responsible for any use that might be made of information contained in this paper.

PROPOSAL FOR THE MANAGEMENT OF TEMPORAL AND SEMANTIC COMPONENTS OF GEOGRAPHIC INFORMATION

Willington Siabato [a], Miguel-Ángel Manso-Callejo [a]

[a] Technical University of Madrid, Mercator Research Group, Autovía de Valencia Km. 7.5, 28031 Madrid, Spain (w.siabato, m.manso)@upm.es

KEY WORDS: spatiotemporal reasoning, GIS, time, geosemantic, dynamic storage, spatial-temporal modelling, temporal reasoning

ABSTRACT:

This paper raises the issue of a research work oriented to the storage, retrieval, representation and analysis of dynamic GI, taking into account the semantic, the temporal and the spatiotemporal components. We intend to define a set of methods, rules and restrictions for the adequate integration of these components into the primary elements of the GI: theme, location, time (Sinton 1978). We intend to establish and incorporate three new structures into the core of data storage by using mark-up languages: a semantic-temporal structure, a geosemantic structure, and an incremental spatiotemporal structure. The ultimate objective is the modelling and representation of the dynamic nature of geographic features, establishing mechanisms to store geometries enriched with a temporal structure (regardless of space) and a set of semantic descriptors detailing and clarifying the nature of the represented features and their temporality. Thus, data would be provided with the capability of pinpointing and expressing their own basic and temporal characteristics, enabling them to interact with each other according to their context, their temporal structure, and the meaning relationships that could be eventually established. All of this with the purpose of enriching GI storing and improving the spatial and temporal analyses.

1. INTRODUCTION

Time and semantics are two broad, general concepts applicable to a considerable number of scenarios (physics, geology, grammar and geography, among others). Time seems to be bound to every performed action; and every one of these actions has a meaning, a sense or a way of being interpreted that may be described through semantics.

In the realm of the Geographic Information Sciences these concepts have evolved following different paths. For over 25 years, time in GIS has been an active research line (Ott and Swiaczny 2001, Peuquet 1984) with important theoretical and conceptual advances having been achieved. The time models developed for the temporal databases (see (Snodgrass 1990)) have chiefly influenced the trends followed for incorporation of temporal structures into GIS. Semantics in turn derives directly from the study of language and meaning; from a computational viewpoint, these concepts are framed within the Natural Language Processing (NLP), a line of the Artificial Intelligence dealing with modelling and processing of human language in which a lot of work has been carried on for over 60 years (Turing 1950). The study of meaning in the realm of Geographic Information (GI) and the Geosciences is widely related to these elements, giving way to specific working areas such as Geosemantics or Geographic Information Retrieval (Jones and Purves 2008).

These constructs have come together leading to initiatives such as the computational processing of temporal expressions, a research area which has aroused the interest of the academic community, as witnessed by the related multiple academic events which have been held (Brandeis University 2006, Brandeis University 2005, DARPA's Information Technology Office 2010, Dipartimento diInformaticai Comunicazione 2010, Pustejovsky and Mani 2003, Office 1995). In this regard relevant advances have been made: initiatives for the generation of temporal mark-up standards (International OrganizationforStandardization -ISO- 2007), annotation tools and systems (Pustejovsky and Mani 2003, James Pustejovsky et al. 2004, Mazul and Dale 2009), temporal annotation corpora (TimeML WG 2008b), annotation languages (Spatio Temporal MITRE 2009, TimeML WG 2008a) and assessment methods (Corporation 2004, Verhagen et al. 2007).

The analysis of temporal expressions allows placing data, facts and events on timelines subjectively, correlating and arranging them chronologically. With a plain temporal description there would be sufficient available elements to solve elementary questions such as When do events occur? How often are they updated? Or which one has occurred before or afterwards? In order to provide data with the ability to describe themselves and relate to one another, taking into account temporality criteria and their representation meaning, it would be necessary to enrich them with a structure enabling identification of their temporal characteristics, their context and their meaning. Such a structure should be computationally usable and should be based on standards to ensure the interoperability of the data enriched with semantic and temporal elements.

A result of this proposal would be the definition of a new storing structure which unlike the current model, would be made up of the three basic components (attributes, location and temporal reference) in addition to three layers to incorporate the semantic and temporal components, these layers are the core and innovation of this proposal (see Figure 1). Initially a file-type storing format is proposed; besides, the possibility of integrating the new concepts into a spatial database engine will be assessed according to the degree of progress of the research work.

This paper is structured as follows; part two presents basic and related concepts of time and geosemantic as well as the initial point of this research. Part three shows the problem that this research is intended to solve, it presents current models and some pros and cons. Part four describes in detail the proposal

and the research scope by giving the hypothesis, objectives, and the metamodel elements. Finally, part five enumerates the conclusion and the anticipated contributions to be gained from this research.

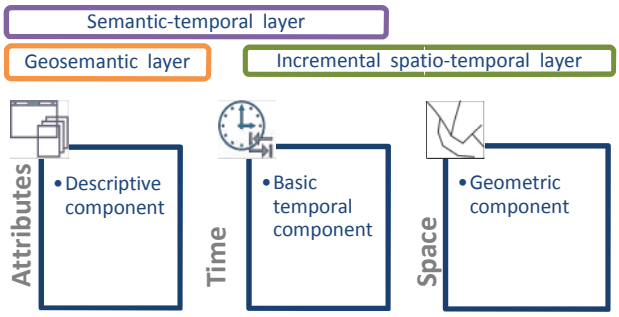


Figure 1. Basic components of the GI and new structures (layers) proposed

2. SOME RELEVANT CONCEPTS

Both time and space appear to be indivisible or at least deeply related to one another; they make up an abstract universal frame, and everything surrounding us is contained therein. In this regard, every geographic feature and geographic entity belongs to that frame, therefore also its characteristics, properties and all possible analyses developed on it. In this context, Couclelis (Couclelis 1998) describes what the spatiotemporal analysis and reasoning is:

> *"Geographic entities, like everything else in the world, exist in time as well as in space;..... Spatio-temporal reasoning is not reasoning about some abstract (x,y,z,t) framework: it is mainly reasoning about the appearance, change, and disappearance of things in space and over time."*

For decades, Geography has emphasised the spatial component, leaving aside the temporal component. From its inception, the GIS development focused on the analysis of the geographic elements (space); this characteristic was inherited and the temporal aspect was consigned to the second place. By following the quantitative revolution of the 60's, Geography was catalogued as the "space science", so that the temporal component was limited to a simple attribute (Ott and Swiaczny 2001). One of the first references to spatiotemporal data is to be found in the study published by Donna Peuquet (Peuquet 1984), who by reflecting on temporal data series, identified the nature and the importance of the temporal dimension setting it apart from the spatial dimension.

Such as Ott and Swiaczny indicate (Ott and Swiaczny 2001), time is a relevant aspect in the subject field of GIScience, an element attracting ever so much attention by its importance, particularly during the last decade. This might be due to the large volume of data we now have at our disposal and to the potential that the historical series of data have generated, whereby the lack of effective methods and tools to deal with these data is shown; there are not methods to process and analyse temporal data appropriately. The time variable provides an additional component in GI management and it needs to be dealt with differently using new methods and models.

2.1 Starting point

The work of many authors[1] has left a wide, robust theoretical base on the aspects to be taken into account for incorporation of time into information systems and the integration of the spatiotemporal variables into the GIS. Others[2] have established the necessary concepts for the extraction and treatment of the temporal expressions and the bases for semantic analysis of the GI. In spite of the significant advances carried out, both the issue of the space-time relationship and the incorporation of the semantic analysis into the GIS are still open research fields, and more importantly, none of the authors has raised the issue as the space-time-semantics triad. There are no previous studies on this matter going beyond modelling based on cognitive analysis (Mennis 2003), the frameworks proposal for semantic interoperability in GIS applications (Stoimenov and Dordevic-Kajan 2002), or the use of general semantic elements (Web Service Modelling Ontology, Internet Reasoning Service) in specific geographic settings such as emergency management (Tanasescu et al. 2006).

A landmark in the research line in which this proposal is framed is the work presented by Gail Langran (Langran 1989, Langran 1990, Langran 1992), who stated that a reasonable objective for the GIS would be to be capable of following up the changes occurring in a certain area by storing the historical data. She pointed out that the Information Systems in the 80's (both spatial and alphanumerical) tended since then to omit the historical record of the data and they did not store previous versions of the system. In order to resolve this weakness, she proposed a conceptual, logical and physical model to enrich systems with temporal capabilities. She introduced the necessary concepts for the changes in geographic features to be stored cumulatively without data duplication at the time of their storage. She later presented (Langran 1993) a series of annotations applicable to the spatiotemporal systems with which the first general bases of this type of systems were settled. Although the incremental model presented by (Langran 1992) is conceptually similar to our proposal, we will explain later on this article (Section 4) how the spatiotemporal factor +/- δt and the semantic descriptor *S* determine that the proposed metamodel is different in concept, hence new and innovative. Following sections will explain in depth our proposal.

3. IDENTIFICATION OF THE PROBLEM

At the present time the models and data storage systems of GI manage the time variable inefficiently in regard to users' needs, most particularly the possibility of carrying out analyses and a follow-up of the geographic features as a continuum. It is possible to register changes that occur in the reality that surround us (real world) and there are methods for implementation of basic temporal analyses allowing us to find

[1] (Langran 1990, Langran 1992, Langran 1993), (Worboys 1994b, Worboys 1998), (Peuquet 1984, Peuquet 1994, Peuquet 2001, Peuquet 2002), (Yuan 1996, Yuan 1997, Yuan 1999, Yuan et al. 2004), (Wachowicz and Healey 1994), (Galton 2001, Galton 2004), (Hornsby and Yuan 2008, Yuan and Hornsby 2007), (Nixon and Hornsby 2010)

[2] (James Pustejovsky 2004), (Verhagen et al. 2005, Verhagen et al. 2007), The MITRE Corporation (Corporation 2004, Spatio Temporal MITRE 2009), (Jones and Purves 2008, Jones et al. 2001, Jones et al. 2004), (Manning and Schütze 2003), Markowetz (Markowetz et al. 2005), (Mennis 2003, Mennis et al. 2000)

out about the changes occurring within a certain period of time. Nonetheless, this task is carried out with a dataset captured at different times and states, what implies the utilisation of multiple photos of the reality. Every one of these photos is generated at a specific time tn and registered independently within the system, ignoring the temporal and spatial correlation of the particular feature and the correlation of the features as a set . This model known as Snapshot (Armstrong 1988) was the first one to incorporate temporality into the spatial databases, and in spite of being so old and inefficient due to the constant duplication of data and attributes, their basic concepts still persist in the present-day systems (ESRI 1998, ESRI 2009). Thus, in a follow-up system one has an independently stored layer set that is kept as historical archives to which additional processes have to be applied in order to check the change zones and other characteristics inherent to the evolution of the territory. From this viewpoint, an initial finite set representing Current Reality (CR) is derived. We define the set CR as follows:

$$CR = \{t_0, t_1, t_2, t_3, \ldots\ldots, t_n \quad / \quad t_0 \neq t_1 \ \ldots\ldots \neq \ldots\ldots, t_n\}$$

where: t_0 → Initial reality. (Photo *0)*
t_n → Reality in a subsequent time *n* (Photo *n*)

Each element t_n corresponds to a graphic output representing one or several levels of the modelled reality (layers) at a specific time. The fact that the set *CR* is defined by multiple elements and the number of these elements increases over time as the changes are registered implies many drawbacks. From the viewpoint of usage and capabilities provided by the system, it is possible to mention:

- Lack of a binding historical register of the represented features.
- Users do not achieve directly (or even indirectly in some cases) the desired answers from the spatiotemporal analyses carried out.
- Inability to develop real spatiotemporal analyses. Michael Worboys (Worboys 1994a) indicates two cases that are not resolved naturally and are related to changes in space, time and attributes: (i) changes in population density in a certain district within a fixed time interval; in this case not only the variation in the number of inhabitants is solely taken into account, but also the changes in the legal and geographic boundary line of the particular zone; (ii) the evolution of the morbidity in a fast growing city in a period of two decades.
- Attribute→space→time relationships without one-to-one matching.
- Inability to find other related levels of information or associated geographic features.
- Time of query and processing longer than really needed in queries related to historical data or temporal characteristics.

From the viewpoint of the control and register of the information status, a possible solution to these deficiencies would be the handling of the versioning, however difficulties arise such as:

- The intrinsic need to manage versions. The historical registry of spatial data is carried out following the methods developed for conventional databases (Snodgrass 1990, Snodgrass 1992).
- Information duplication (geometries and attributes) in zones not going through changes. A difficulty derived from versioning style with which the data are currently treated within the spatial databases. As example, the GeoDataBase model of ESRI® (ESRI 2009).
- Lack of data versions (releases). There is evidence of the registered versions but this does not imply identifying how many versions an entity has in store; there might be one or as many as existing versions in the database.
- Possible incompatibilities between the spatial and the descriptive components, taking into account a possible update of the attributes (alphanumerical data) but not of the spatial component (geometry). Example: the census.
- Unnecessary package traffic on the computer networks derived from data duplication.

In addition to these deficiencies, we have to mention the inability of the GI to be comprehended from a semantic perspective, i.e. its meaning. The data are not prepared to interact with users in natural language and they are not ready to define neither their own context nor their meaning. This is due to the fact that the natural flow of information treatment in the GIS is User→System←Data, the system being the one interpreting the user's commands (requests) to process subsequently the assigned user's commands taking the related data. In this scenario, everything revolves around the system by having the ability to interpret user commands and data structure; the user does not need to directly interact with the stored data, it just proposes tasks and processes. Under these conditions the user loses self-reliance, remains assigned to a second place and sees him(her)self constrained by the own capabilities of the system. If data were better fitted semantically, able to describe their context and inform who, what and how they are, defining the set they belong to, then, they could interact among them, relate to one another and set up natural subsets (e.g. buildings, rivers, ways, trees), independent of their storage, going beyond the data model controlling them. This would help the user with procedures since the system would not have to transform every order, and part of it could be comprehended directly by the data. In this case a flow of the type User←System←→Data would be generated.

4. DESCRIPTION OF THE PROPOSAL

The above-mentioned deficiencies are indeed a problem requiring attention and needing solutions. To this effect, we have considered interesting to present a proposal improving the storage of the dynamic GI in the spatiotemporal (δt) and semantic (S) domains, with the purpose of optimising its retrieval, management, analysis, and general tasks based on spatial and spatiotemporal reasoning. Therefore, we propose the definition of a metamodel for storage of the dynamic GI –DGI– to be appropriate for different application domains through specific models and apt for materialisation through mark-up languages. As a result, a new robust, dynamic and flexible storage format is anticipated that integrates the spatial, the temporal and the semantic components. Based on the presented bibliographic reference framework and having exposed the issue, the starting hypothesis for definition of this research work is as follows.

The hypothesis on which this research is based on is the lack of the semantic and temporal components in the current structures of Geographic Information storage and which causes the spatiotemporal analyses to be deficient. The proposal of a new model incorporating an independent temporal structure and a

semantic meaning would optimise such storage and would allow improving GI retrieval, processing and analysis capability. If this hypothesis is substantiated, the integration of the geographic, semantic and temporal components through standards would allow:

- exploiting efficiently the dynamic nature of GI;
- optimising the modelling, analysis and transfer of spatiotemporal GI;
- changing the current GI storage methods that do not appropriately fit the dynamic reality they represent;
- relying on well structured data to carry out geographic analysis taking into account the meaning of the data and the time variable;
- taking the first steps toward compatibility in the representation and analysis of data that include the time variable between the different available platforms, hence moving forward toward the semantic and temporal interoperability of the spatiotemporal data.

In order to substantiate our hypothesis, we expect to propose and implement a metamodel for the enriched storage of GI involving temporal and semantic structure, also enabling the optimization of retrieval and dynamic representation of geographic features, the ability to carry out spatiotemporal analyses, and the interaction and exploitation of data in natural language. Broadly speaking, the objective of this work is to define a model of GI storage with the following characteristics:

- Integration of the dynamic nature of GI.
- Incorporation of semantic components describing the meaning and the temporal aspect of the stored data.
- Incorporation of semantic elements enabling the user to naturally interact with them (Natural language interaction).
- Integration of spatiotemporal structures that will allow registering incrementally the changes occurred in geographic features.
- Definition of rules for interaction of any stored dataset.
- Design to be implemented with mark-up languages.

4.1 Proposal and methodology

4.1.1 Proposal and anticipated advantages.

This work would optimise the register of geographic information through a new storing structure, incorporating the succeeding geometric and alphanumerical changes occurring over time on the geographic features, hence avoiding information duplication; furthermore, this work would permit to know the reality of a registered geographic feature at any time. We intend to provide data with semantic elements that should in turn describe them by using mainly available standards and specifications. We propose to generate a model describing a unitary set (singleton) for dynamic data storing, semantically enriched and called *Proposed Reality –PR–* defined as follows:

$PR_0 = t_0 \pm \delta t;$

$PR = \{PR_0, S\};$ ➔

$PR = \{(t_0 \pm \delta t) / S \in PR \wedge \exists$ one and only one $S\}$

where: t_0 → Initial reality

δt → Spatiotemporal change of reality (t_x)
Spatiotemporal factor

The spatiotemporal factor δt would make possible to know the reality (current state) of a geographic feature at a certain time and to represent the changes of the modelled object, hence of the stored feature. The changes in attributes and geometry may be registered and looked up at any point of the temporal scale in which the information has been registered; a discretisation of the model may therefore be inferred. The set contains a semantic descriptor (S) that will grant the stored data the necessary description so that data are self-describable and may relate according to their geographic and temporal nature. The following conditions are proposed for this set:

- $\forall t_x \in PR_0, \exists \rightarrow S \in PR$
- $PR_0 \subset S \quad v \quad S \in PR_0$
- $t_0 = dom(t)$
- $t_x = t_{x-1} + \delta t \quad \wedge \quad t_x = t_{x+1} - \delta t$
- $\sum_{X=0}^{n} t_x = t_n$
- $\sum_{X=0}^{n} \delta t_x = \delta t_n$

This proposal could optimise the current process through the creation of a model for GI storage in which the dynamic component of the represented reality is registered, thus reducing the problems exposed for the finite set *CR*. With the *PR* model, in addition to the above-mentioned aspects in the hypothesis, it is anticipated that the outcome of this research study allow:

- Optimising the volume of storage due to the elimination of coincident geometries, hence improving transfer time of geographic data belonging to time series or collections.
- The dynamic representation of the temporal changes registered in geometry and attributes owing to the data intrinsic timeline. These changes may be linear or cyclic depending on the nature of the modelled feature.
- Optimising response time in queries and spatiotemporal analyses.
- Optimising processes applied to spatial and temporal reasoning.
- Improving the interaction and relationship of data of the same nature through semantic descriptors.
- Enabling human-data interaction in natural language.

4.1.2 Methodology.

This research work will be based on standard mark-up languages. Among the likely useful languages for implementation of the proposal, the following stand out: Geography Mark-up Language –GML– (OGC 2007), Keyhole Mark-up Language –KML– (OGC 2008), SpatialML (Spatio Temporal MITRE 2009), TimeML (ISO 2007), DARPA Agent Mark-up Language (DARPA's Information Exploitation Office 2006), Web Ontology Language –OWL–, Resource Description Framework (RDF), SPARQL Query Language for RDF, in addition to other *de facto* or *ad hoc* standards related to semantic, temporal, and/or GI storage aspects. To substantiate the defined hypothesis, the following activities are put forward:

- Finding an efficient way of fusing together and uniting space and time in the structures of storage with mark-up languages (i.e. GML and TimeML). Here the storing methods based on mark-ups and binaries, as well as the coincidences and differences between methods and time mark-up standards will be assessed.
- Evaluation of GML-based temporal models in other specific application fields. The analysis of the temporal

elements of the Aeronautical Information eXchange Model (AIXM), and of the models applied to the Geological Time Systems (Solid Earth GRID 2010) and CHRONOS (Iowa State University 2010) are proposed.
- Analysis of the appropriate form for incorporation of semantic, temporal and descriptive annotations of data. The concepts applied here will be based on widely disseminated methodologies such as the Geographic Information Retrieval (Markowetz et al. 2005, The Information Retrieval Facility 2010) and the Natural Language Processing (Manning and Schütze 2003).
- Definition of the general concepts of the GI storing metamodel. Setting up the model layers and their characteristics.
- In order to validate the proposal, the metamodel will be implemented through two specific models. The implementation will be able to answer questions that have not been possible to solve so far. We will expose the deficiencies, difficulties and weak points as well as the strengths and opportunities offered.

5. CONCLUSION AND EXPECTED CONTRIBUTIONS

As a conclusion, it is possible to state that the semantic and temporal enrichment of the GI and its implementation through mark-up languages for its integration into the GI management systems is, we believe, the next natural step in the research carried out on the space-time issue in the Information Systems. It is necessary to carry on with the work of Peuquet, Langran, Armstrong, Snodgrass, Worboys, Yuan, Wachowicz, Galton, Hornsby, Frank, Jones, Shen, and all the researchers who have contributed some element to lay the foundations of the change in the paradigm toward the temporal analysis of GIS and the semantic interpretation of the GI. This proposed research work is warranted since it is anticipated to make progress, (i) setting up mark-up languages as an integrating element of the spatial, temporal and semantic elements; (ii) adding further descriptions to data so that they will be able to interact, to describe themselves and to provide the system with temporal and spatiotemporal dynamism as well as meaning. Temporal dynamism through a robust mark-up language; spatiotemporal dynamism by identifying incremental variations; enhancement of meaning by identifying the represented feature type; and (iii) developing concepts enabling progress in geosemantics.

The innovation of this research work lies in the proposal of a metamodel for representation, retrieval, reasoning and spatiotemporal and semantic analysis of GI. The anticipated contributions to be gained from this research are:

- Proposal of a metamodel enabling exploitation of the GI dynamic component.
- Definition of the method for incorporation of the semantic component into the GI storage structure.
- Definition of the method for integration of an independent temporal component into the GI.
- Proposal of a new format for the GI integrating the temporal, spatial, and attribute components, providing a semantic-temporal support.
- Proof of concept through implementation of mark-up language-based storing spatiotemporal structures.

REFERENCES

Armstrong, M. P. 1988. "Temporality in spatial databases" GIS/LIS 88: Accessing the World, The Urban and Regional Information Systems Association, Falls Church - USA, 880-889.

Brandeis University 2005. Annotating, Extracting and Reasoning about Time and Events, Dagstuhl - Germany. http://www.dagstuhl.de/de/programm/kalender/semhp/?semnr=05151 Schloss Dagstuhl (Accessed 16/04/2009)

Brandeis University. 2006. Annotating and reasoning about time and events - ARTE, http://www.timeml.org/acl2006time/ (Accessed 15/04/2009)

Corporation, T. M. 2004. "Time Expression Recognition and Normalization Evaluation" *TERN-2004 Evaluation Workshop*.

Couclelis, H. 1998. "Aristotelian Spatial Dynamics in the Age of Geographic Information Systems" Spatial and temporal reasoning in geographic information systems, M. J. Egenhofer and R. G. Colledge, eds., Oxford University Press, New York - USA, 109-118.

DARPA's Information Exploitation Office. 2006. The DARPA Agent Markup Language Homepage, http://www.daml.org BBN Technologies (Accessed 17/05/2010)

DARPA's Information Technology Office. 2010. Conference on Message Understanding, Morristown - USA. http://en.wikipedia.org/wiki/Message_Understanding_Conference Wikimedia Foundation (Accessed 30/07/2010)

Dipartimento diInformaticai Comunicazione. 2010. TIME International Symposium, Milan - Italy. http://time.dico.unimi.it/TIME_Home.html Universitat degli Studi di Milano (Accessed 30/07/2010)

ESRI. 1998. "ESRI Shapefile Technical Description" *J-7855*, Environmental Systems Research Institute Inc., Redlands - California - USA.

ESRI. 2009. Geodatabase | Spatial Data & Information Management | GIS Data Storage, Redlands - CA - USA. http://www.esri.com/software/arcgis/geodatabase/index.html ESRI (Accessed 08/10/2010)

Galton, A. 2001. "Space, time, and the representation of geographical reality" *Topoi*, 20(2), 173-187.

Galton, A. 2004. "Fields and objects in space, time, and space-time" *Spatial Cognition and Computation*, 4(1), 39-68.

Hornsby, K. S., and Yuan, M. 2008. *Understanding Dynamics of Geographic Domains*, CRC Press, Boca Raton - USA.

ISO. 2007. "Language resource management – Semantic Annotation Framework (SemAF) – Part1: Time and events" *ISO/CD 24617-1*, International Organization for Standardization -ISO-, Geneva - Switzerland.

Iowa State University and NSF. 2010. CHRONOS, http://chronos.org/index.html National Science Foundation (Accessed 26/05/2010)

James Pustejovsky, Inderjeet Mani, and Jerry Hobbs. 2004. Temporal Awareness and Reasoning Systems for Question Interpretation, http://www.timeml.org/site/tarsqi/index.html TimeML Working Group (Accessed 10/03/2009)

Jones, C. B., Abdelmoty, A. I., Finch, D., Fu, G., and Vaid, S. 2004. "The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing" Geographic Information

Science, M. J. Egenhofer, C. Freksa, and H. J. Miller, eds., Springer-Verlag, Berlin - Germany, 125-139.

Jones, C. B., Alani, H., and Tudhope, D. 2001. "Geographical Information Retrieval with Ontologies of Place" Spatial Information Theory. Foundations of Geographic Information Science, International Conference COSIT 2001, D. R. Montello, ed., Springer-Verlag, Berlin - Germany, 322-335.

Jones, C. B., and Purves, R. S. 2008. "Geographical Information Retrieval" *Int J of Geogr Inf Science*, 22(3), 219-228.

Langran, G. 1989. "A review of temporal database research and its use in GIS applications" *Int J of Geogr Inf Systems*, 3(3), 215-232.

Langran, G. 1990. "Temporal GIS design tradeoffs" *URISA Journal*, 2(2), 16-25.

Langran, G. 1992. *Time in geographic information systems*, Taylor & Francis, London - UK.

Langran, G. 1993. "Issues of implementing a spatiotemporal system" *Int J of Geogr Inf Systems*, 7(4), 305-314.

Manning, C., and Schütze, H. 2003. *Foundations of statistical natural language processing*, MIT Press, Cambridge - MA.

Markowetz, A., *et al.* 2005. "Geographic information retrieval" Next generation geospatial information: from digital image analysis to spatio temporal databases, P. Agouris and A. Croitoru, eds., Taylor & Francis, 5-17.

Mazul, P., and Dale, R. 2009. "The DANTE temporal expression tagger" Human Language Technology. Challenges of the Information Society, Z. Vetulani and H. Uszkoreit, eds., Springer-Verlag, Berlin - Germany, 245-257.

Mennis, J. L. 2003. "Derivation and implementation of a semantic GIS data model informed by principles of cognition" *Computers, Environment and Urban Systems*, 27(5), 455-479.

Mennis, J. L., Peuquet, D. J., and Qian, L. 2000. "A conceptual framework for incorporating cognitive principles into geographical database representation" *Int J of Geogr Inf Science*, 14(6), 501-520.

Nixon, V., and Hornsby, K. S. 2010. "Using geolifespans to model dynamic geographic domains" *Int J of Geogr Inf Science*, 24(9), 1289-1308.

Office, D. A. 1995. "Proceedings of the 6th conference on Message understanding" *MUC6 '95*, Morristown - USA.

OGC. 2007. "OpenGIS® Geography Markup Language (GML) Encoding Standard 3.2.1" *OGC 07-036*, OGC Inc.

OGC. 2008. "OGC® KML" *OGC 07-147r2*, OGC Inc.

Ott, T., and Swiaczny, F. 2001. Time-integrative Geographic Information Systems - Management and Analysis of Spatio-Temporal Data, Springer-Verlag, Berlin - Germany.

Peuquet, D. J. 1984. "A conceptual framework and comparison of spatial data models" *Cartographica: The Int J for Geographic Information and Geovisualization*, 21(4), 66-113.

Peuquet, D. J. 1994. "It's About Time: A Conceptual Framework for the Representation of Temporal Dynamics in Geographic Information Systems" *Annals of the Association of American Geographers*, 84(3), 441-461.

Peuquet, D. J. 2001. "Making space for time: Issues in space-time data representation" *GeoInformatica*, 5(1), 11-32.

Peuquet, D. J. 2002. *Representations of space and time*, The Guilford Press, London - UK.

Pustejovsky J., and Mani I. 2003. TANGO, http://www.timeml.org/site/tango/index.html ARDA Workshop (Accessed 10/03/2009)

Sinton, D. F. 1978. "The inherent structure of information as a constraint to analysis: Mapped thematic data as a case study" *First International Advanced Study Symposium on topological data structures for GIS*, Cambridge - USA, 1-17.

Snodgrass, R. T. 1990. "Temporal databases status and research directions" *ACM SIGMOD Record*, 19(4), 83-89.

Snodgrass, R. T. 1992. "Temporal databases" Theories and Methods of Spatio-Temporal Reasoning in Geographic Space, A. Frank, I. Campari, and U. Formentini, eds., Springer-Verlag, Berlin - Germany, 22-64.

Solid Earth GRID. 2010. Geological Time Systems, https://www.seegrid.csiro.au/twiki/bin/view/CGIModel/GeologicTime CSIRO (Accessed 26/05/2010)

Spatio Temporal MITRE. 2009. "SpatialML: Annotation Scheme for Marking Spatial Expressions in Natural Language 3.0", ©The MITRE Corporation.

Stoimenov, L., and Dordevic-Kajan, S. 2002. "Framework for semantic GIS interoperability" *Facta Universitatis (Series: Mathematics and Informatics)*, 17(1), 107-125.

Tanasescu, V., *et al*. 2006. "A Semantic Web GIS based emergency management system" *Semantic Web for eGovernment 2006*, Athens - Greece, 1-12.

The Information Retrieval Facility. 2010. Home Web Page, http://www.ir-facility.org/ IRF org (Accessed 15/04/2010)

TimeML Working Group. 2008a. Markup Language for Temporal and Event Expressions, http://www.timeml.org (Accessed 09/03/2009)

TimeML Working Group. 2008b. TimeML Corpora, http://www.timeml.org/site/timebank/timebank.html (Accessed 10/03/2009)

Turing, A. M. 1950. "Computing machinery and intelligence" *MIND*, 59(236), 443-460.

Verhagen, M., *et al*. 2007. "Semeval-2007 task 15: Tempeval temporal relation identification" *4th International Workshop on Semantic Evaluations*, Morristown - USA, 75-80.

Verhagen, M., *et al*. 2005. "Automating Temporal Annotation with TARSQI" *The ACL Interactive Poster and Demonstration Sessions*, Michigan - USA, 81-84.

Wachowicz, M., and Healey, R. G. 1994. "Towards temporality in GIS" Innovations in GIS: selected papers from the first National Conference on GIS Research UK, M. Worboys, ed., CRC Press, London - UK, 105-115.

Worboys, M. 1994a. "A unified model for spatial and temporal information" *The Computer Journal*, 37(1), 26-34.

Worboys, M. 1994b. "Unifying the spatial and temporal components of geographic information" Proceedings of the Sixth Int Symp on Spatial Data Handling, London, 505-517.

Worboys, M. 1998. "A generic model for spatio-bitemporal geographic information" Spatial and temporal reasoning in GIS, M. J. Egenhofer, Oxford University Press, NY - USA, 25-39.

Yuan, M. 1996. "Temporal GIS and spatio-temporal modeling" *3rd International Conference on Integrating GIS and Environmental Modeling*, Santa Barbara - USA, 21-26.

Yuan, M. 1997. "Modeling semantical, temporal and spatial information in geographic information systems" Geographic Information Research: Bridging the Atlantic, M. Craglia and H. Couclelis, eds., Taylor & Francis, London - UK, 334-347.

Yuan, M. 1999. "Use of a Three-Domain Representation to Enhance GIS Support for Complex Spatiotemporal Queries" *Transactions in GIS*, 3(2), 137-159.

Yuan, M., and Hornsby, K. S. 2007. Computation and visualization for understanding dynamics in geographic domains: a research agenda, CRC Press, Boca Raton - USA.

Yuan, M., Mark, D. M., Egenhofer, M. J., and Peuquet, D. J. 2004. "Extensions to Geographic Representation" A Research Agenda for Geographic Information Science, R. B. McMaster and E. L. Usery, eds., CRC Press, Boca Raton - USA, 129-156.

CONCEPTUAL DESIGN OF A STAR-SCHEMA TO SUPPORT MULTI-DIMENSIONAL TRAFFIC ANALYSIS

Garavig Tanaksaranond [a], Tao Cheng [a], Andy Chow [a], and Alexandre Santacreu [b]

[a] Dept. of Civil Environmental & Geomatic Engineering, University College London, Gower Street, London, WC1E 6BT UK – ucesgta@ucl.ac.uk
[b] Road Network Performance & Research, Transport for London – alexandre.santacreu@tfl.gov.uk

KEY WORDS: Star schema, Traffic analysis, Multidimensional

ABSTRACT:

Huge amount of traffic data distracts multi-dimensional data analysis. The process can take up to a minute or more of powerful computational time when the data size approaches, normally, a hundred million records. This paper designed a star schema to support multidimensional analysis of our link travel time data in Central London. The star schema is more suitable for data analysis, and will be implemented within data warehouse and online analytical processing (OLAP).

1. INTRODUCTION

Huge amount of traffic data are collected every day and stored in the database. They are invaluable as they can provide us with means to identify traffic trends and patterns. With such a huge amount of historical data, it is very difficult and time consuming to perform multidimensional data analysis by using ad hoc SQL queries (i.e. analysis of data related to two or more categories (IBM, 2005)). For example, we have a large amount of five minute aggregated data, and we wanted to query a data set consisting of traffic data of selected links over a couple of months. The database will search each row until it finds all the rows that meet the criteria—the selected links and the selected months. The query is becoming more complex and time consuming should we want a summarized data; for example, the daily average. As a result, the process can take up to a minute or more of powerful computational time when the data size approaches, normally, a hundred million records. Therefore it is nearly impossible to change swiftly between different levels of granularity; such as from time of the day to day of the week or to month of the year. The efficiency deteriorates should we want to present spatial data on maps that are usually contained within the traffic data, and since the database also has to query the spatial data.

Data warehouse (DW) and Online Analytical Processing (OLAP) have been employed successfully to tackle the problems of huge amount of data or multidimensional data. A DW is a data repository for data analysis, rather than transaction processing and data manipulation (Chaudhuri et al., 1997). DW is often maintained separately from the operational databases which are optimized for inserting, deleting, and updating of data. OLAP is a tool to dig into the DW (Bédard et al., 2001). It provides 'online' analysis — users can manipulate data view, intuitively, by such mean as drilling down to the lower level of aggregation. Hierarchies define level of aggregation e.g. day is a subgroup of a month, and month is a sub group of a year.

There are only few research works that applied DW and OLAP to manage traffic data. Shekhar (2002) designed a data cube called 'cube view' to support traffic data analysis and visualization within 'Advanced Interactive Traffic Visualization System' (AITVS). The cube defines the perspective of traffic parameters. It defines how traffic parameters are (i.e. occupancy and volume of traffic on the road) aggregated across two dimensions: time (day, month, or year) and space (highway station, county, or region). The analysis is limited to these two dimensions.

Tang (2010) created an OLAP cube called 'congestion cube'. Congestion cube separated the data into two types of table: 'fact' and 'dimensions'. Fact is a group of measures of traffic congestion (duration and speed). Dimensions are types of categorization of measures such as time and road section. The fact table is linked with dimensions table by key attributes. The work developed by Tang (2010) cannot be applied to support analysis of traffic data directly since the congestion cube was designed for congestion data only—it was not for analyzing normal traffic data. Moreover, the physical direction of traffic was never been taken into account.

This paper developed a conceptual design of a star schema to support multidimensional analysis of link travel time data in Central London. The star schema defines how traffic data are structured. Directions of the travel time data are consolidated as a dimension of travel time. The star schema will be implemented within our OLAP and Data Warehouse.

2. LINK TRAVEL TIME DATA

Link travel time data, supplied by Transport for London (TfL), were captured as part of the current London Congestion Analysis Program (LCAP). A record of link travel time data consists of an average travel time of vehicles within five minute interval and attributes associated with the link travel time; i.e. link number, core, road type, and capture date and capture time. It also contains the number of vehicles when the travel time are collected. There are 1034 road links (according to data retrieve from TfL in July 2010) within the M25 motorway that encircled greater London. The size of travel data is indeed very large. We have 2.5 years of historical data; that amounts to 100 million records in the database.

The raw travel time data is tied to road network table with link number. The road network table contains information of each link on the network—such as link number, road name, link length, direction (inbound, outbound, clockwise, or anti-clockwise), and type of road defined by TfL (outer, inner, or central London). The relationship between the travel time table and the road network table is shown in the below:

Figure 1. Relationship between road network table and travel time table

3. STAR SCHEMA

Before implementing DW and OLAP, we need to design multidimensional model. The multidimensional model will be stored in the DW as star schema in order to support OLAP. The star schema divides information into two groups: fact and dimension. The fact consists of traffic measures including the numerical values we are interested in. Dimension represents possible ways of aggregating traffic measures. Fact and dimensions are stored as tables in the Relational Database System (RDBMS) while levels within dimensions are stored as columns within the dimension table. In case a star schema contains many levels, we can normalize directions table into multiple related tables in order to improve query performance; and this type of schema is called *snowflake schema*. Here, we present the star schema with normalized dimensions in order to clarify levels within each dimension. However, levels within each dimension will be stored as columns within dimension tables. The star schema is shown in Figure 2.

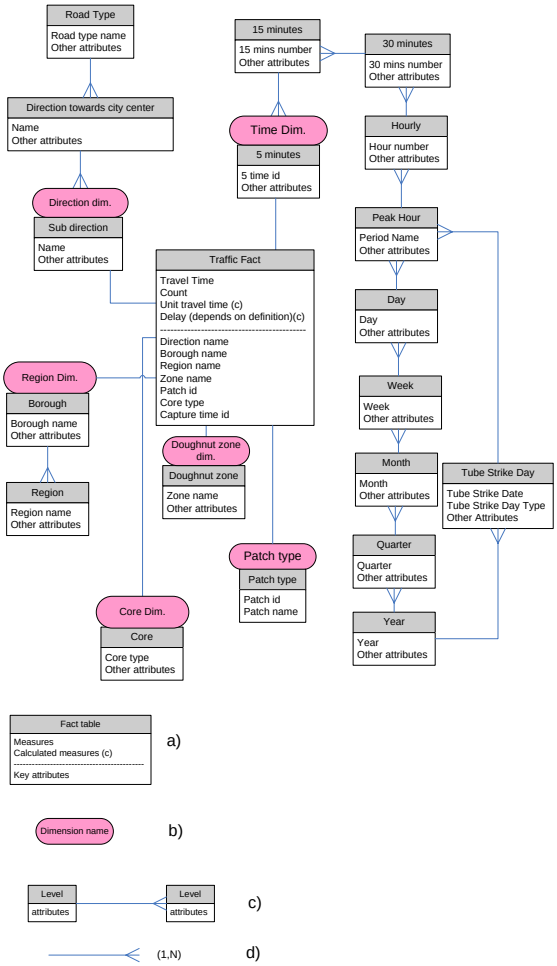


Figure 2. Multidimensional model for link travel time Notations are also shown: : a) fact table with measures, calculated measures (c), and key attributes linked with dimension tables, b) dimension name, c) levels with hierarchies, d) cardinalities between hierarchies.

The traffic measures derived from the raw data are *link travel time* (in seconds) and *vehicle count* (the number of vehicles that can be captured during five minutes capture time interval). Traffic measures also include derived measures calculated from the raw data i.e. *unit travel time* (travel time per link length) and *delay* (difference between actual unit travel time and base line unit travel time (as defined by TfL as the travel time between 2:00-6:00 am)).

There are five dimensions within our multidimensional model: *time*, *direction*, *region*, *core*, and *doughnut zone*. A dimension contains levels and hierarchies. Levels determine possible aggregation levels of measures. For example, Time dimensions can consist of 15 minute interval, 30 minutes interval, hourly, period of the day, daily, monthly, quarter, year sorted by smallest to largest levels. Hierarchies represent relationship of levels. One dimension can contain more than one hierarchy. As shown in Figure 2, time hierarchy consists of 3 hierarchies. One hierarchy ranges from 5 minutes to a year. Another hierarchy ranges from 5 minutes of a tube strike day to a year.

The *direction* dimension consisted of three levels: *road type*, *direction towards city centre*, and *sub-direction*. Road type defines a road link whether it is an 'orbital', 'radial', or 'a road within a city centre'. We represent road type with line set. *Direction towards city centre* dimension represent sub group of *road type* level. The members of *direction towards city centre* level can also be 'clockwise' or 'anti-clockwise' directions for an 'orbital' road type. There is no direction category within the city centre. We represent directions towards city centre level with line set. *Sub-direction* level categorized *direction towards city centre* level. For example, the direction of 'inbound' road can be divided into 'northeast bound', 'southeast bound', 'northwest bound', and 'southwest bound'. Figure 3 shows the hierarchy of *direction* dimension. Figure 4 shows how *sub-direction* is defined.

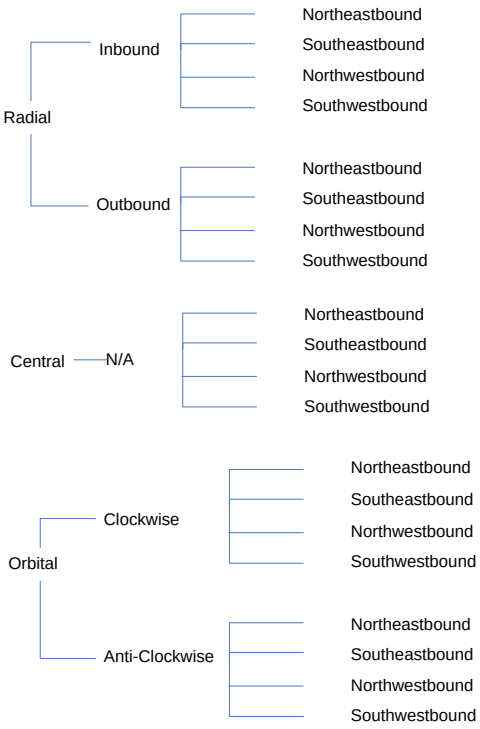


Figure 3. Hierarchy of direction dimension

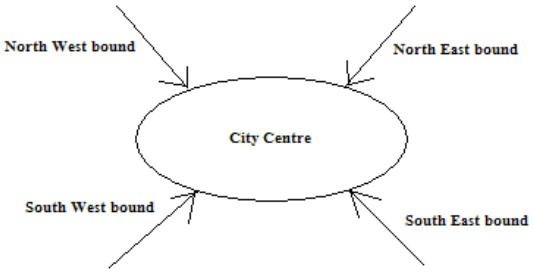


Figure 4. Inbound roads with their sub-groups

4. CONCLUSIONS AND FUTURE WORK

This paper designed a multidimensional model for link travel time data. The model also defines spatial data types and their topological relationships to spatial dimensions. We will implement spatial data warehouse and spatial online analytical processing to support multidimensional analysis. A visualization system will also be developed.

REFERENCES

Bédard, Y., Merrett, T. and Han, J. (2001). Fundamentals of spatial datawarehousing for geographic knowledge discovery. In: *Geographic Data Mining and Knowledge Discovery*. Taylor & Francis, pp. 53-73.

Chaudhuri, S. and Dayal, U. (1997), An Overview of Data Warehousing and OLAP Technology. SIGMOD Rec., 26(1): 65-74.

Golfarelli, M. , Rizzi, S, (2009). *Data Warehouse Design: Modern Principles and Methodologies*. McGraw Hill. pp. 242-256.

IBM (2005). IBM Informix Dynamic Server v10 Information Center. http://publib.boulder.ibm.com-/infocenter/idshelp/v10/index.jsp?topic=/com.ibm.ddi.doc/ddi236.htm. (accessed 14 June 2011)

Malinowski, E., Zimányi, E. (2004). Representing spatiality in a conceptual multidimensional model. In: *The 12th annual ACM international workshop on Geographic information systems*, Washington DC, USA, ACM: 12-22.

Malinowski, E. and Zimanyi, E. (2007). Spatial DataWarehouses: Some Solutions and Unresolved Problems. In: *IEEE International Workshop on Databases for Next Generation Researchers (SWOD 2007)*.

Michael,S., Vlamis, D., Nader, M., Claterbos, C.,Collins, C.,Campbell, M.,Conrad, F. (2010). *Oracle Essbase & Oracle OLAP: The Guide to Oracle's Multidimensional Solution*. McGraw Hill. pp.32-33.

Shekhar, S., Lu, C. T., Liu, R. and Zhou, C. (2002). CubeView: a system for traffic data visualization. In: *The IEEE 5th International Conference on Intelligent Transportation Systems*.

Tang, L. A., Yu, X., Sun, Y., Han, J., Peng, W. C., Kim, S., Gonzalez, H. and Seith, S. (2010), Multidimensional Traffic Data Analysis: A Congestion Cube Approach.

A SELF-ADAPTING FUZZY INFERENCE SYSTEM FOR THE EVALUATION OF AGRICULTURAL LAND

Y.L. Liu[a,b] L.M. Jiao[a,b*] Y.F. Liu[a,b] J.H. He[a,b]

[a] School of Resource and Environment Science, Wuhan University, 129 Luoyu Road, Wuhan 430079, China – (yaolin610, lmjiao027, yf610, hjianh)@163.com

[b] Key Laboratory of Geographic Information System, Ministry of Education, Wuhan University, 129 Luoyu Road, Wuhan 430079, China.

Commission VI, WG VI/4

KEY WORDS: Agricultural land evaluation, Fuzzy inference system, Self-adapting, Genetic algorithm

ABSTRACT:

The inference rules relating land characteristics to suitability class are crucial to the estimation of agricultural land suitability. In fuzzy logic modeling for agricultural land evaluation, inference rules based on membership functions and rule aggregation are constructed on predetermined evaluation criteria, including value ranges for fuzzy linguistic terms and weights of land variables. However, most existing evaluation criteria systems are built on the basis of expert knowledge or previous research, and could be highly subjective and contain uncertainty. This study integrates a genetic algorithm with a multi-criteria evaluation based fuzzy inference system to construct a self-adapting system that calibrates its evaluation criteria by self-learning from land samples. In the GA-optimized fuzzy inference system, the criteria for land evaluation are encoded into chromosomes, and the performance of the fuzzy inference system on a training set is used as the fitness of an individual chromosome representing a solution for the evaluation criteria system. The genetic algorithm repeatedly modifies a population of chromosomes through selection, crossover, and mutation. The optimized evaluation criteria are produced by decoding the final best chromosome generated at the end of the generational process of GA. To reduce the violation of the constraints for chromosomes and the destruction of excellent genes, the subsets in chromosomes are used as basic units in crossover and mutation. The reciprocal of the error rate, instead of the accuracy of the fuzzy inference model, is used as the fitness measure for chromosomes, which accelerates the convergence of the model. In the application of the model to the case study area, the error rate of the evaluation criteria system for the training set decreased from 27.93% to 1.81%. The accuracy of the GA-optimized fuzzy inference model for the test set was 93.22%, much higher than the accuracy of the original model, 72.07%. The results indicated that 74.8% of the cultivated land area was highly suitable for paddy. This is consistent with the empirical understanding of the quality of the cultivated land in the case study area. However, a total of 15 land units, 1.2% of the land area, were classified as not suitable.

1. INTRODUCTION

1.1 General Instructions

Agricultural land evaluation is carried out to estimate the suitability of land for an agricultural use such as arable farming or the planting of a specific crop. Based on the analysis of various land conditions, including climate, soil, topography, water supply, and other influential factors, agricultural land evaluation can be implemented by matching the land characteristics with the requirements of specified crops (FAO, 1976; Ahamed et al., 2000). Crucial to the estimation of agricultural land suitability is the building of the inference system relating land characteristics to agricultural land suitability. However, most of the existing inference systems were constructed based on artificial rules transformed from expert knowledge, and were highly subjective and contained uncertainty (Braimoh et al., 2004; Joss et al., 2008). Investigating the self-adjustment of the rules base for agricultural land evaluation will improve the accuracy of the evaluation result and facilitate the automation of the evaluation process.

The conventional methodological steps adopted in agricultural land suitability evaluation are: (1) Definition of types of farming or target crops; (2) Identification of environmental requirements of crops; (3) Determination of the quantitative relationship between each considered environmental factor and the potential productivity of the considered target crop; (4) Calculation of a suitability class or a score for a single factor for each evaluation unit; and (5) Combination of the classes or scores from all the factors and determining the overall suitability class for each evaluation unit. A central and critical issue of the process is how to parameterize and combine land characteristics in order to model the productive response of the target crop to a given set of environmental factors (Corona et al., 2008). Conventionally, a rule set for agricultural land evaluation is constructed based on expert knowledge to address this issue. The rule set is comprised of a series of decision rules that are used to define the factors' range of values for a given suitability class and a weighting system that gives the degree of importance of each factor. The rule-based approaches are highly dependent on expert knowledge and are subjective.

Alternative methods based on fuzzy set theory or fuzzy logic (Zadeh 1965) started to appear in land evaluation studies in the 1980s (Chang and Burrough, 1987; Burrough, 1989). In a fuzzy set, all objects belong to the set in membership values corresponding to their closeness to the defined class. The membership values range from 0 to 1, ranging from non-membership to complete membership, respectively. In contrast to classical methods, which assume land patches are crisply delineated in both attribute and geographic space resulting in homogenous polygons with single attribute values or suitability class, fuzzy logic provides a means to handle ambiguity and uncertainty to generate realistic continuous classifications

(Burrough et al., 1992, 1997; Tang et al., 1991; Tang and Van Ranst, 1992; Zhang et al., 2004; Corona et al., 2008). The fuzzy logic modeling for land evaluation consists of three main stages: fuzzification, fuzzy rule inference, and defuzzification. Fuzzification converts quantitative values into linguistic variables, such as highly suitable, moderately suitable, marginally suitable, and not suitable (FAO, 1976). Membership functions are defined to represent linguistic variables and determine the degree of class membership (i.e., from 0 to 1) for each of the linguistic variables (Joss et al., 2008). In fuzzy rule inference, Joss et al. (2008), Avdagic et al. (2008) and Reshmidevi et al. (2009) built production rules based inference systems, while Ahamed et al. (2000), Ceballos-Silva and Lopez-Blanco (2003), Corona et al. (2008), Kurtener et al. (2008), and Chen et al. (2010) employed the multi-criteria evaluation method. In a production rules based inference system, a fuzzy rule-base with rules in "IF…THEN" form was developed based on expert's or farmer's knowledge, and then the minimum-maximum (Min-Max) fuzzy inference method was employed to aggregate the rules. In the multi-criteria evaluation method, using land factors as criteria, after determining the criterion weights, weighted linear combinations or ordered weighted averaging were commonly used to derive the membership degrees for different classes for each evaluation unit (Corona et al., 2008). Defuzzification converts the fuzzy output to a single value or a single linguistic variable. Some of the frequently used methods are maximum membership principle, the weighted average method, and the center of maximum (Joss et al., 2008; Reshmidevi et al., 2009).

It can be seen from the process of fuzzy logic modeling for land evaluation that membership functions and weights play key roles in the modeling. However, the need for expert knowledge is an important constraint when determining the membership functions and weights (Mouton et al., 2011). In the process, several membership functions should be defined for each input environmental variable, one for each linguistic variable. The membership function for each linguistic variable of an environmental variable is developed on the basis of a range of the variable value. The values in the range have the membership degree greater than 0.5 to the linguistic variable, and the lower and upper limits of the range are marginal for the linguistic variable. The lower and upper values of the range are also called crossover points, since they are usually the intersection points of membership function graphs. Thus the value ranges and weights are the basis of the fuzzy logic modeling. However, they are usually determined subjectively. In most research, the value ranges were directly given on the basis of experience or relevant studies (Joss et al., 2008; Ahamed et al., 2000). McBratney and Odeh (1997) pointed out that the choice of crossover points specified for fuzzy membership computation should be based on data, expert knowledge, or conventionally imposed criteria. Self-learning and adjusting from samples of the crossover points, which were rarely reported by others, will be discussed in this study. Different land parameters have different relative importance to land suitability and, therefore, it is necessary to assign appropriate weights to the parameters (Ahamed et al., 2000), and the choice of weights is crucial for determining the overall land suitability index (Braimoh et al., 2004). Fugger (1999) and Braimoh et al. (2004) used simple ranking to rate land characteristics from 1 (least important) to n (most important). Ahamed et al. (2000) employed the Analytical Hierarchy Process (Saaty, 1980) method to determine the weights of land parameters. Davidson et al. (1994) suggest that the weights of environmental variables should be based on data and knowledge of the relative importance of differentiating land characteristics to crop growth. However, most studies adopted the weights determined by knowledge, and few investigated or calibrated the weights based on data or samples. Some research developed self-learning models for land evaluation such as artificial neural networks, and fuzzy neural networks (Jiao and Liu, 2004, 2007; Jiao, 2006; Liu and Jiao, 2008; Pradhan and Lee, 2010). But these models were more like "black boxes" and the learned rules can not be interpreted completely.

This study employs a genetic algorithm (GA) to calibrate the value ranges specified for fuzzy membership computation and the weights of land parameters in fuzzy inference, and constructs a genetic optimized fuzzy inference model for agricultural land evaluation. The original fuzzy inference model built subjectively by experts is self-adjusted using the genetic algorithm on the basis of the samples representing land suitability. The model integrates knowledge and data and becomes self-adapting.

2. DATA AND METHODS

2.1 Data

The study area is Anlu County in Hubei Province, China. It is located between longitude 113° 10′ E and 113° 57′ E, and latitude 31° 04′ N and 31° 29′ N, as shown in Fig. 1. The total area is 1355 km^2. The study area is under the north subtropical monsoon climate with an annual average rainfall of about 1100 mm. It is a generally slightly undulating area, and has a small river plain located in the center accounting for 10% of the total area. Three major land use categories are cultivated land (37.1%), forest (21.7%), and unused land (17.2%). Anlu is one of the important counties of commercial food production in Hubei province. Cultivated land in this area is mainly cropped with paddy as the principal crop, and paddy field accounts for 85% of the total cultivated area. This study will assess the suitability of cultivated land for paddy.

The land use map at the scale of 1:50,000 was collected from the Bureau of Land Management of Anlu. The cultivated land parcels on the map were used as evaluation units. We collected data required for land evaluation from various sources, including:

Soil map (1:50,000), soil survey report, soil organic matter content map (1:50,000) and soil pH map (1:50,000) from the Bureau of Agriculture of Anlu. Soil profile form, soil texture, and soil depth were extracted from the soil map and soil survey report.

Topographic map (1:10,000) and geomorphic map (1:100,000) from the Bureau of Land Resource Management of Anlu.
The distribution of the water table, irrigation guarantee rate, and drainage condition were delineated on the land use map and topographic map by consulting the experts and referencing the documents of the groundwater survey and irrigation works at the Bureau of Water Resources of Anlu.

Eight land variables influencing paddy yield were selected on the basis of the discussions of the panel for land suitability evaluation consisting of experts from universities, the Bureau of Agriculture, the Bureau of Land Resource Management, and the Bureau of Water Resources of Anlu. Summary statistics of the dataset are presented in Table 1. To identify the characteristics of the land units, overlay analysis was implemented between the land characteristic maps and land use map.

A total of 346 land samples with site and paddy yield information were collected from a field survey with the assistance of the local land management division. Fig.1 shows the distribution of the samples and Table 1 presents the statistics of the samples.

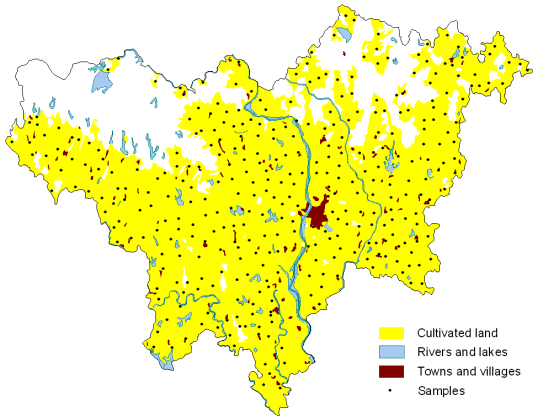


Fig. 1. The study area and samples.

Table 1. Statistics of land variables and samples.

Variables		Min	Max	Mean	Std. Dev.
Samples	Paddy yield (t/ha)	6.00	10.52	8.80	1.02
Land characteristics	Soil organic content (%)	0.34	3.24	1.97	0.58
	Soil depth (cm)	20	110	85.69	27.30
	Soil pH	3.9	8.5	6.82	0.92
	Water table (cm)	25	110	76.49	10.81
	Irrigation guarantee rate (%)	65	100	81.37	4.87
	Soil surface texture	String: loam; clay loam; sandy loam; gravelly clay			
	Drainage class	String: good; moderate; marginal; poor			
	Soil texture profile	String: loam in all horizons; sand in all horizons; clay in all horizons; loam/sand/loam; loam/clay/loam; clay/sand/clay; loam/clay/clay; loam/sand/sand			

Land samples were graded into four suitability classes, highly suitable (S1), moderately suitable (S2), marginally suitable (S3), and not suitable (N), on the basis of paddy yield and suggestions of the experts and local farmers.

Table 2. Suitability class of land samples.

Class	Count of samples	Paddy yield (t/ha)	
		Range	Average
S1	265	9.2–10.5	9.9
S2	62	8.4–9.2	9.0
S3	12	7.5–8.4	8.1
N	6	6.0–7.5	7.2
(Total)	346	6.0–10.5	8.8

2.2 ISPRS Affiliation (optional)

Following the multi-criteria method based fuzzy logic model, the steps of an ordinary fuzzy inference for agricultural land evaluation are as follows.

(1) Determination of the value ranges of grades and weights for land characteristics. One of the national standards for agricultural land evaluation, Regulations for Classification on Agricultural Land (Ministry of Land and Resources of China, 2003), provides an indicator of the ranges of the selected land variables associated with the study area. The variable ranges were refined by consulting the local agricultural experts. Empirical values of each land variable were graded into four linguistic variables (i.e., suitability classes): S1, S2, S3, and N. We employed the Delphi method (Rowe and Wright, 1999) to estimate the weights of land variables. The Delphi method was conducted to combine and refine the opinions of a panel of experts from agricultural and land resource research to establish the weights based on a merging of the information collectively available to the experts. Table 3 shows the value ranges and weights of land variables for paddy field evaluation.

Table 3. Original criteria for the evaluation for paddy fields.

Variables	weights	S1	S2	S3	N
Soil organic content (%)	0.22	3-4	2-3	1-2	0-1
Soil depth (cm)	0.10	100−140	60−100	30−60	0−30
Soil pH	0.05	6.5-7.0	6.0−6.5,7.0−7.5	5.0−6.0,7.5−8.0	4.0−5.0, 8.0−9.0
Water table (cm)	0.05	100−120	85−100	70−85	40−70
Irrigation guarantee rate (%)	0.15	80−100	70−80	60−70	50−60
Soil surface texture	0.20	Loam	Clay loam	Sandy loam	Gravelly clay
Drainage class	0.13	Good	Moderate	Marginal	Poor
Soil texture profile	0.10	loam in all horizons; loam/sand/loam; loam/clay/loam	loam/clay/clay; loam/sand/sand	clay/sand/clay; clay in all horizons	sand in all horizons

(2) Fuzzification

Because of its smoothness, non-zero, and concise notation, Gaussian membership function is a popular method for specifying fuzzy sets (Avdagic et al., 2008). The Gaussian membership function was employed in this study and defined by the following formula:

$$f(x) = e^{-(\frac{x-m}{c})^2} \quad (1)$$

where x is the value of a variable, m is the position of the center of the peak of the Gaussian "bell curve", c is a real constant, and e is Euler's number.

There is one membership function associated with each linguistic variable, meaning that there are four functions corresponding to four linguistic variables (S1, S2, S3, and N) of each land characteristic. The membership functions of S2 and S3 are in "full bell" shape, and the ones of S1 and N are in "semi-bell" shape (Fig. 2). However, Gaussian membership functions are not applied to the variables with string values. For character variables, if a value of a land unit belongs to a

linguistic term, then the membership degree of the linguistic term is set to 1, and the others are set to 0. For example, if the soil texture of a land unit is "loam", the membership degrees of four linguistic terms are 1, 0, 0, and 0, respectively.
According to the definition of the Gaussian function, m in the function can be given by:

$$m = (x_l + x_u)/2 \quad (2)$$

where x_l and x_u are the lower and upper limits of the value range of a linguistic variable. The crossover points, corresponding to c_2, c_3, and c_4 in Fig. 2, represent situations where the value of the characteristic is marginal for belonging to two related linguistic terms. At these points, the membership value of the land variable is:

$$f(x) = 0.5, \quad x = x_l, x_u \quad (3)$$

The constant c in the membership function can be deduced using Eq. (1), (2), and (3), as shown in Eq. (4). All membership functions for numeric land variables can be calculated in the same way:

$$c = \sqrt{(x_l - x_u)^2/(4\ln 2)} \quad (4)$$

where ln is the natural logarithm function.

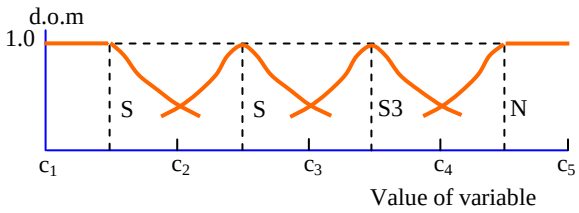


Fig. 2. Membership functions of a land variable.

(3) Fuzzy inference
The weights of land characteristics compose a weight matrix:

$$A = [a_i]_{1\times n} \quad (5)$$

where a_i is the weight of land characteristic i, n is the number of land characteristics.
A membership matrix is produced consisting of membership degrees of land variables to linguistic terms,

$$R = [r_{ij}]_{n\times m} \quad (6)$$

where r_{ij} is the membership degree of land variable i to linguistic term j, n is the number of land characteristics, and m is the number of linguistic terms.
The final evaluation matrix of a land unit is calculated as:

$$B = A \times R = [b_j]_{1\times m} \quad (7)$$

where b_j is the aggregated membership degree of a land unit to linguistic term j, and m is the number of linguistic terms.

(3) Defuzzification
The maximum membership principle is applied in defuzzification, i.e., the final class of a land unit is set to the linguistic term to which the land unit has the maximum membership degree.

2.3 A self-adapting fuzzy inference system

2.3.1 Framework: Genetic algorithm is employed to optimize the subjectively determined original criteria in the ordinary fuzzy inference system by self-learning from land samples. Genetic algorithm is based on natural selection in biological evolution, and can be used to solve various optimization problems in which the objective function is discontinuous (or with discrete values), non-differentiable, stochastic or non-linear (Avdagic et al., 2008). In a GA-optimized fuzzy inference system, the criteria for land evaluation are encoded into a chromosome. Genetic algorithm repeatedly modifies a population of chromosomes through selection, crossover, and mutation. Over successive generations, the population evolves toward an optimal solution (optimal criteria for paddy field evaluation).

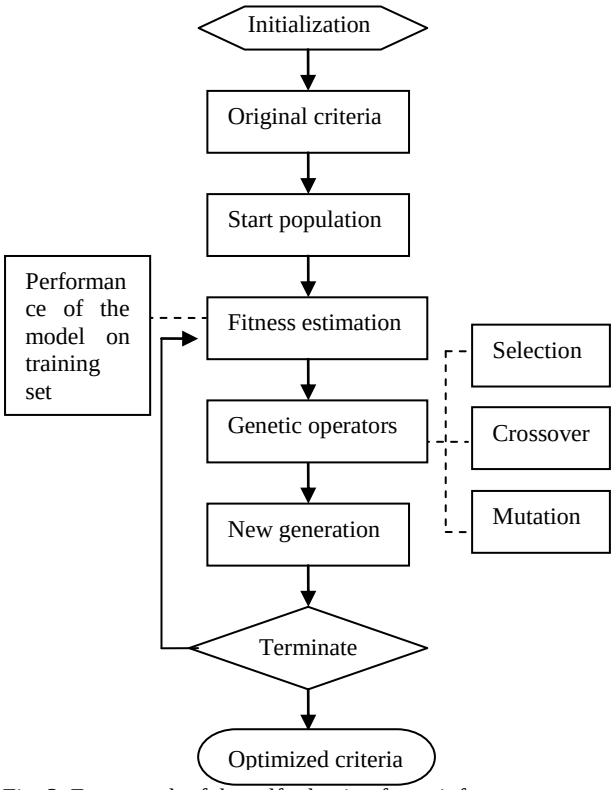


Fig. 3. Framework of the self-adapting fuzzy inference system.

2.3.2 Encoding: The parameters for a land variable include range limits, c_1, c_2, c_3, c_4, c_5 (Fig. 2), and weight a. A chromosome is constructed by linearly combining the range limits and weights of all land variables (Fig. 4). In a chromosome, the range limits associated with a variable compose a subset, and the weights of all variables compose a subset. In Fig. 4, subsets 1 to *n* (the number of land variables) represent range limits for land variables, and subset *n*+1 represent the weights. In this study, a chromosome has eight subsets for range limits and one for weights.

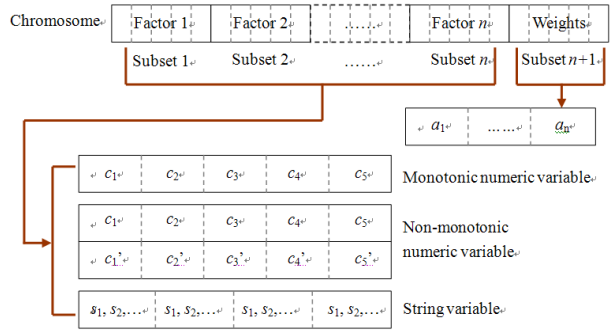


Fig. 4. Chromosome and logic subsets.

Since the subsets in chromosome have specific meanings, they should always meet the following constraints in genetic operations:

$$c_1 < c_2 < c_3 < c_4 < c_5 \text{ or } c_1 > c_2 > c_3 > c_4 > c_5 \quad (8)$$

$$a_1 + a_2 + \cdots + a_n = 1, \quad a_1, a_2, \cdots, a_n > 0 \quad (9)$$

2.3.3 Fitness function: The reciprocal of the error rate on the training set is used as the fitness for a chromosome. The fitness function can be described as:

$$F(x_i) = \frac{1}{n(x_i)/N} = N / n(x_i) \quad (10)$$

where x_i is a chromosome, $n(x_i)$ represents the number of the samples whose evaluation results generated by the fuzzy inference system are not consistent with their expectations, and N is the number of samples in the training set.

2.3.4 Genetic operators: The fitness proportionate selection, also known as roulette-wheel selection, is employed to select potentially useful solutions for reproduction (i.e., crossover and mutation). A chromosome's probability of being selected is:

$$P(x_i) = f(x_i) \Big/ \sum_{j=1}^{N} f(x_j) \quad (11)$$

where $f(x_i)$ represents the fitness of the chromosome i in a population, N is the number of chromosomes in a population. Crossover is used to produce new chromosomes for the next generation. The subsets in chromosome are used as basic units in crossover in order to avoid violating the constraints for chromosomes, as described in Eqs. (8) and (9). One-point crossover is applied. When a single point on both parent chromosomes is selected randomly, all data beyond the point in either chromosome is swapped between the two parent chromosomes (Fig. 5). The resulting chromosomes are the children.

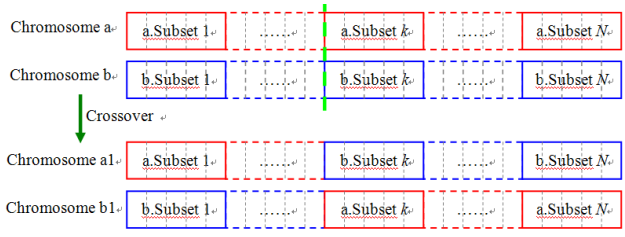


Fig. 5. Crossover of chromosomes.

Mutation is also used to produce new chromosomes. The purpose of mutation is preserving and introducing genetic diversity. One subset randomly selected from the chromosome to be mutated will be changed randomly under the constraints described in Eqs. (8) and (9).

If the selected subset to be mutated is associated with the range limits of a land variable, then the mutation will be implemented as follows:
Numeric variable:

$$c_j' = (c_{j+1} + c_{j-1})/2 + (rand - 0.5) \times (c_{j+1} - c_{j-1}) \quad (12)$$

where c_j is value of the range limit j, c_j' is the mutated range limit, j is a stochastic natural number between 1 and 5, i.e., j = 2, 3 or 4, and rand represents a stochastic real number between 0 and 1.

String variable:
A substring in the chromosome (a value of the land variable) is selected randomly and moved into an adjacent subset.
If the selected subset to be mutated is the last subset that represents the weights of variables, then the mutation can be described as:

$$w_j' = \begin{cases} w_j + (rand - 0.5) \times \frac{2}{n}, & w_j + (rand - 0.5) \times \frac{2}{n} > 0 \\ rand \cdot w_j, & w_j + (rand - 0.5) \times \frac{2}{n} \le 0 \end{cases} \quad (13)$$

where w_j is the original weight, w_j' is the mutated weight, j is a stochastic natural number from 1 to *n*, and *n* is the number of land variables. In the process, the original weight is added to a random real number ranging from 0 to the average weight. If the original weight is less than the average weight, the mutated weight may be less than 0. In this situation, an alternative modification is simply to multiply the original weight by a stochastic real number in (0, 1). To ensure the sum of the weights is equal to 1, each weight is transformed by:

$$w_j' = w_j \Big/ \sum_{i=1}^{n} w_i \quad (14)$$

where w_j', w_j are the weights before and after mutation, respectively, and *n* is the number of land variables.

3. RESULTS AND DISCUSSION

3.1 Self-adapting adjusting of original rules

The sample set was randomly divided into two groups in the proportion of 7:3, training set and test set. The training set had 242 samples, and the test set had 104 samples. The number of individuals in a population was set to 20. The best solution in the parent generation was directly copied to the next generation. The crossover probability was set to 70%, i.e., total 13 chromosomes in the new generation were created by crossover. The remaining 6 chromosomes in the new generation were created by mutation. The maximum number of generations of the genetic algorithm was set to 2000. The start population was

composed of one original chromosome, which was based on the original criteria system (Table 3), and 19 other individuals generated by mutating the original chromosome. By applying the original criteria into fuzzy evaluation of the samples, a total 110 samples were misclassified, i.e., the fitness of the original chromosome was 3.58 (=394/110). At the end of the training, the fitness of the final best solution was 55.20, meaning that the error rate on the training set of the evaluation criteria system decreased from 27.93% (= 1/3.58) to 1.81% (= 1/55.20). Fig. 6 shows the changes of fitness values with generations. The upper line represents the fitness value of the best solution, and the lower line represents the average fitness value of the population. The optimized criteria system for paddy field evaluation was generated by decoding the final best chromosome, as shown in Table 4.

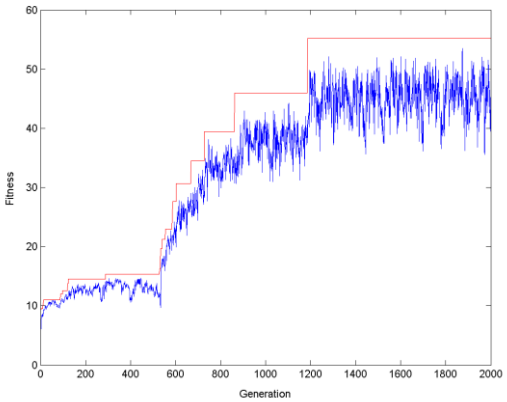


Fig. 6. Fitness value changing with generations.

3.2 Reliability and accuracy of the model

The test set is independent of the training set, and is used to test the reliability of the training result. If the accuracy of the trained fuzzy inference system on the test set does not decrease remarkably compared to the accuracy on the training set, then the trained system is deemed to be reliable. There were a total of 23 best individual solutions generated during the training. Table 5 shows the fitness and accuracy of these best individuals on the training set and test set. Fig. 7 shows the change in accuracy of the best individuals in sequence.

It can be seen that the accuracy on the training set increases monotonously during the training. The accuracy on the test set does not change between several sequential individuals, such as No. 2 and No. 3, No. 4 and No. 5, No. 7 and No. 8, No. 12 and No. 13, and No. 14 and No. 15. The accuracy on test set even decreases at No. 11, No. 14, No. 17, No. 18, and No. 22. Nevertheless, the accuracy on the test set increases in a general trend, and reaches the largest value finally at the end of the training, 93.22%. The changing of the accuracy on the test set shows that the fuzzy inference system was approaching the ideal form by self-adjustment in the training.

Table 4. GA-optimized criteria for paddy field evaluation.

Variables	weights	S1	S2	S3	N
Soil organic content (%)	0.15	2.14–4.00	1.83–2.14	0.41–1.83	0–0.41
Soil depth (cm)	0.06	60–140	41–60	15–41	0–15
Soil pH	0.06	6.3–7.0	5.0–6.3,7.0–7.5	4.8–5.0,7.5–8.0	<4.8, >8
Water table (cm)	0.05	100–120	86–100	53–86	0–53
Irrigation guarantee rate (%)	0.23	95–100	73–95	55–73	0–55
Soil surface texture	0.12	Loam	Clay	Sand	Gravelly clay
Drainage class	0.20	Good	Moderate	Marginal	Poor
Soil texture profile	0.13	loam in all horizons;	loam/sand/loam; loam/clay/loam; loam/clay/clay	loam/sand/sand; clay/sand/clay; clay in all horizons	sand in all horizons

Table 5. Fitness and accuracy of best individuals on the training set and test set

No.	Generation	Training set		Test set	
		Fitness	Accuracy	Fitness	Accuracy
1	1	8.36	88.04	4.72	78.81
2	2	9.20	89.13	4.92	79.66
3	3	9.52	89.49	4.92	79.66
4	10	9.86	89.86	5.13	80.51
5	11	10.22	90.22	5.13	80.51
6	13	11.04	90.94	5.62	82.20
7	89	12.00	91.67	5.90	83.05
8	98	12.55	92.03	5.90	83.05
9	121	13.80	92.75	8.43	88.14
10	124	14.53	93.12	9.83	89.83
11	290	15.33	93.48	7.87	87.29
12	529	16.24	93.84	9.83	89.83
13	532	17.25	94.20	9.83	89.83
14	533	19.71	94.93	9.08	88.98
15	541	21.23	95.29	9.08	88.98
16	555	23.00	95.65	10.73	90.68
17	585	25.09	96.01	7.87	87.29
18	587	27.60	96.38	7.38	86.44
19	603	30.67	96.74	8.43	88.14
20	669	34.50	97.10	9.08	88.98
21	728	39.43	97.46	13.11	92.37
22	862	46.00	97.83	11.80	91.53
23	1186	55.20	98.19	14.75	93.22

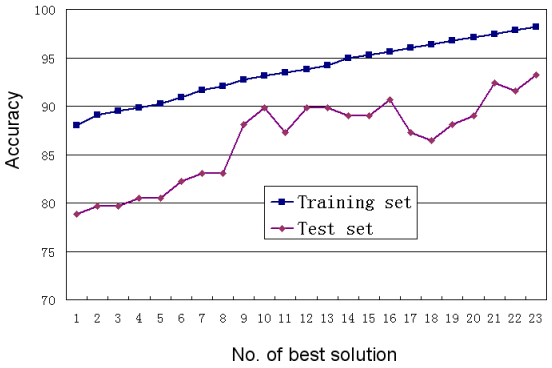


Fig. 7. Changing accuracy of training set and test set.

3.3 Comparison of original criteria and adjusted criteria

The adjusting of the value ranges of land variables can be found by comparing Table 3 and Table 4. The value ranges of land organic content, soil depth, soil texture profile and irrigation guarantee rate were changed significantly. The value ranges of water table, soil pH, soil surface texture, and drainage class were changed slightly or not changed. The adjusting of the original value ranges indicated that there were relatively large biases existing in evaluation rules of almost half of the land variables.

The weights of soil organic content and soil surface texture were decreased significantly, and the weights of irrigation guarantee rate and drainage class were substantially increased. Drought and water logging are two of the major agricultural disasters in the study area. The variance of annual precipitation is large. The maximum annual precipitation was 1772.6 mm in 1954, and the minimum was 652.9 mm in 1978. The temporal distribution of rainfall in a single year is also often very unbalanced. The average rainfall from April to August is 746.3 mm, occupying a proportion of 67.6% of that for the whole year.

3.4 Fitness function

The reciprocal of the error rate of the fuzzy inference model on the training set was used as fitness in the GA-optimized fuzzy inference model. Theoretically, the accuracy of the model, which is equal to 1 minus the error rate, can also be used as a fitness function. Fig. 8 shows the changes in the accuracy of two models in training, which use the inverse of error rate and accuracy as fitness functions, respectively. The final accuracies on the test set of the two models are both larger than 90%, and both could be regarded as being trained successfully. However, the model that employed the inverse of the error rate as its fitness function achieved a higher accuracy and converged faster than the other one. The inverse of the error rate amplified the differences between individuals and gave the individuals with higher accuracy more predominant opportunities to be selected to have children. Thus, the model using the inverse of error rate as its fitness function was more efficient.

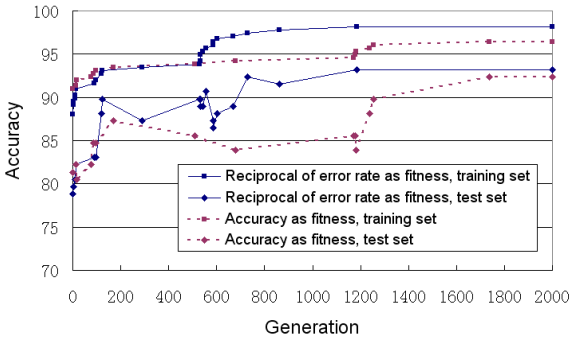
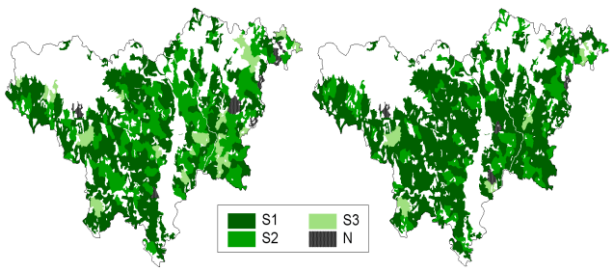


Fig. 8. Performance of two models using different fitness functions

3.5 Suitability class rating of agricultural land

The classified land suitability maps for paddy, generated using the original criteria and GA-optimized criteria, are shown in Figure 9. The number of land units and percentage area under each suitability class are tabulated in Table 6. The suitability classes of a unit, evaluated by applying the original criteria and GA-optimized criteria, could be the same or different. Statistics of the difference between these two evaluation results are shown in Table 7.



(a) (b)

Fig. 9. Classified land suitability maps for paddy: (a) original criteria, (b) GA-optimized criteria

Table 6. Statistics of suitability classes

Suitability class	Original criteria		GA-optimized criteria	
	Number of units	Percentage area (%)	Number of units	Percentage area (%)
S1	602	56.8	809	74.8
S2	345	33.1	188	20.9
S3	80	7.7	35	3.0
N	20	2.3	15	1.2
Total	1047	100.0	1047	100.0

Table 7. Difference between classified maps generated by original criteria and GA-optimized criteria.

Difference of suitability classes for a unit (C1-C2)	Number of units	Percentage area (%)
-3	4	0.2
-2	4	0.3
-1	96	12.2
0	598	53.1
1	318	30.9
2	21	2.1
3	6	1.2
Total	1047	100.0

C1: the numeric code of a suitability class generated by original criteria (1 for S1, 2 for S2, 3 for S3, and 4 for S4)
C2: the numeric code of a suitability class generated by GA-optimized criteria (1 for S1, 2 for S2, 3 for S3, and 4 for S4)

It can be seen from Table 7 that 53.1% of the area was classified into the same suitability class by both the original criteria and GA-optimized criteria. This implies that these two results are similar in their general trend, as shown in the land suitability maps (Fig. 8). However, 12.2% of the area was classified into an immediate lower class, and 30.9% of the area was classified into an immediate higher class. This indicates that there are more units classified into a higher class after GA-optimization. Table 6 shows that a large portion of the area was classified into high suitability in the final result. This is consistent with the fact that the study area is a commercial food base with high food productivity and a large portion of the cultivated land is considered very suitable for paddy. Only 4.2% of the area (a total of 50 land units) was classified as marginally suitable or not suitable.

4. CONCLUSIONS

This study discusses the self-adjusting of the criteria for agricultural land evaluation, which has seldom been studied in previous research, and presents a self-adapting fuzzy inference system. In an ordinary fuzzy inference system for agricultural land evaluation, the value intervals and weights of land variables, which are used in constructing membership functions and fuzzy inference, are often determined subjectively. This study integrates genetic algorithm with a fuzzy inference system to introduce machine learning into fuzzy evaluation of the suitability of agricultural land. The criteria for evaluation were coded as chromosomes. The performance of the fuzzy inference system on a training set was used as the fitness of an individual chromosome representing a solution for the evaluation criteria system. The generational process of GA, including selection, crossover and mutation, was repeated until a termination condition was reached. The optimized evaluation criteria were produced by decoding the final best chromosome generated by the genetic algorithm.

The fitness function and genetic operators were constructed, and the application in the study area verifies the effectiveness and reliability of the proposed model. The subsets in chromosomes were used as basic units in crossover and mutation. They could reduce the violation of the constraints for chromosome and the destruction of excellent genes, thus accelerating the convergence of the genetic algorithm. Some transformations were implemented in mutation to ensure the new individuals accorded with the constraints. The reciprocal of error rate of the fuzzy inference model on the training sample set was used as fitness for chromosomes. It could improve the efficiency of the genetic algorithm compared to using accuracy from the training set as fitness. In the application of the model to the case study area, the error rate on the training set of the evaluation criteria system decreased from 27.93% to 1.81%. The accuracy of the GA-optimized fuzzy inference model on the test set was 93.22%, much higher than the accuracy of the original model, 72.07% (=1-27.93%). This indicates that the proposed model is effective and reliable.

The GA-optimized fuzzy inference model adjusted the original criteria by self-learning, and generates a more accurate evaluation result. Value ranges of land organic content, soil depth, soil texture profile and irrigation guarantee rate were changed considerably. The result showed that irrigation guarantee rate and drainage class were lower ranked in the original weighting system. The results illustrated that 74.8% of the cultivated land area is highly suitable for paddy. This is consistent with the empirical understanding of the quality of the cultivated land in the case study area. However, a total 15 land units, 1.2% of the land area, were classified as not suitable. The generated land suitability map can be further used for the allocation of paddy cultivation and developing priority-based supplementary irrigation and drainage plans.

Future work entails investigating the generalization of the model when applying the model in more case study areas with different original criteria systems and sample sets. This study trained the fuzzy inference model from a moderate evaluation criteria system. The method presented in this study will not be applicable under the condition of having a very poor original criteria system, for example, when some important land variables are missing in the system. Fortunately, we can construct a moderate evaluation criteria system by consulting experts in most cases, so we can adjust the imperfect criteria by using the model in this study. Same to other machine learning methods, the model in this study needs representative and sufficient samples. Quantitive evaluation of the sampling scheme and the influence of the quality of the sample set on the performance of the model should be investigated in the future.

REFERENCES

Ahamed, T.R.N., Rao, K.G., Murthy, J.S.R., 2000. GIS-based fuzzy membership model for crop-land suitability analysis. *Agricultural Systems* 63(2), pp.75–95.

Avdagic, Z., Karabegovic, A., Ponjavic, M., 2008. Fuzzy Logic and Genetic Algorithm Application for Multi Criteria Land Valorization in Spatial Planning, in: Plemenos, D., Miaoulis, G., *Artificial Intelligence Techniques for Computer Graphics*. Springer-Verlag, Berlin, Heidelberg, pp. 175–198.

Braimoh, A.K., Vlek, P.L.G., Stein, A., 2004. Land evaluation for maize based on fuzzy set and interpolation. *Environmental Management* 33(2), 226–238.

Burrough, P.A., van Gaans, P.F.M., Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. *Geoderma* 77, pp.115–135.

Burrough, P.A., 1989. Fuzzy mathematical methods for soil survey and land evaluation. *Journal of Soil Sciences* 40, pp.477–492.

Burrough, R.A., Macrnillan, R.A., Van Deusen, W., 1992. Fuzzy classification methods for determining land suitability from soil profile observation and topography. *Journal of Soil Sciences* 43, pp.193–210.

Ceballos-Silva, A., Lopez-Blanco, J., 2003. Evaluating biophysical variables to identify suitable areas for oat in Central Mexico: a multi-criteria and GIS approach. *Agriculture Ecosystems & Environment* 95(1), pp.371–377.

Chang, L., Burrough, P.A., 1987. Fuzzy reasoning: a new quantitative aid for land evaluation. *Soil Survey and Land Evaluation* 7, pp.69–80.

Chen, Y., Khan, S., Paydar, Z., 2010. To Retire or Expand? A Fuzzy GIS-Based Spatial Multi-Criteria Evaluation Framework for Irrigated Agriculture. *Irrigation and Drainage* 59(2), pp.174–188.

Corona, P., Salvati, R., Barbati, A., Chirici, G., 2008. Land Suitability for Short Rotation Coppices Assessed through Fuzzy Membership Functions. *Patterns and Processes in Forest Landscapes: Multiple Use and Sustainable Management*, pp.191–211

Davidson, D.A., Theocharopoulos, S.P., Bloksma, R.J., 1994. A land evaluation project in Greece using GIS and based on Boolean and fuzzy set methodologies. *International Journal of Geographic Information Systems* 8(4), pp.369–384.

FAO, 1976. A framework for land evaluation. Soils Bulletin 32. FAO, Rome.

Fugger, W.-D., 1999. Evaluation of potential indicators for soil quality in Savanna soils in Northern Ghana. Ph.D. thesis, Georg-August University, Gottingen.

Jiao Limin, Liu Yaolin, 2007. The Model of Land Suitability Evaluation Based on Computational Intelligence. *Geo-Spatial Information Science*10(2), pp.151–156.

Jiao, L.M., 2006. Land evaluation models based on computational intelligence. Doctorial dissertation of Wuhan University.

Jiao, L.M., Liu, Y.L., 2004. Application of Fuzzy Neural Networks to Land Suitability Evaluation. *Geomatics and Information Science of Wuhan University* 6, pp.513–516.

Joss, B.N., Hall, R.J., Sidders, D.M., Keddy, T.J., 2008. Fuzzy-logic modeling of land suitability for hybrid poplar across the Prairie Provinces of Canada. *Environmental Monitoring and Assessment* 141(1-3), pp.79–96.

Kurtener, D., Torbert, H.A., Krueger, E., 2008. Evaluation of agricultural land suitability: Application of fuzzy indicators. *Computational Science and Its Applications - Iccsa 2008, Pt 1, Proceedings* 5072, pp.475–490.

Liu, Y.L., Jiao, L.M., 2008. *Principles, methods, and software for land evaluation*. Beijing: Science Press (In Chinese).

McBratney, A.B., Odeh, I.O.A., 1997. Application of fuzzy sets in soil science: Fuzzy logic, fuzzy measurement and fuzzy decisions. *Geoderma* 77, pp.85–113.

Ministry of Land and Resources, China, 2003. Regulations for classification on agricultural land. Beijing: Ministry of Land and Resources, China (In Chinese).

Mouton, A.M., Alcaraz-Hernández, J.D., De Baets, B., Goethals, P.L.M., Martínez-Capel, F., 2011. Data-driven fuzzy habitat suitability models for brown trout in Spanish Mediterranean rivers. *Environmental Modelling & Software* 26, pp.615-622

Pradhan, B, Lee, S, 2010. Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modeling, *Environmental Modelling & Software* 25, pp.747-759.

Reshmidevi, T.V., Eldho, T.I., Janna, R., 2009. A GIS-integrated fuzzy rule-based inference system for land suitability evaluation in agricultural watersheds. *Agricultural Systems* 101(1-2), pp.101–109.

Rowe, G., Wright, G., 1999. The Delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting* 15(4), pp.353–375.

Saaty, T.L., 1980. The Analytical Hierarchy Process. McGraw Hill, New York.

Tang, H.J., Debavye, J., Van Ranst, E., 1991. Land suitability classification based on Fuzzy set theory. *Pedologie* 41, 277–290.

Tang, H.J., Van Ranst, E., 1992. Testing of fuzzy set theory in land suitability assessment for rainfed grain maize production. *Pedologie* 42, pp.129–147.

Zadeh, L.A., 1965. Fuzzy sets. *Information and Control* 8, pp.338–353.

Zhang, B., Zhang, Y., Chen, D., White, R.E., Lu, Y., 2004. A quantitative evaluation system of soil productivity for intensive agriculture in China. *Geoderma* 123(3-4), pp.319–331.

ACKNOWLEDGEMENTS

This research was funded by the National Natural Science Foundation of China (40721001, 40901188), National High Technology Research and Development Program (2007AA12Z225), and the Fundamental Research Funds for the Central Universities (4082002).

Porphyry Copper Mineral Prospectivity Mapping using Interval Valued Fuzzy Sets TOPSIS method in central Iran

Ali Reza Jafari Rad[1], Wolfgang Busch[2]

Institute of Geoinformation and Mine Surveying, Clausthal University of Technology, Erz Street 18, 38678 Clausthal-Zellerfeld, Germany
Telephone: (49) 5323 722076
Fax: (49) 5323 722479
Email[1]: alirad@yahoo.com
Email[2]: wolfgang.busch@tu-clausthal.de

1. Introduction

Geospatial Information System (GIS) provide tools to quantitatively analysis and combination of datasets from geological, geophysical, remote sensing and geochemical surveys for decision-making processes. Excellent coverage of well-documented and good quality data enables testing of variable exploration modeling in an efficient way.

The study area of this research is the most important part of Cu (Mo) porphyry – type mineralization belt in Iran. There are some well–known porphyry copper deposits in this region like Sarcheshmeh and Meiduk mines, but certainly there are same grounds to search for new porphyry deposits.

The risks of developing mineral resources need to be known as accurately as possible, with regarding to all features those are effective in mineralization. These features can be recognized respect to Critical Genetic Factors (CGF's) using Critical Recognition Criteria (CRC) for each type of mineralization. CGF's can be employed for designing a Conceptual Genetic Model (CGM). Evidence maps create basis on CGM and then integrate together for production of Mineral Prospectivity Map (MPM). This map categorizes the areas based on their exploration importance.

There are several techniques for creation of MPM. Interval Valued Fuzzy Sets (IVFSs) TOPSIS method was applied in this research. This method as a knowledge-driven method, allocate appropriate weights to layers basis on the effective membership, non membership, and non-certainty. The fundamental concept of TOPSIS is that the chosen alternatives should have the shortest distance from the positive ideal points (A^*) and the farthest distance from negative ideal points (A^-).

2. Study area, Data and software

Iran is located in Alpine-Himalaya orogenic and metallogenetic belt formed after Tethys collision, and therefore has a high potential for different types of minerals. Conventionally a unique Volcano–Plutonic–Arc (VPA) is considered to be formed by subduction of Mesozoic Tethys oceanic crust, but new evidences show that there are different oceanic basins, and associated arcs. One of the most important VPA is Kalkafi Sarcheshmeh–Kharestan (Samani & Ashtari 1992), where the study area of this research

is a part of this VPA. The study area is located at northwest of Kerman province in central Iran (fig. 1).

The utilized data include digital geology maps, ASTER and LANDSAT ETM imagery, airborne geophysics (magnetics, radiometrics, and electromagnetics), geochemical stream samples and heavy minerals data.

ARCGIS, ENVI and GEOSOFT software were developed for data preparation, analysis and modeling in this research.

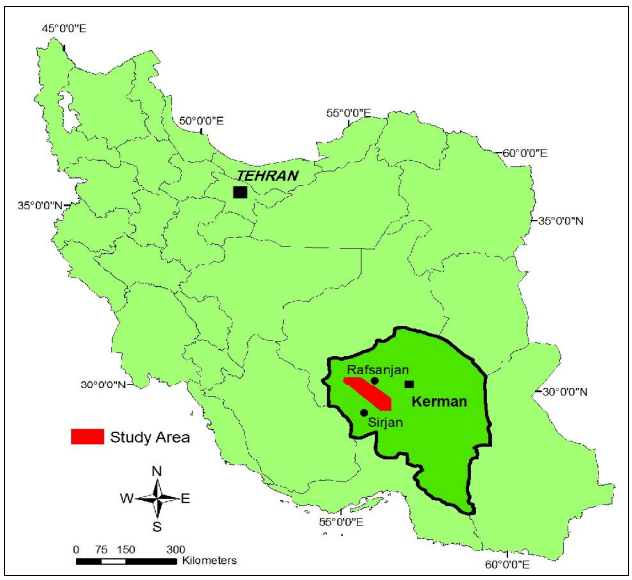


Figure 1. Location of study area in Iran

3. Geological and metallogenetic setting

The geological formations of the study area consist of ranging from the Cretaceous up to the very recent Quaternary sediments.

The most significant features, related to mineralization, are the sedimentation, magmatic activity and structural displacement that occurred during the Tertiary. The granodiorite and diorite are the most common intrusive rocks. The porphyry copper mineralization is related to regional scale faults (more than 20 km length) and the most important trends for mineralization are N-S, NE-SW, E-W, and NW-SE respectively. At places, where two fault systems intersect, the intrusive bodies are frequently hydrothermally altered. These locations have the best situation for porphyry mineralization. Hydrothermal alteration zoning follows the Lowell and Guilbert pattern.

4. Data layers processing and analysis

According to geological and metallogenetic setting, all features that are important in porphyry copper mineralization were recognized and CGM was presented and the predictor maps were performed. These maps consisting of five thematic layers as follows:

1) Geological thematic layer: The original geology map contained 80 different lithologies. Basis on the rock types and age, these units were classified to 32 groups and then were reclassified to 6 classes according to their importance in mineralization.

2) Structural thematic layer: The structure features were extracted from: a) geological maps; b) satellite imagery and c) geophysics data. After selecting regional faults and calculating the azimuth, they were divided to different trending and buffered up to 1500 meter. According to buffer distance and trending, they were classified to 7 classes.

3) Alteration thematic layer: As a result of satellite imagery interpretation, phyllic, advance argillic, argillic and propylitic alteration zones were identified and classified according to Lowell and Guilbert model.

4) Geochemistry anomalies thematic layer: Several geochemical anomalies were found out during geochemical data analysis and classified basis on the zonation of paragenesis elements.

5) Geophysics thematic layer: The results of geophysical data interpretation consist of: intrusive bodies, alteration areas; and lineaments. The anomalies were classified to 4 classes' base on the existence of zonation, and the correlation of anomalies with geological features.

5. IVFSs TOPSIS method for MPM

Ting-Yu Chen and Chueh-Yung Taso (2007) applied the IVFSs TOPSIS in decision analysis. Also there are some papers that the authors applied multiple criteria decision-making in their research like: Chen et al. (2000), Jahanshahloo et al. (2006).

IVFSs TOPSIS method has been specified in support of MPM in this research for the first time. Using this method all significant factors for a knowledge driven modeling system, like allocation Fuzzy Membership (FM), Priority Weights (PW) and predefined targets can be considered.

This technique can be performed in following steps:

1) Classification of each thematic data layers basis on the GCM.

2) Allocation of FM to each class of data layers (table 1 as an example for geology layer).

FM_Geology		
Class	Fuzzy_1	Fuzzy_2
1	0.7	0.9
2	0.5	0.7
3	0.3	0.5
4	0.2	0.25
5	0.05	0.1

Table 1. FM of geology layer

3) Assign PW to each data layer (table 2).

W_Geology	
W_1	W_2
0.8	0.9

Table 2. PW for geology layer

4) Multiplication of FM and PW (table 3).

Geology		
class	F1*W1	F2*W2
1	0.56	0.81
2	0.4	0.63
3	0.24	0.45
4	0.16	0.225
5	0.04	0.09

Table 3. Multiplication of FM and PW for geology layer

5) Calculation effective membership (a value), non-membership (b value) and non-certainty (c value) for each class of data layers (table 4).

	Geology		
	a	b	c
	0.56	0.19	0.25
	0.4	0.37	0.23
	0.24	0.55	0.21
	0.16	0.775	0.065
	0.04	0.91	0.05
A*	0.56	0.19	0.25
A-	0.04	0.91	0.05

Table 4. “a”, “b” and “c” values for geology layer

6) Creation several Raster Images (RI) according to “a”, “b” and “c” values for each data layer (fig. 2).
7) Calculation of positive ideal point (A*) and negative ideal point (A^-) for “a”, “b” and “c” values in each data layer (table 4).
8) Measurement distance from (A*) and (A^-) for each layer. For measuring distance in this research, Szmidt and Kacprzyk’s equation was specified in the form of equation 1 and 2.

S* = 1/2 [|(RI of “a value” - A*) + (RI of “b value” - A*) - (RI of “c value” - A*)|] (1)

S^- = 1/2 [|(RI of “a value”- A^-) + (RI of “b value”- A^-) - (RI of “c value” - A^-)|] (2)

9) Calculation of closeness for preparation MPM using equation 3:

$$C_i^* = \frac{S_i^-}{S^* + S^-} \quad (3)$$

5. Conclusion

Developing mineral resources should start at the pre-discovery stage and continue through feasibility to the development stage. Integrating of predictor maps using GIS allows more probabilistic data analysis techniques and reduces costs and time. Basis on the IVFSs TOPSIS method, calculation of closeness at the end step of procedure present a MPM that demonstrates the favorable area for pre-discovery exploration (fig. 3). The original MPM includes different numerical classes. It can be reclassified to descriptive values based on the big jumps in numerical values (fig. 4). First class targets of this research contain 22 regions (0.76 percent of study area); include 13 old mining areas and

9 new areas. Setting of all old mining areas inside the first class targets, and field observations proves the efficiency of this method. This method can be used in the similar geological and metallogenetic locations in northwestwards and southeastwards of the study area in Iran.

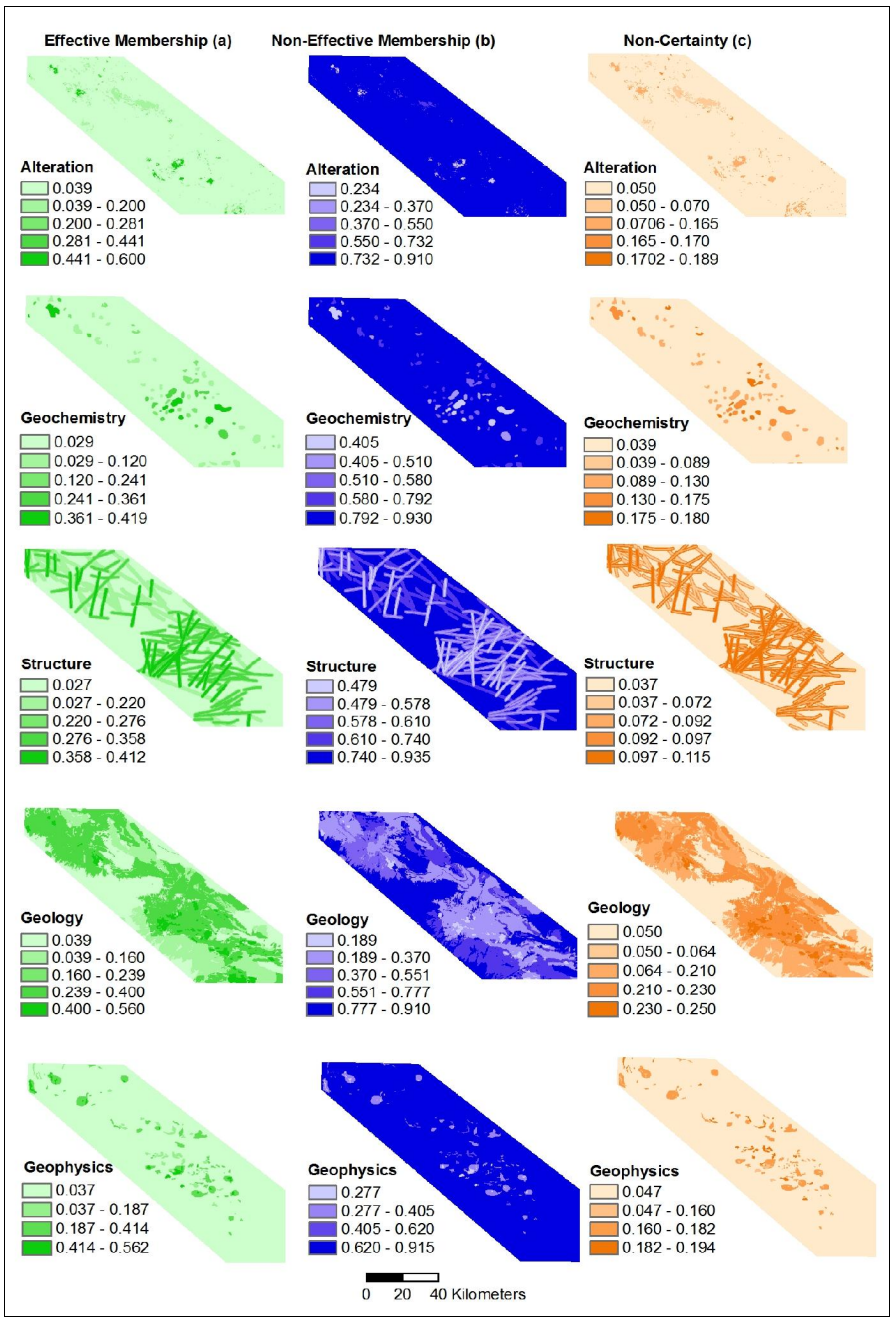


Figure 2. Raster images based on “a”, “b” and “c” values

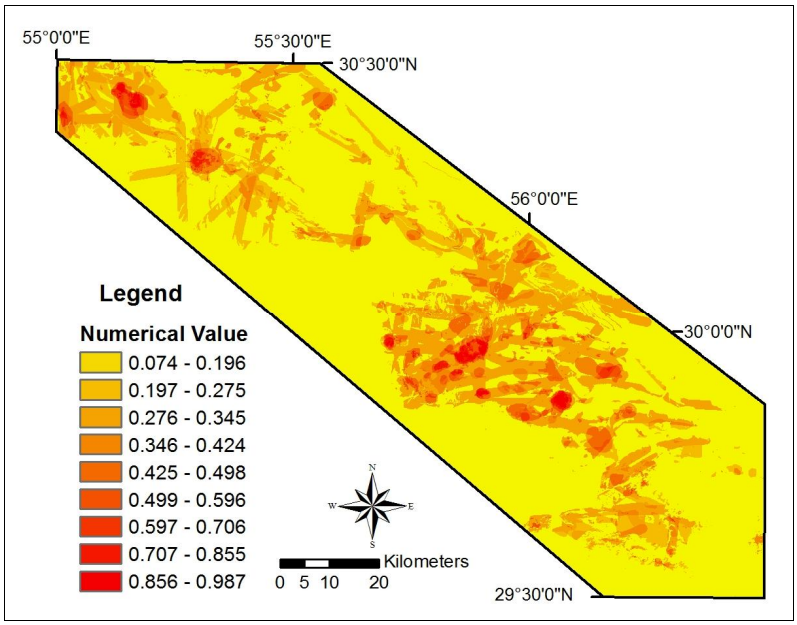


Figure 3. MPM using IVFSs TOPSIS method

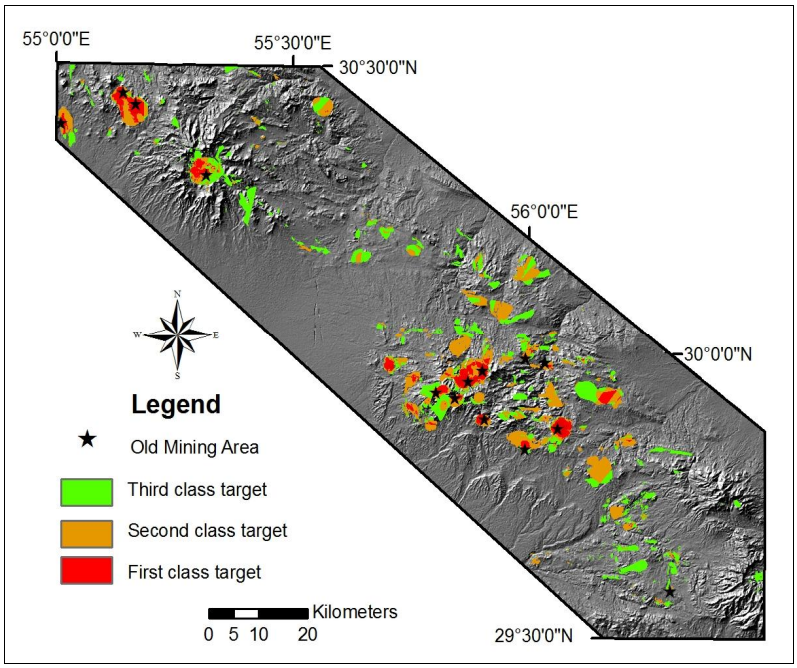


Figure 4. Descriptive priority map

6. References

Chen T. Y. & Tsao C. Y., 2007, "The interval-valued fuzzy TOPSIS method and experimental analysis", Fuzzy Sets and Systems, 159, 1410-1428.

Chen C., 2000, "Extensions of the TOPSIS for group decision-making under fuzzy environment", Fuzzy Sets and Systems 114, 1–9.

Jahanshahloo G. R., Hosseinzadeh Lotfi F. & Izadikhah M., 2006, "An algorithmic method to extend TOPSIS for decision-making problems with interval data", Appl. Math. Comput. 175, 1375–1384.

Rutkowski L., 2008, "Computational Intelligence, methods and techniques, Springer.

Samani, B. & Ashtari s., 1992, "Evaluation of Volcano – Plutonic – Arc in central Iran, formed by subduction of Mesozoic Tethys oceanic crust", Geosciences. GSI 1 (4).

Yang T. & Hung C. C., 2007, "Multiple-attribute decision making methods for plant layout design problem", Robotics Comput. Intregrated Manufacturing 23, 126–137.

INTEGRATED USE OF MULTI-TEMPORAL LANDSAT IMAGES &GIS FOR MAPPING & MONTRING WATERLOGGING &SALINITY IN IRRIGATED LANDS

Azhar Al-Baldawi **Dr Rafea A. Hasan**
Baghdad/Unv. **Mustansryh /Unv.**
Az_rts@yahoo.com
gis_rs2008@yahoo.com

KEYWORDS : Salinity , monitoring, landsat images , GIS, NDVI,SI, WI

ABSTRACT

Salinity has always been a major issue in both old Mesopotamian and modern-day Iraq and it was already recorded a cause of crop yield destroying & reductions, it was estimated that half of the irrigated areas in central and southern Iraq were degraded due to water logging and salinity .since the country's economic development is linked to agriculture, therefore resource planners are always looking for efficient and economic methods to bring the saline areas under cultivation. Therefore, this study is intending to improve the monitoring capability afforded by remote sensing to monitoring, mapping, and analyzing the desertification in study area in Mesopotamian plain by using available temporal landsat images for three periods MSS 1976, TM 1990, and ETM+ 2002 scenes covering the study in addition to the ancillary data and field observations.

In order to identify the vegetation cover, water bodies, and saline soil cover and there changes which took place over the three periods, Normalized Difference Vegetation Index(NDVI), Salinity Index (SI), and water body index (WI) , were used and two methods for changes identification. were implemented Firstly, direct detection of change in indices images between different years analysed by use of visual

interpretation in addition to statistical analysis. Secondly, differencing change detection analysis was applied to determine and analyses the land cover changes over the three periods..

GIS analytical tools, were used such as buffering, overlaying the related land cover layers for monitoring the changes took place during the addressed periods.

Findings of the present study show, the rate change of desertification processing and land degradation during the periods 1976 and 1990 , are positive, while the rate of change for period 1990 to 2002 negative changing, 1976 represent the heavily damage and threats of desertification process , while 1990 shows positive statues and re-growth period. In addition the central and southern part of the study area classified as severe of desertification and the other part subjected to desertified .

REFERENCES

- Al Baldawi.A.(2004)using Remote Sensing Techniques to study the salt affected soil ,Journal of Remote Sensing Issue no. ,17Dec2004GORS,Ddamscus ,Syria
- AL Baldawi A. , Hassan R. (2008) Integrated Geo-Spatial Techniques For Land Degradation Assessment in Mesopotamian Plain (Iraq) , paper submitted to 5th International Conference on Land Degredation Bari, Italy 18-22 Sep.2008
- AL Baldawi A., Hassan R. (2009) Integration of RS & GIS to study and mapping sand encorchment and mobile dunes in Iraq ,paper submitted to 4th International Congress Geotunis 2009, 16-20 Dec .,Tunisia .
- AL Baldawi A., (2010) , Desertification study of Dalmaj area in Mesopotamian plain using Remote Sensing techniques , PhD thesis , Baghdad University .

- Al-Khaier, F.,(2003). Soil salinity defection using satellite remote sensing, M.Sc. thesis, ITC, p 49.
- Briggs, M., (1996). Riparian ecosystems recovery in arid lands, strategies and references. Univ. of Arizona Press, Tucson.
- Fadhil, A.M., (2004). Land degradation detection using geo-information technology for some sites in Iraq. Soil and water Sci.Dept. Agriculure College, Salahaddin – Erbil-Iraq
- Foody, G. M., (2001), Accuracy of thematic maps derived from remote sensing. Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Science, edited by G. B. M. Heuvelink and M. J. P. M. Lemmens (Delft, Holland: Delft University Press), pp. 217–224.
- Goovaerts P., (1997). Geostatistics for natural resources evaluation. Oxford University Press, New York. Harahsheh and Tateishi (2001), Environmental GIS Database for Desertification Studies in West Asia , Center for Environmental Remote Sensing (CEReS), Chiba University

- Hick P.T. & Russell W.G.R., (1990). Some spectral considerations for remote sensing of soil salinity. Australian Journal of soil research, 28(3), 417-431.
- karavanova, E.I., Shrestha, D.P., and Orlov, D.S., (1993). Application of remote sensing techniques for the study of soil salinity in semi arid Uzbekistan. Soil science department .Lomonosov, Moscow state university, 119899, Russia Stocking, M., (1995). Soil erosion and land degradation. In:

O'Riordan, T. (Ed) Environmental science for environmental management. London, Longman, 233-243.

SPACE-TIME TREND ANALYSIS OF HEALTH OUTCOMES: PROSTATE CANCER LATE-STAGE DIAGNOSIS IN FLORIDA

P. Goovaerts [a*], H. Xiao [b]

[a] BioMedware, 3526 W Liberty, Suite 100, Ann Arbor, MI 48103 – Goovaerts@biomedware.com
[b] College of Pharmacy, Florida A&M University, Tallahassee, FL 32307 – hong.xiao@famu.edu

Commission VI, WG VI/4

KEY WORDS: Joinpoint regression, binomial kriging, cluster analysis

ABSTRACT:

The increase in computational power and storage capacity of computers, combined with the growing availability of geocoded data, has increased dramatically the amount of information processed in health studies, making it difficult to understand, to explore, and to discover interesting patterns within the data. This paper introduces a suite of techniques for the visualization and analysis of timeseries of health data, including 3D display of health outcomes in a combined time and geography space, binomial kriging to filter noise in the data, joinpoint regression analysis to model time trends, and hierarchical cluster analysis to classify geographical units according to their temporal trends. The approach is applied to annual percentages of prostate cancer late-stage diagnosis recorded in Florida since 1980s. White males have experienced a 50% decline in the state-average proportion of late-stage diagnosis. This drop started in the early 1990s when prostate-specific antigen (PSA) test became widely available. Geographical disparities were substantial at that time, which suggests disparities in the impact of the new screening procedure, in particular as it began available. The gap among Florida counties is narrowing with time as the percentage of late-stage diagnosis is decreasing. One outlier is the Big Bend region of Florida where the decline in percentage of late-stage diagnosis has been the slowest in all Florida.

1. INTRODUCTION

Interpretation of cancer incidence and mortality rates in a defined population requires an understanding of multiple complex factors that likely change through time and space, and interact with the different types and scales of places where people live. These factors include the prevalence of risk factors in the population, changes in the use of medical interventions to screen and treat the disease, and changes in how data are collected and reported. Analyzing temporal trends in cancer incidence and mortality rates can provide a more comprehensive picture of the burden of the disease and generate new insights about the impact of various interventions (Potosky *et al*., 2001).

Prostate cancer is the most frequently diagnosed non-skin cancer and the second leading cause of male cancer-related death in the US. As exemplified for Florida (Fig. 1,) prostate cancer mortality and late–stage diagnosis started declining after 1991. According to some studies, this decline in mortality is due to early detection (PSA screening) although screening for prostate cancer is still controversial. The analysis of temporal trends outside a spatial framework is however unsatisfactory, since it has long been recognized that there is significant variation among U.S. counties and states with regard to the incidence of cancer. Figure 2 highlights that the statewide decrease in prostate cancer late-stage diagnosis over the period 1981-2007 encompass significant disparities among counties. Visualizing, analyzing and interpreting these geographical disparities will bring important information and knowledge that will benefit substantially cancer epidemiology, control and surveillance (Goovaerts, 2010). For example, to what extent could the geographic disparities in Figure 2 be explained by differences in socio-economic status and screening practices?

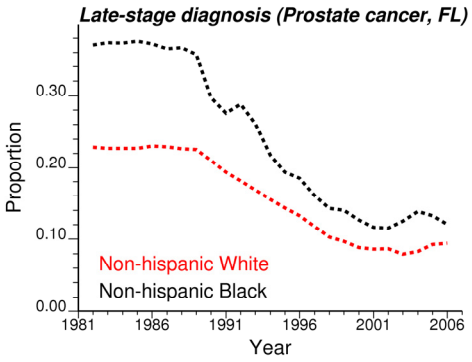


Figure 1. Evolution of the percentage of late-stage cancer diagnosis recorded annually for prostate cancer for the entire state of Florida (period 1981-2007).

2. 3D VISUALIZATION OF THE DATA

Number of cases of prostate cancer and associated stage at diagnosis recorded yearly from 1981 through 2007 for each county of Florida and two ethnic subgroups (White, Black) were downloaded from the Florida Cancer Data System website. Percentages of late-stage diagnosis were computed for each year, county and ethnic group. These rates were then processed using binomial kriging (Goovaerts, 2009) to filter the noise caused by the small number problem. Geographical and temporal changes were visualized using three-dimensional space-time displays (Goovaerts, 2010) of the data to take full advantage of human visual perception that is fundamentally three dimensional; see Figure 2.

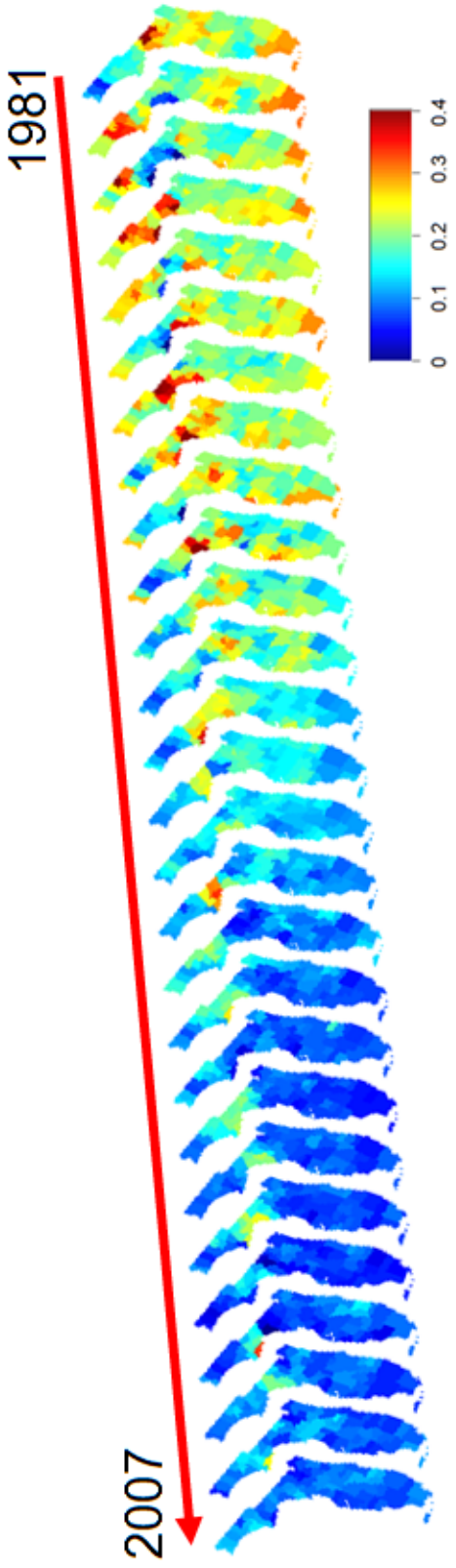


Figure 2. Three-dimensional representation of yearly percentages of late-stage prostate cancer (white males 65 years or older) which were noise-filtered at the county level using binomial kriging (Goovaerts, 2009).

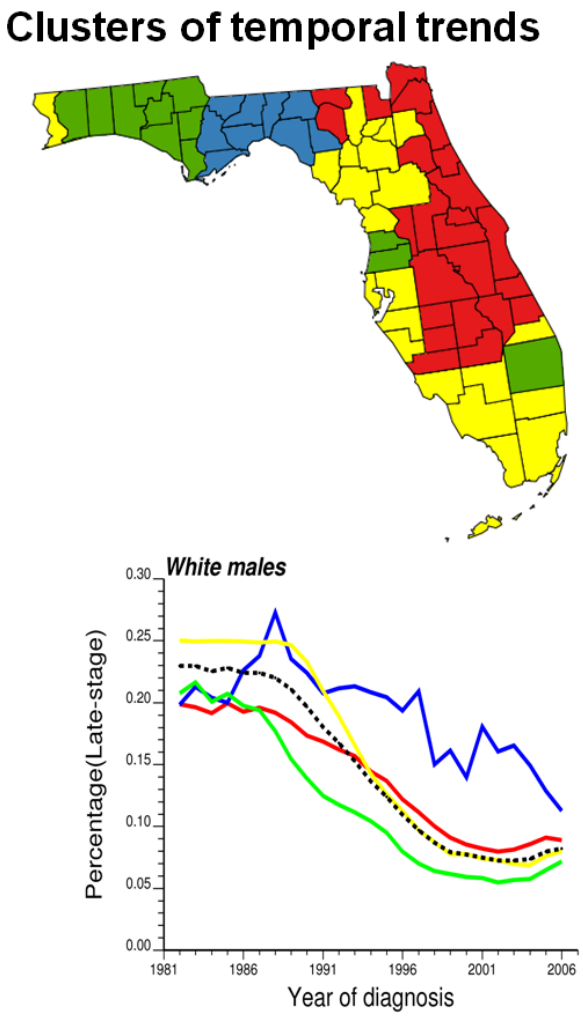
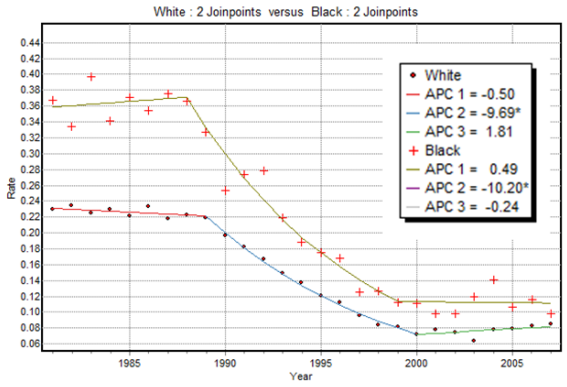


Figure 3. Grouping of counties (Cluster analysis, Ward's minimum-variance method) based on the similarity of their temporal trends in percentages of late-stage diagnosis for white males (dashed black curve represents Florida).

3. CLUSTER ANALYSIS

During the analysis of large space–time–attribute datasets, users may have difficulty perceiving, tracking and comprehending numerous visual elements that change simultaneously, such as the 67 timeseries in Figure 2. One solution (Ward, 2004) illustrated in Figure 3 is to reduce the data size being displayed by grouping timeseries into subsets (e.g. aggregation or clustering). In this case, collective characteristics of the grouped data are visualized, revealing clear differences among regions of Florida. While some regions experienced a decrease coinciding with the introduction of PSA test, others (in particular the area around Tallahassee, blue color) display a flat trend and higher rates of late-stage diagnosis in this last decade. Note also the recent increase in percentage of late-stage diagnosis in most regions.



Cohort	Joinpoint	Estimate	Lower CI	Upper CI
White	1	1989	1988	1990
White	2	2000	1999	2002
Black	1	1988	1986	1993
Black	2	1999	1996	2002

Figure 4. Joinpoint regression models fitted to the timeseries of percentage of late-stage prostate cancer diagnosis for the entire state of Florida shown in Fig. 1. Two joinpoints were fitted to each curve and the timing of significant changes for the two races is fairly similar. Both ethnic groups display a significant decline in the 90's (Annual Percentage Change=-9.69 and -10.20).

4. JOINPOINT REGRESSION

Figure 1 provides a visual assessment of the temporal trends in percentage of late-stage diagnosis for two ethnic groups. A quantitative analysis of temporal trends in incidence rates can be conducted using Joinpoint regression (Kim *et al*., 2000) that is currently implemented in the public-domain software developed at NCI (http://srab.cancer.gov/joinpoint/). The basic idea is to model the time series using a few continuous linear segments (see an example for the Florida timeseries in Fig. 4). Line segments are joined at points called joinpoints which represent the timing for a statistically significant change in rate trend (e.g. 1989 and 2000 for white males, Fig. 4). The number of joinpoints, as well as the parameters of the piecewise linear regression, are estimated through an iterative procedure that tests whether models of increasing complexity (i.e. including more joinpoints) provide a significantly better goodness-of-fit than simpler models.

Joinpoint regression was also conducted for each county and they were grouped based on the similarity of their temporal trends in the annual rate percentage change (APC), which corresponds to the slope of the linear segments in the piecewise linear regression (e.g. see Fig. 4). Unlike in Figure 3, individual timeseries are not aggregated but rather are displayed according to their cluster allocation in Figure 5. Each row corresponds to a particular county and each pixel to a particular year; the color scale indicates the APC value that ranges between -24 (decrease in late-stage) to 47.8 (increase in late-stage). For example, the percentage started declining early for the counties in Cluster #5 (Panhandle) followed by a recent increase. Cluster #3 shows a brief and steep decline in the mid nineties, while this decline occurred at a slower pace and over a longer time period for counties in Cluster #4.

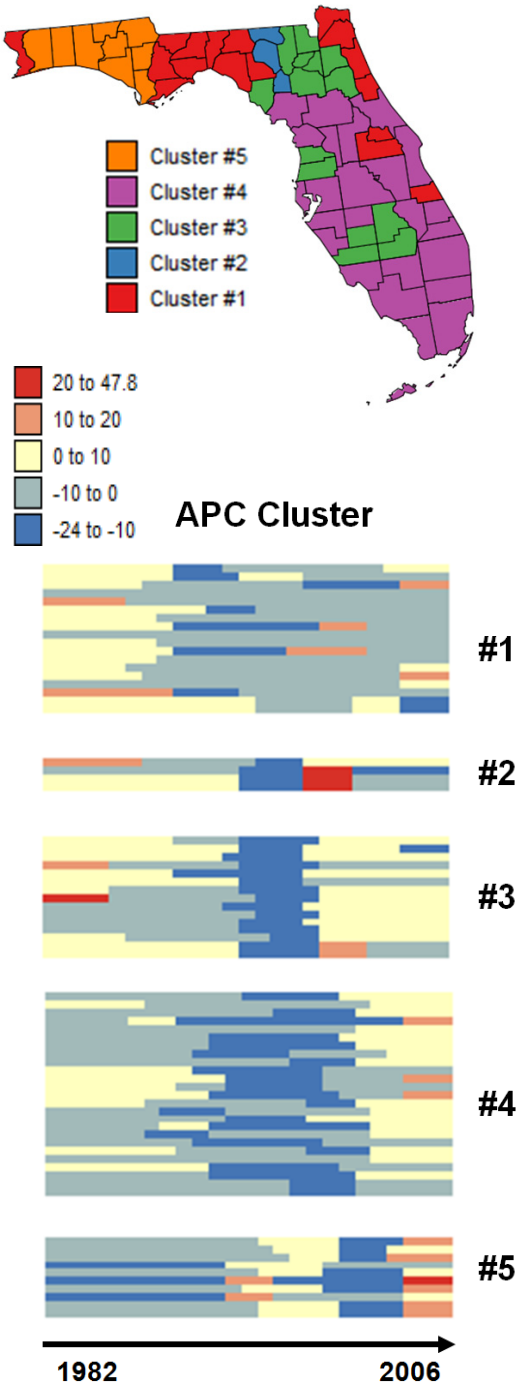


Figure 5. Grouping of Florida counties based on the similarity of their temporal trends in the annual rate percentage change (APC) in prostate cancer late-stage diagnosis estimated by joinpoint regression. Individual time series of APC are displayed as horizontal strings and ordered according to their allocation to one of the five clusters.

5. CONCLUSIONS

This paper presents two approaches (analysis of original timeseries and joinpoint regression models) for summarizing and visualizing spatio-temporal changes in health outcomes.

Figure 5 shows that the onset and rate of decrease in late-stage diagnosis for prostate cancer varies greatly across Florida, in particular the decline started early and was of larger magnitude in the Panhandle while the decline is much smaller in the Big Bend region. Current research is using individual-level data to explore in mode details the impact of covariates such as distance to urologists and radio oncologists, and socio-economic status. This approach can be easily generalized to other states and cancer sites, with clear applications in (a) monitoring and surveillance of cancer incidence and mortality, (b) the generation of hypotheses for in depth individual studies of risk factors that are causal, or impact survival; and (c) establishing the rationale for targeted cancer control interventions.

6. REFERENCES

Goovaerts, P., 2009. Combining area-based and individual-level data in the geostatistical mapping of late-stage cancer incidence. *Spatial and Spatio-temporal Epidemiology*, 1, pp. 61-71.

Goovaerts, P., 2010. Three-dimensional visualization, interactive analysis and contextual mapping of space-time cancer data. *Proceedings of 13th Agile International conference*, Guimarães, Portugal, May 2010. http://home.comcast.net/~pgoovaerts/Agile_2010.pdf (accessed 28 Jan. 2011).

Kim, H.J., Fay, M.P., Feuer, E.J. and D.N. Midthune, 2000. Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine*, 19, pp. 335-351.

Potosky, A.L., Feuer, E.J. and D.L. Levin, 2001. Impact of screening on incidence and mortality of prostate cancer in the United States. *Epidemiologic Review*, 23(1), pp. 181-186.

Ward, M.O., 2004. Finding needles in large-scale multivariate data haystacks. *Computer Graphics and Applications*, 24(5), pp. 16-9.

7. ACKNOWLEDGEMENTS

This research was funded by grants R43CA150496-01 and R44CA132347-02 from the National Cancer Institute, as well as grant #RSGT-10-082-01-CPHPS from the American Cancer Society. The views stated in this publication are those of the authors and do not necessarily represent the official views of the NCI.

COUPLING OUTBREAK DETECTION OF SPATIALLY CLUSTERED ASSOCIATIONS AND DATA REDUCTION PRINCIPLES

D. G. Leibovici [a], J. Swan [b], M. Jackson [a]

[a] Centre for Geospatial Science, University of Nottingham, Innovation Park, Nottingham, NG7 2TU – (didier.leibovici, mike.jackson)@nottingham.ac.uk
[b] Dept. of Computing Science, University of Nottingham, Jubilee Campus, Nottingham, NG7 – jerry.swan@nottingham.ac.uk

KEY WORDS: **S**patio-temporal analysis, co-occurrence count data, spatial association, data reduction, spatial entropy

ABSTRACT:

A dream goal of spatio-temporal event data analysis is to provide evidence and description of spatially and temporally clustered associations of attributes coming from a potentially large amount variables that the thematic specialist in environmental sciences, including social science, is often looking at for monitoring purposes. The motivation is in explaining these clusters by the variables used and from cofactors or covariates measurements, which can be used to manage or mitigate their occurrences. Disease outbreaks and associated factors, biodiversity losses and ecological conditions, crime spots and social descriptions are few examples. Recently Leibovici et al. (2011b) proposed an exploratory approach to discover and locate spatially clustered associations (Scan*k*OO analysis). An extension of this methodology for outbreak detections of multivariate multinomial events has been also proposed. If this method provides a spatio-temporal detection, the "mining" is limited to the chosen spatio-temporal paradigm. No help to select or describe the best variables and attributes responsible of the clusters are devised, besides *a priori* and *a posteriori* analyses. Using the same approach with a focus on spatial entropy and conditional spatial entropy, Leibovici et al. (2011a) provided another method (SelSOO*k*) allowing variable selection but without localisation of the spatial associations. In this paper we investigated the combination of variable selection (variable mining) and/or data reduction with spatial clustering principles illustrated in the Scan*k*OO method.

1. INTRODUCTION

1.1 Spatio-Temporal Monitoring

A monitoring study implicitly refers to temporal data, and focuses also on the spatial variations and on the spatial patterns of the records. Besides the sampling strategy, or the data capture available, for monitoring, which is an issue in itself, it means that the support (the spatio-temporal realm) is intrinsically interacting with the monitored variables. Discovering the spatial associations of the variables, or, of their attributes when considering discrete data, is the goal of the scientist. Often a large amount of variables is looked for, either intentionally measured as part of the design or gathered during or after the monitoring if it has stopped, for example with data available on the web. If a *plethora* of spatial statistical methods to discover global and local spatial autocorrelation or clustering of events, have been successfully extended to the spatio-temporal domain, e.g. Hardisty and Klippel (2010), Tango et al. (2011) for some recent references, they do not deal with variable interactions (besides covariate adjustment), neither with data reduction techniques in order to reduce the number of variables to consider. The challenge is double as the method used needs to provide evidence of a spatial-temporal association but also a description of the variables responsible of it.

For event data, the evidence of local variable association, observed spatially, can be termed clustered associations of attributes as in Leibovici et al. (2011b). The idea of analysing co-located events has also been used in spatio-temporal databases, introducing also a spatial statistic viewpoint (Huang et al. 2004). The motivation is in explaining the clusters by describing the associations of the variables, but also from cofactors and covariates measurements, which can be used to manage or mitigate the occurrences of the clusters. Disease outbreaks and associated factors, biodiversity losses and ecological conditions, crime spots and social descriptions are few examples.

1.2 Outbreak Detection

The management and decision making is based on the data time series and on detecting special events, and more likely an outbreak for pre-specified events. These outbreaks can be commonly determined by a threshold on the number of events or cases (in epidemiology), or, by comparison to the remainder of the data: time specific outbreak and spatio-temporal specific outbreak. The first type of outbreaks is more related to alarms, set for a specified level defining a degree of abnormality, while the second type can be semantically defined as a stochastic outbreak. The clustering methods mentioned above deal with stochastic outbreak detection, nonetheless to our knowledge only few methods such as the Scan*k*OO method described in Leibovici et al. (2011b) can analyse more than one variable at a time; the SaTscan method can analyse a multinonmial data (Jung *et al.* 2010).

1.3 Data reduction

If the Scan*k*OO or its extension to outbreak detection deals with multivariable data, it doesn't explicitly provide a variable description analysis of the obtained clusters, neither a mean to discard not determinant or not influential variables during the optimisation. Nonetheless, the SelSOO*k* method (Leibovici et al., 2010b), which is based also on a spatial entropy index derived from multiway co-occurrence data table, proposes a regression tree like variable selection linked to the global spatial

association criterion, and the CA*k*OO method (Leibovici, 2010) uses a generalisation of correspondence analysis to the analysis of a multiway table to describe in a data reduction way (correspondence analysis can be seen as a particular singular value decomposition) the global spatial association of *k* variables.
The aim of this paper is to take benefit of these different methods, as they are based on the same approach of using a multivariable multivariate co-occurrences data, in order to propose solutions to detect locally and explore the multivariate-spatio-temporal interactions.

2. COUPLED MINING

The methodology can be characterised as a doubled data mining technique because of the different mining domain: the spatio-temporal domain and the variable domain. The term coupled may be more appropriate as these domains are interacting and the optimisation looked for has to be inter-dependent on these two fronts.

2.1 A Priori and A Posteriori Coupling

A low coupling procedure is to perform an a priori analysis in order to select the best variables to be analysed by the Scan*k*OO method then to perform an a posteriori analysis to describe the hot-spots highlighted. A basic coupling approach implies chaining methods in a workflow alternatively focusing on one domain and another.

2.2 Coupling Optimisation Principle

A higher coupling procedure will be performed if the workflow described above can be iterated either at a macro level (a direct extension of a priori a posteriori coupling), or, within the optimisation when the different methods cannot be fully separated: coupling at a micro level. In the next two subsections we describe macro and micro level analyses for a high coupling between Scan*k*OO and respectively the SelSOO*k* and CA*k*OO methods alternating variable selection/description and sampling selection.

2.3 The ScanSelSOO*k* method

2.3.1 Macro level: After initialisation of the choice of a set of *v* variables (among an original set of $p > v$ variables) from a SelSOO*k* analysis (see Figure 1), that is performing a selection of the best variables optimising a joint or a conditional spatial entropy index (see Leibovici et al. 2011a for details), the algorithm is an iterative process between selecting hot-spots used as a spatio(-temporal) sampling and re-selecting the best variables based on these local observations.

2.3.2 Micro level: A pseudo-micro level analysis, deepening the above algorithm, can be performed by discarding the *v* variables for a second set of best variables. That is performing the macro ScanSelSOO*k* on $(p - v)$ variables as new original set, then leading to a *v'* second best set.

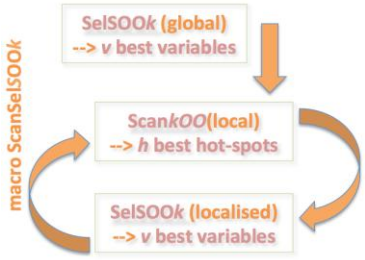


Figure 1. Algorithm of high coupling at macro level between SelSOO*k* and Scan*k*OO

2.4 The ScanCA*k*OO method

2.4.1 Macro level: This analysis is very similar to the algorithm described on Figure 1 but using CA*k*OO instead of SelSOO*k*. The selection would come after selecting (like in a PCA) the best components and then the variables building these components.

2.4.2 Micro level: The data reduction method allows a micro level integration by coming back to the principal tensor algorithm at the base of the CA*k*OO method (Leibovici, 2010). In few words, the Scan*k*OO is performed within the principal tensor decomposition optimisation, which is therefore constrained at each step before extracting a new principal tensor: the orthogonal projection is dependent on the obtained sampling (due to hot-spot). The resulting final decomposition cannot be called a CA*k*OO decomposition but can be classified as a non-linear data reduction method.

3. DISCUSSION AND CONCLUSION

Results of the different methods briefly described here will be presented with examples on spatial data (used in the quoted papers as a comparison) and spatio-temporal data. Each iterative methods presented is subject to the convergence/stopping rules for the hot-spots and variable selection stability. This has a direct consequence in the micro level analyses.

REFERENCES

Hardisty, F. and Klippel, A. 2010. Analysing spatio-temporal autocorrelation with LISTA-Viz. *International Journal of Geographical Information Science*, 24(10), pp. 1515 -1526.

Huang, Y., Shekhar, S., and Xiong, H., 2004. Discovering colocation patterns from spatial data sets: a general approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(12), pp. 1472-1485.

Jung, I., Kulldorff, M., and Richard, O.J., 2010. A spatial scan statistic for multinomial data. *Statistics in Medicine*, 29(18), pp. 1910-1918.

Leibovici, D.G., 2010. Spatio-temporal Multiway Decomposition using Principal Tensor Analysis on k-modes: the R package PTAk. *Journal of Statistical Software*, 34(10), pp. 1-34.

Leibovici, D.G., Bastin, L., and Jackson, M., 2011a. Higher Order Cooccurrences in Point Pattern Analysis and Decision Tree Clustering. *Computers & Geosciences*, 37(3): 382-389.

Leibovici, D.G., Bastin, L., Anand, S., Hobona, G., and Jackson, M., 2011b. Spatially Clustered Associations in Health related geospatial data. *Transactions in GIS*, 15(3): 347-364.

Tango, T., Takahashi, K., and Kohriyama, K., 2011. A Space-Time Scan Statistic for Detecting Emerging Outbreaks. *Biometrics*, 67(1), pp. 106-115.

THE AGENT BASED MODELING FOR HIV TRANSMISSION AND IT'S INTEGRATION WITH GIS

Yang Kun [a, b, *], Wang Jiasheng [a]

[a] School of Tourism and Geographical Science, Yunnan Normal University, 121st Street, Kunming, P.R.China
[b] Engineering Research Center of GIS Technology in West China, Ministry of Education, 121st Street, Kunming, P.R.China,
kmdcynu@163.com

KEY WORDS: Agent Based Model, HIV Transmission, GIS, Model Integration

ABSTRACT:

The HIV transmission model research has undergone three stages of mathematical model, spatial analysis model and the recent Agent based modeling stages. The former two methods belong to static and macro models. However, the agent based model belongs to dynamic and micro model and is very suitable for simulating HIV/AIDS transmission due to that the transmission reason is individual contact with spatial and temporal distribution. Based on the relevant spatial and aids epidemic data of Kunming City of China, this paper focused on the principles and methods for establishing the agent based models of HIV transmission and their integration with GIS.

1. THE SPATIAL AND TEMPORAL INFORMATION MECHANISM OF HIV TRANSMISSION

Within specific geographical space, there are three main approaches for HIV/AIDS spread among people which are sexual contact, injected drug use and mother-to-child transmission routes respectively. The sexual transmission mainly occurs in some commercial occasions such as hotels and leisure places among high risk people groups. The injection drug users usually have special groups within specific geographic space. The above high risk people spread the HIV virus to common people through their definite sex partners of spouses. The HIV/AIDS transmission mechanism can be illustrated as that in Figure 1.

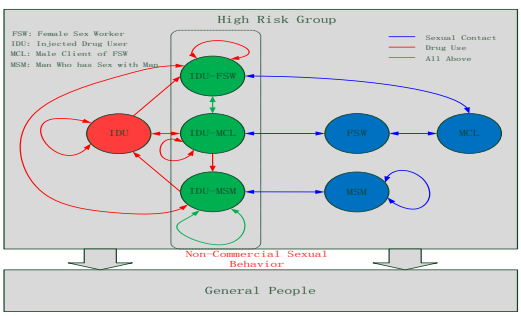


Figure 1 The HIV transmission mechanism among people

2. THE AGENT BASED MODEL FOR HIV TRANSMISSION

For establishing the agent based models of HIV transmission, the relevant agents and environments should be defined first with interactive rules. According to the HIV transmission mechanism and AIDS epidemic data of Kunming City, the injected drug use and sexual contact are the main approaches for HIV transmission which interferes with five class peoples and they will be defined as five class agents. On the other hands, the clinical process of AIDS may be simplified as that in Figure 2 due to that the research focus is HIV carriers and AIDS patients.

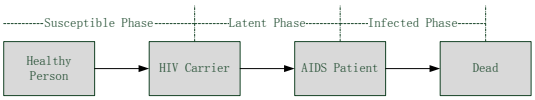


Figure 2 The simplified clinical process of AIDS

According to the characteristics of the high risk people groups, the attributes and behaviors may be extracted from the above five class peoples as that in table 1. Then the five class agents will be defined according to the relevant attributes and behaviors

People Group	Attributes	Behaviors
Female Sex Worker（FSW）	Age、Working Place	Sex Working Frequency、Back to Residential Place、Replacing Work Place
Male Client of FSW（MCL）	Age、Income Level、Sex Trade Interval	Selecting Trade Sites、Selecting Sex Workers、Back to Residential Place
Man who has sex with Man（MSM）	Age、Male Sex Partner	Homo-Sexual Partner Selection and Separation
Injected Drug User（IDU）	Sex、Age、Group	Group Meeting、Use Common Needles、

* Corresponding author:kmdcynu@163.com

		Replacing Groups
General People	Sex、Age、Educated Level、Sex Partners、Residential Place、Health Status	Sex Partner Formation and Separation

Table 1 the attributes and behaviors of the five class people

According to the geographical data sets of Kunming city and the defined five class agents with relevant behaviors and rules, the agent based model for HIV transmission may be established by using the agent based modeling platform Repast J and developing tools in Java. The Model structure may be illustrated as that in Figure 3, which is constructed by threes function layers as User Interface, Model and Data layers.

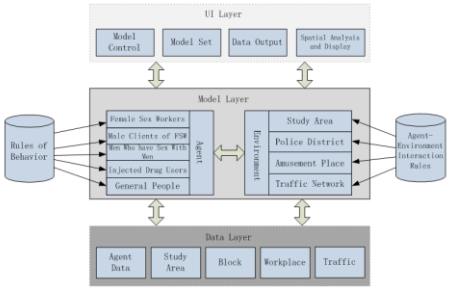


Figure 3 The model structure of agent based models for HIV transmission

3. THE INTEGRATION OF GIS AND AGENT BASED MODELS FOR HIV TRANSMISSION

There are three approaches for integrating GIS with agent based models, which are loose, moderate and tight integration. At present, the data models of GIS mainly include raster and vector data models and the vector data model include point, line and polygon feature data sets. For the integration of vector GIS and agent based models, the relevant data mapping relations should be established between GIS and agent based models, and the principles for integrating vector GIS and agent based models may be illustrated as that in Figure 4.

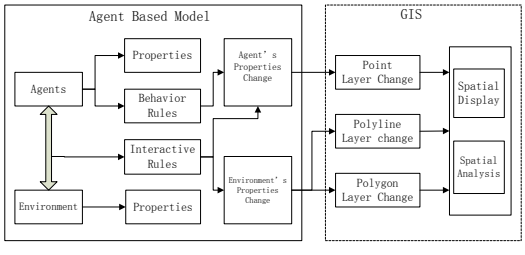


Figure 4 The principles for integrating vector GIS and agent based models

In this research, the loose integration of vector GIS with agent based models has been implemented by using ArcMap and Repast J based on Shapfiles. Then the simulation result may be displayed and spatially analyzed within the ArcMap environment. The integrated platform may be illustrated as that in Figure 5.

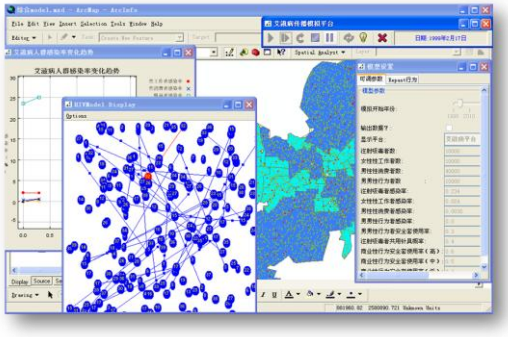


Figure 5 The loose integration of vector GIS with ABM by using ArcMap and Repast J.

The relatively tight integration of vector GIS and agent based models has also been implemented in this research by using ArcGIS Engine 9.2 and Repast J to form the spatial and temporal decision support systems for HIV transmission, which has functions about data operation, query, aids epidemic simulation analysis, statistical analysis, medical resource allocation, AIDS epidemic alarming, AIDS epidemic scenarios and map output. The user interface of the AIDS epidemic simulation and spatial decision support platform may be illustrated as that in Figure 6.

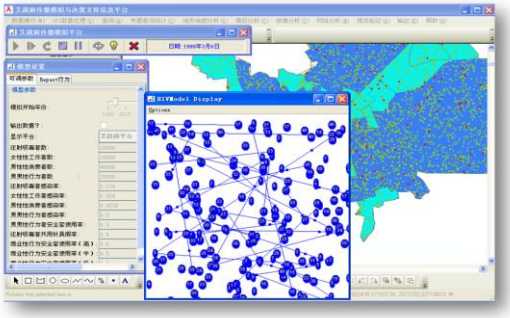


Figure 6 The user interface of HIV epidemic simulation and spatial decision support platform

4. MODEL TESTING

The model testing of Agent Based Model includes simulation program verification and result validation. For the program verification of agent based models for HIV transmission, the software engineering principles and methods have been implemented thoroughly in this research and keep the simulation process in the right status. For the simulating result validation, the linear fitting and Z value testing methods have been implemented in this research by using the AIDS epidemic data of Kunming Center of Disease Control and Prevention from the year 1990 to 2009, which has proven that the simulating results of agent based models for HIV transmission are effective and valid.

5. THE ANALYSIS OF SIMULATING RESULTS

The sensitive analysis about the condom use rate of commercial sexual behavior and needle share rate of injection drug users have proven that HIV infection rate are highly related to these

parameters and the simulating results have been illustrated in Figure 7,Figure 8 and Figure 9.

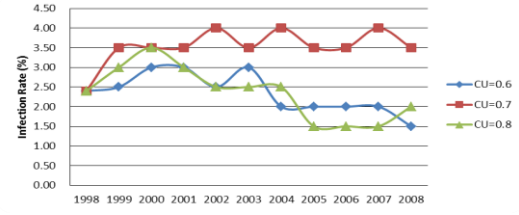


Figure 7 The Effects of Condom Use Rate on Female Sex Worker HIV Infection Rate（CU：Condom Use Rate）

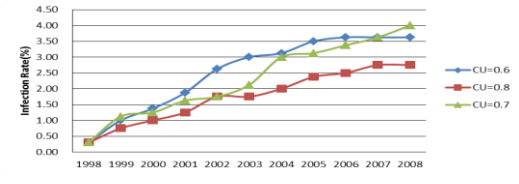


Figure.8 The Effects of Condom Use Rate on Male Sex Consumer HIV Infection Rate（CU：Condom Use Rate）

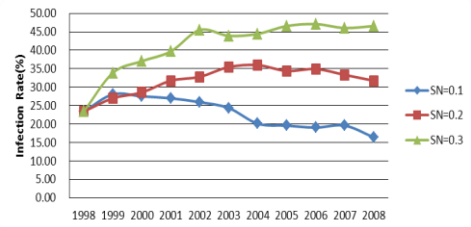


Figure 9 The Effects of Needle Share Rate on Injection Drug User HIV Infection Rate（SN：Needle Share Rate）

6. CONCLUSIONS

Agent Based Modeling is established on the individual behaviors and environments, which may reflect the overall effects of individual activity and is very suitable for simulating HIV transmission due to individual contact. The agent based models for HIV transmission and their integration with vector GIS to form the AIDS epidemic simulation and spatial decision support platform have been applied in Kunming CDC and the simulating results are relatively right compared with the baseline investigation data in Kunming City.

REFERENCES

Lu L, Jia M, Ma Y, Yang L, Chen Z, Ho DD, et al 2008 . The changing face of HIV in China. Nature 2008, 455, pp.609-611.

Jia M, Luo H, Ma Y, Wang N, Smith K, Mei J, Lu R, Lu J, Fu L, Zhang Q, Wu Z, Lu L, 2010. The HIV Epidemic in Yunnan Province, China, 1989-2007. Journal of Acquired Immune Deficiency Syndromes, 53, pp.34-40.

Yang K, Li J, Cui Q & Cheng X, 2008. The Integrated Models of ABM and GIS for HIV /AIDS Transmission, Journal of Yunnan Normal University, 40(4), pp.14-20.

GeanuracosC, CunninghamS, WeissG, ForteD, ReidL, & EllenJ, 2007. Use of Geographic Information Systems for Planning HIV Prevention Interventions for High-Risk Youths. American Journal of Public Health, 97(11), pp.1974-1981.

Namulanda, Gonza N, 2005. HIV/AIDS, GIS and the Internet: Efficient health care planning and equitable resource allocation for HIV/AIDS health and social services. M.S. dissertation, University of Missouri - Columbia, United States -- Missouri.

Brown D, Riolo R, Robinson D, North M, Rand W, 2005. Spatial process and data models: Toward integration of agent-based models and GIS. Journal of Geographical Systems [serial online]. 7(1), pp.25-47.

Groff E, 2007. 'Situating' Simulation to Model Human Spatio-Temporal Interactions: An Example Using Crime Events. Transactions in GIS [serial online]. 11(4), pp.507-530.

Simoes J, 2005. Spatial Epidemic Modelling in Social Networks. AIP Conference Proceedings [serial online]. 776(1), pp.287-297.

Jackson J, Forest B, Sengupta R, 2008. Agent-Based Simulation of Urban Residential Dynamics and Land Rent Change in a Gentrifying Area of Boston. Transactions in GIS [serial online]. 12(4), pp.475-491.

Wang J, Xiong J, Yang K, Peng S, Xu Q, 2010. Use of GIS and agent-based modeling to simulate the spread of influenza. Geoinformatics 2010, pp.1-6.

Xiong J, Wang J, Yang K, Peng S, Xu Q, 2010. Multiagent-based simulation of the HIV/AIDS spatial and temporal transmission among injection drug users. Geoinformatics 2010, pp.1-6.

Peng S, Yang K,Xu Q, Wang J, Xiong J, Liu L,2010. A simulation study of H1N1 space-time epidemic based on agent-based modeling. Geoinformatics 2010, pp.1-4.

Carpenter C, 2004, Agent-based modeling of seasonal population movement and the spread of the 1918-1919 flu, University of Missouri-Columbia, Missouri.

Rhee, AJ, 2006. An agent-based approach to HIV/AIDS Epidemic modeling: a case study of Papua New Guinea epidemic, Massachusetts Institute of Technology, United States -- Massachusetts.

Deng H, Chi Y, Tang Y, 2004. Multiagent-based simulation of disease infection, Computer Simulation, 21, pp.167-175.

Gong J, Sun Z& Li X, 2003. Simulation and analysis of control of severe acute respiratory syndrome,Journal of Remote Sensing, 4, pp.260-265.

Perez L& Dragicevic S, 2009. An agent-based approach for modeling dynamics of contagious disease spread, International Journal of Health Geographics, 2009. http://www.ij-healthgeographics.com/content/8/1/50.

ACKNOWLEDGEMENTS

This research is financially supported by the National 863 Project(2007AA12Z231).

Managing Dutch elm disease: an agent-based model approach

Bruce Mitchell [a], Joana Barros [b]

Department of Geography, Environment and Development Studies,
Birkbeck, University of London,
Malet Street, London, WC1E 7HX

Telephone: (+440) 207 079 0644
Fax: (+44) 207 6316 498
Email [a]: bruce.birkbeck@ntlworld.com
[b]: j.barros@bbk.ac.uk

Keywords: Dutch elm disease control, agent-based modelling, local vulnerability

1. Introduction

Dutch elm disease (DED) is the most destructive tree disease known, transported around the globe by international trade. DED has been studied since 1918, and its epidemiology and lifecycle are by now well known. A number of models of DED have been developed, focusing on either the biological aspects of the disease (i.e. Castro and Bolke, 2004) or on the spread of the disease (i.e. Swinton and Gilligan, 1996). Thus, DED has been studied mostly as an aspatial phenomenon despite it clearly being a geographical phenomenon in which the patterns and locations of different elements are of fundamental importance. In particular, topological features above a certain height are believed to restrict the spread of the disease, making elevation a key spatial factor in DED's epidemiology.

The paper presents an agent-based model (ABM) of Dutch Elm Disease (DED), applied to the Isle of Man (IoM). The objective of this paper is to determine how the accurate targeting of resources towards at-risk areas might affect the results of a DED control campaign.

2. The DED-IoM Model

The DED-IoM model was initially developed as an exploratory spatial model of DED for the IoM, focusing on the identification of areas of enhanced vulnerability to the disease within the IoM.

For the second stage of the model development, the emphasis is on redeveloping the model in order to use it as an analytical tool for resource management. This is done by

using the simulation as a model for determining the impact of accurate targeting resources towards at-risk areas might affect the results of a DED control campaign.

The model uses the 2.5-D agent-based modelling environment StarLogo TNG (MIT Media Laboratory, 2008-10) to represent the presence of DED upon the Isle of Man. Despite being a relatively simple user-friendly framework, TNG has proved to be well up to the task.

2.5D is a graphical projection technique where two-dimensional images or scenes are so presented as to stand in for a three-dimensional reality. The DED-IoM model represents the topographical space of the Isle over which a set of agents interact. The third dimension of height was added to the model by the importation into StarLogo, as a raster, of a 75m DEM of the Isle of Man from EDINA (generalised to 400 m2 cells). This environment was populated with four agents (elm trees, beetles, foresters and birds), and the model played out over the surface of the raster.

2.1 Agent's behaviour

Agents (tree, forester and bird) were generated at the start of the simulation and distributed randomly on the virtual landscape. The model ran as described in the flowchart. Beetle agents issued from infected elms (<=1.5% of the 800) and sought out nearby healthy Elm in which to lay eggs and propagate the infection. Foresters sought to fell diseased Elm before the next generation of Beetle hatches. Boids followed and ate Beetles. These behaviours were modified by a 'sense' of where the respective prey was.

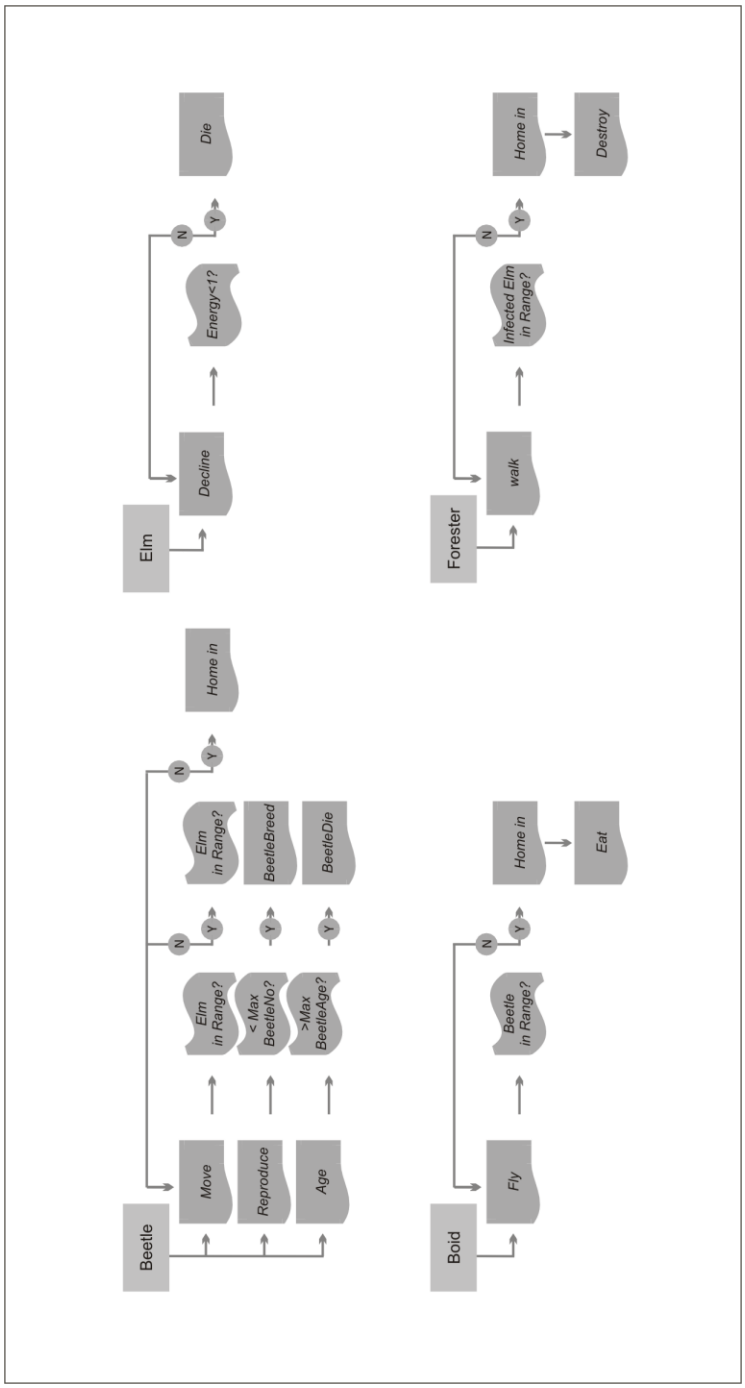


Figure 1: Agent Flowchart, Version 3

2.2 Model Mechanics

The model paused at game turn (GT) 60 to record the number of beetles and where they were located (they always operate in a single cluster for each run). It then ran on for an arbitrary 1,000 game turns, recording and exporting results as it went, then resets and ran again repeatedly, producing a distribution of data. This distribution was combined with a terrain analysis and ported into ArcGIS Spatial Analyst, Mapinfo and SPSS for analysis (Mitchell et al., 2009, 2010).

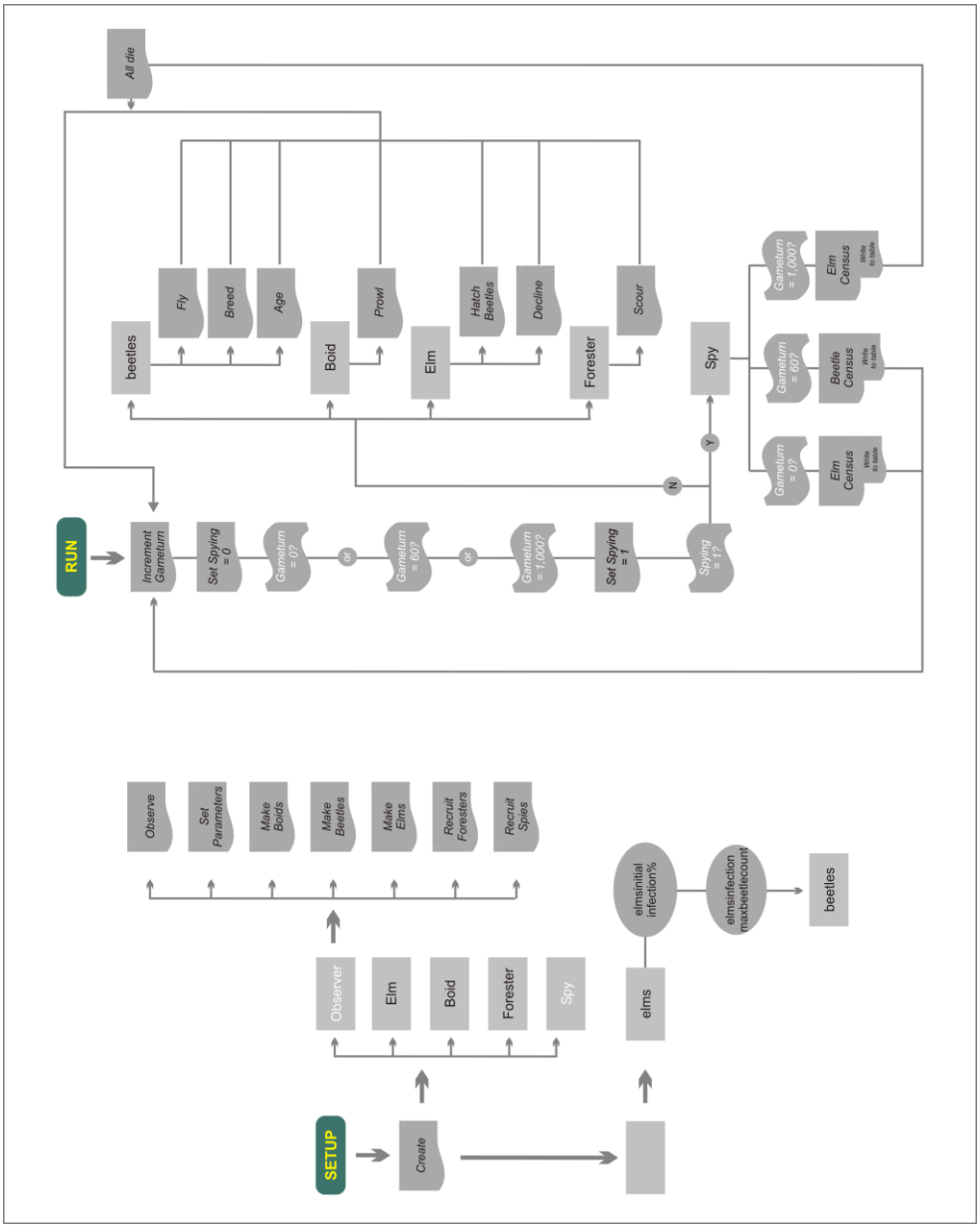


Figure 2: System Flowchart, Version 3

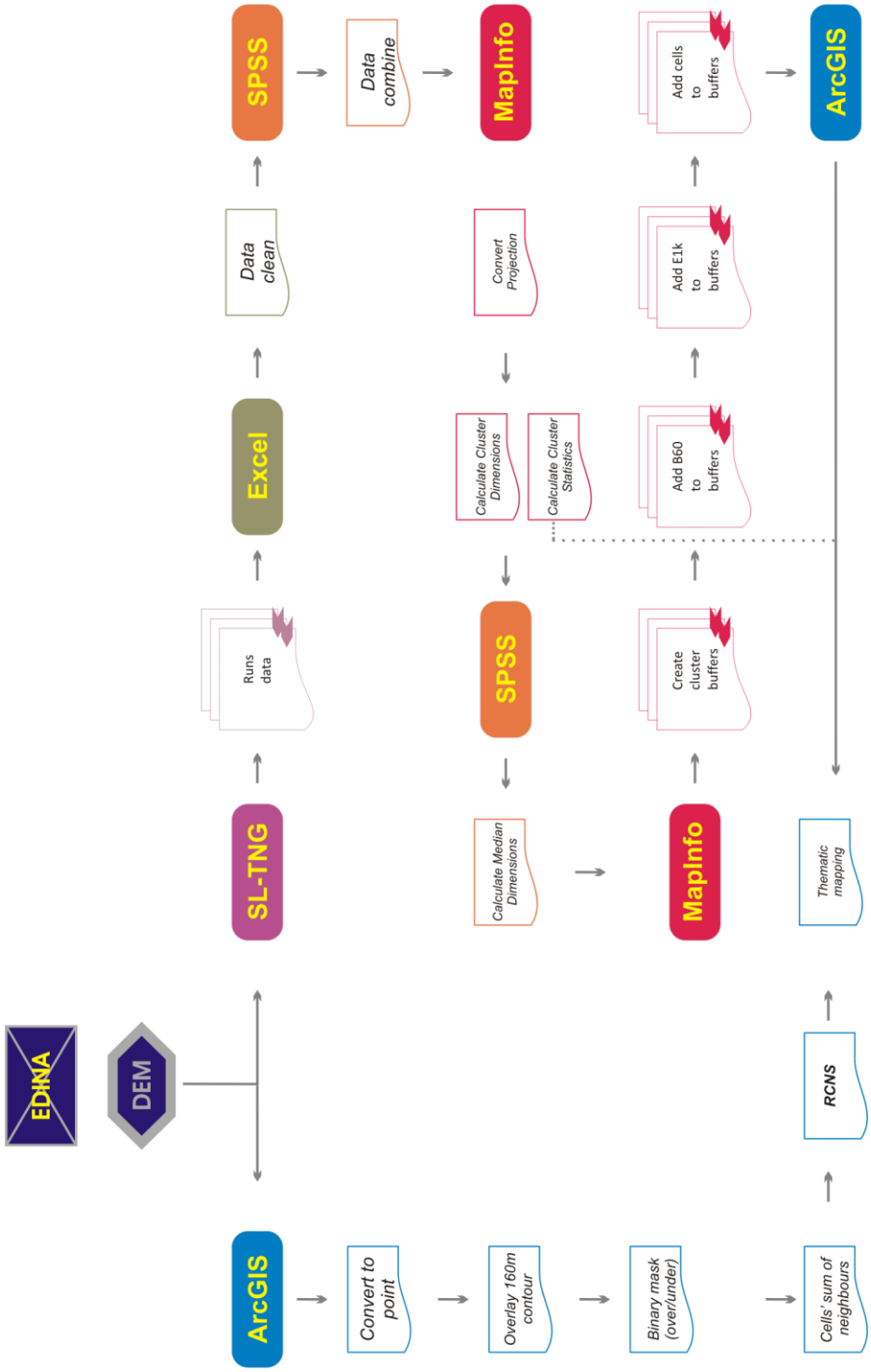


Figure 3: Data Processing Flowchart, Version 3

3. The DED-IoM Model as a resource management tool

The current version of the model investigates how programmed variation in the behaviour of the human Agent Forester might bring about different outcomes. In the original model, foresters were allowed to range freely across the island in search of diseased elm trees. This meant that they could cross the hilly, elm-free districts. But their progress while in the highlands was random and aimless until they came within range of diseased trees.

In this version, the movements of Agents Forester are restricted so that they remain more, or less, close to the elm population. The real-world Scolytus beetles that spread the disease do not fly much beyond the maximum elevation locally colonised by elms (the 'elm-line'). Landscape above that elm line is not penetrated by the beetles. Manx data support an elm line of 160 metres – and on the Isle, two ranges of hills exceed that height and restrict the flow of beetles and disease across the island. The 'beetle-line' in the model was set at an arbitrary 25% higher than the elm-line (200m).

The model investigates two limits on forester movement: a) a higher one (240m) which grants them greater freedom of movement than have the beetles (permitting them to cross higher ridges) and b) a lower one (160m) which constrains them to within the elm line. There is a trade-off between time spent passing over treeless ridges and time spent among elms but hemmed within valleys while circumnavigating spurs of higher land.

Two distributions of data (of 500 runs per scenario) were carried over into, compared and analysed in ArcGIS Spatial Analyst, Mapinfo and SPSS (Mitchell et al., 2009, 2010).

4. Discussion

The present study investigates the use of a exploratory agent-based model as a spatial analytical tool for resource management in the process of containment of the DED. Preliminary results suggest that considering the limitations of an agent-based approach and the difficulties in validating the DED-IoM model, the model's results are informative to the process of resource management in the Isle of Man. However, more sensitivity tests, coupled with calibration of the model are required in order to determine the accuracy of such results and the utility the model might provide for resource management in areas of DED risk.

5. References

Bar**ros, J. & Alves Jr., S. (2003),** Simulating Rapid Urbanisation in Latin American Cities, *in* Longley, P. and Batty, M., eds, *Advanced Spatial Analysis: The CASA Book of GIS*, ESRI Press, London.

EDINA Digimap. Available online at: http://digimap.edina.ac.uk/main/index.jsp

Jones, P. 1981. The Geography of Dutch elm disease in Britain. 1981. The Royal Geographical Society (with the Institute of British Geographers).

Kammeraad,, J.W., Brewer, R. 1963. Dispersal Rate and Elm Density as Factors in the Occurrence of Dutch Elm Disease. American Midland Naturalist, Vol. 70, No. 1 (Jul., 1963), pp. 159-163.

MIT Media Laboratory (2008-10), StarLogo TNG 1.5 (Software). Available online at http://education.mit.edu/starlogo/

Mitchell, B. Markovic D., Rotheray D. and Measho, S. (2009). Modelling the Epidemiology of Dutch Elm Disease. A case study of the Isle of Man. Paper submitted at GeoComputation 2009, Sydney, Australia.

Mitchell, B. Barros, J. Wendel, D. (2010). Identifying Dutch elm disease 'danger-spots' on the Isle of Man with an agent-based model. Paper submitted at GISRUK Conference, 2010, UCL, London, UK

Rose, D. 2000. Dutch Elm Disease in the Isle of Man. A review of current practices and recommendations for the future. Forest Research, Disease Diagnostic and Advisory Service

Swinton, J. and Gilligan, C.A. 1996. Dutch Elm Disease and the Future of the Elm in the UK: A Quantitative Analysis. Philosophical Transactions, Biological Sciences. Vol. 351, No. 1340. The Royal Society.

Wooldridge, M. J. (2002), *An Introduction to Multiagent Systems,* J. Wiley, Chichester.

DISCOVERING SPATIO-TEMPORAL PATTERNS OF ELECTORAL SUPPORT WITH CONTRAST DATA MINING

T. F. Stepinski [a, *], J. Salazar [b]

[a] Department of Geography, University of Cincinnati, Cincinnati, OH 45221-0131, USA - stepintz@uc.edu
[b] Lunar and Planetary Institute, Houston, TX 77058, USA - salazar@lpi.usra.edu

Commission II, WG3

KEY WORDS: Data Mining, Temporal, Generalization, Sociology, Surveying

ABSTRACT:
A novel strategy for exploration of complex spatio-temporal datasets is introduced. This model-free, brute computational power-based strategy relies on the concept of contrast data mining. Given an input consisting of predictor variables and a two-class response variable, the method yields an exhaustive set of predictors' patterns associated with each of the two classes. Pattern summarization is used to reduce this set of patterns to an interpretable insight. This strategy is demonstrated on a dataset combining US census data with the results of Presidential election in 2004 and 2008. Socio-economic blocks of voters' support for each of the two main parties are identified together with their geographical distributions and temporal stabilities.

1. INTRODUCTION

The complexity of spatio-temporal datasets consisting of response and predictors variables hides knowledge that may be discovered by exploring their overall structure. A common approach to such discovery is to build a regression model and to infer the structure of the dataset from the form of the model (White and Sifneos, 2002; Demsar et al.,2008). However, regression models are not the best tools for knowledge discovery. Recently, we have proposed (Stepinski et al., 2010a; Stepinski et al., 2010b) a data mining-based method for exploration of spatial response/predictors datasets. Our method relies on the concept of contrast data mining and uses no models.

The method described here is for datasets with binary response variable; further extension, to datasets with multivalued response variable is left to future research. The novel contribution of this work is an extension of our original method from purely spatial to spatio-temporal analysis. The working of the method is illustrated by applying it to a dataset consisting of socio-economic census data (predictor variables) for all the counties in the contiguous United States and the results of Presidential elections in 2004 and 2008. The inclusion of two distinct election years in the dataset requires modification of the original method, so it can handle spatio-temporal datasets.

2. METHOD

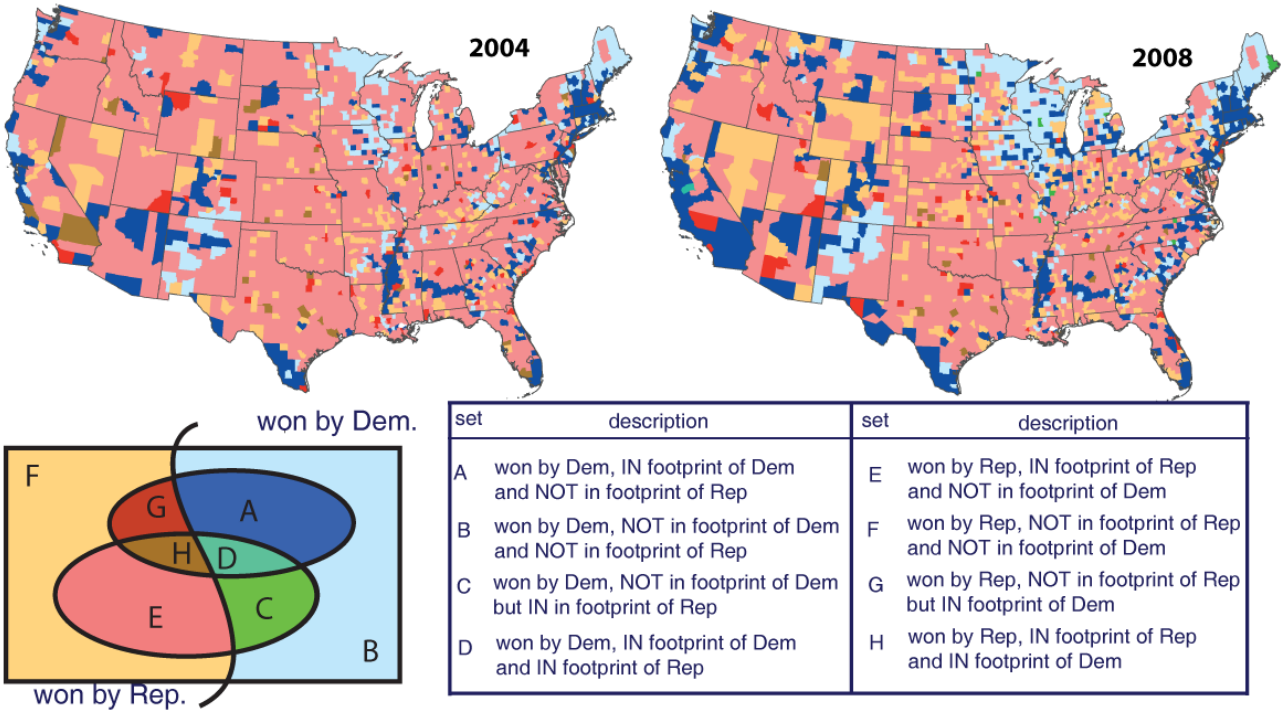
A dataset is organized around spatio-temporal objects (o) (counties at different years) and structured as follows $o_i = \{id; f_1, f_2, \ldots, f_m; c\}$, $i=1, \ldots, N$; where id is county's label indicating its number *and* a year of election, $\{f_1, f_2, \ldots, f_m\}$ are the ordinal values of m socio-economic indicators, c (either 1 if a county was won by the democrats or 0 if it was won by the republicans) indicates an outcome of the election for a given object, and N is the number of objects (counties) in the dataset. The inclusion of an election year in the id is the *only* change required to modify our original method for application to a spatio-temporal domain.

From the point of view of association analysis, such dataset is a set of N fixed-length transactions $\{f_1, f_2, \ldots, f_m\}$. Our method mines for itemsets that discriminate between the two possible outcomes of a vote. An itemset is a set of values of socio-economic indicators common to a group of transactions (their common factor). A discriminative itemset is an itemset that is frequent among $c=1$ transactions but rare among $c=0$ transactions or vice versa. A collection of all discriminative itemsets provides a complete description of all existing socio-economic blocks of voters preferring one party over the other. A geographical extent of an itemset is called its footprint. It is common for this complete enumeration-based method to find thousands, sometimes barely distinct, discriminative itemsets. We employ a custom-designed similarity measure (Stepinski et al., 2010b) to cluster them into aggregates large enough to provide interpretable insight.

3. RESULTS

The dataset consists of 3107 counties listing 13 socio-economic indicators selected from the US Census Bureau data and the results of Presidential election in 2004 and 2008. Contrast data mining is performed separately for Democrats (Dem.) vs. Republicans (Rep.) and vice versa. A given county may belong to a footprint of Dem., to a footprint of Rep., to both, or to neither. Taking into account that a county was actually won by either Dem. or Rep., there are eight possible outcomes as illustrated on Fig.1. The sets of counties indicated as A and E on Fig.1 represent a core support for each party, sets B and F represent an additional, non-core support for each party. Overall, the patterns of support look similar in 2004 and 2008 indicating a relative stability of voter preferences. However, the differences observed on Fig.1 are important as they are responsible for a different outcome of the two elections.

* Corresponding author.



set	description	set	description
A	won by Dem, IN footprint of Dem and NOT in footprint of Rep	E	won by Rep, IN footprint of Rep and NOT in footprint of Dem
B	won by Dem, NOT in footprint of Dem and NOT in footprint of Rep	F	won by Rep, NOT in footprint of Rep and NOT in footprint of Dem
C	won by Dem, NOT in footprint of Dem but IN in footprint of Rep	G	won by Rep, NOT in footprint of Rep but IN footprint of Dem
D	won by Dem, IN footprint of Dem and IN footprint of Rep	H	won by Rep, IN footprint of Rep and IN footprint of Dem

Figure 1. Geographic distribution of US counties divided into eight categories based on itemset footprints and results of Presidential elections.

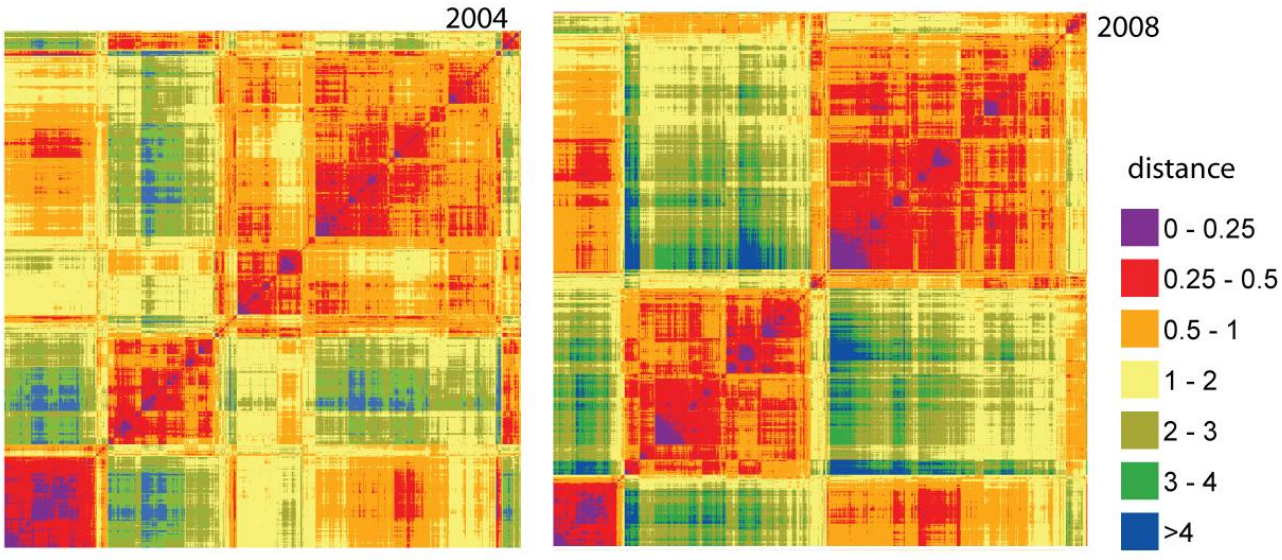


Figure 2. Heat maps illustrating proximity relations between 6746 discriminative itemsets mined from the 2004 dataset and 6970 discriminative itemsets mined from the 2008 dataset. Heat maps shown here are symmetric.

Visualization of proximity relationships between all discriminative itemsets are given by the so-called heat map; heat maps for itemsets mined in 2004 and 2008 datasets are showed in Fig.2. A heat map is a distance matrix organized by the order of dendrogram. The heat maps shown in Fig.2 visualize similarities between all itemsets from groups A, B, E and F (arranged consecutively from left to right and from bottom to top in the distance matrix); purple and red colours indicate similarity, green and blue colours indicate dissimilarity. The two heat maps shown in Fig.2 have different sizes because the numbers of itemsets mined in 2004 and 2008 are different. We use heat maps to summarize information contained in a large number of itemsets to a few interpretable clusters.

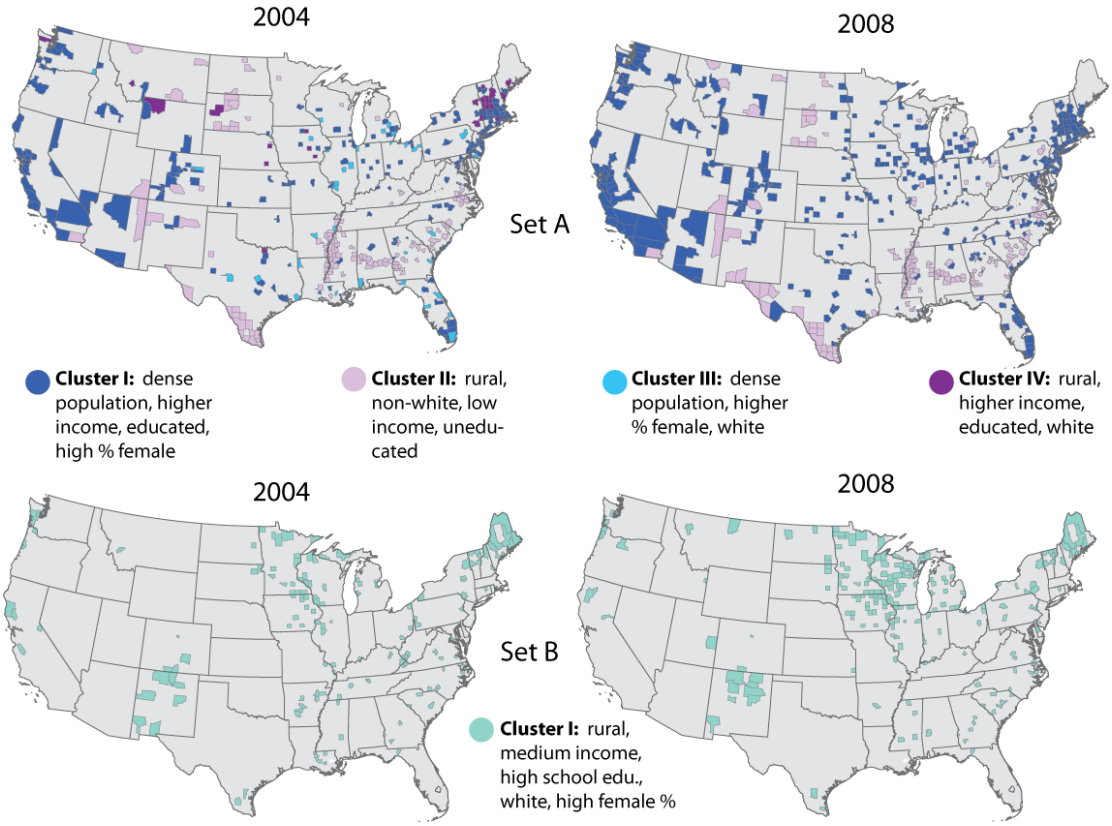


Figure 3. Geographic distribution of socio-economic blocks of counties voting for democrats.

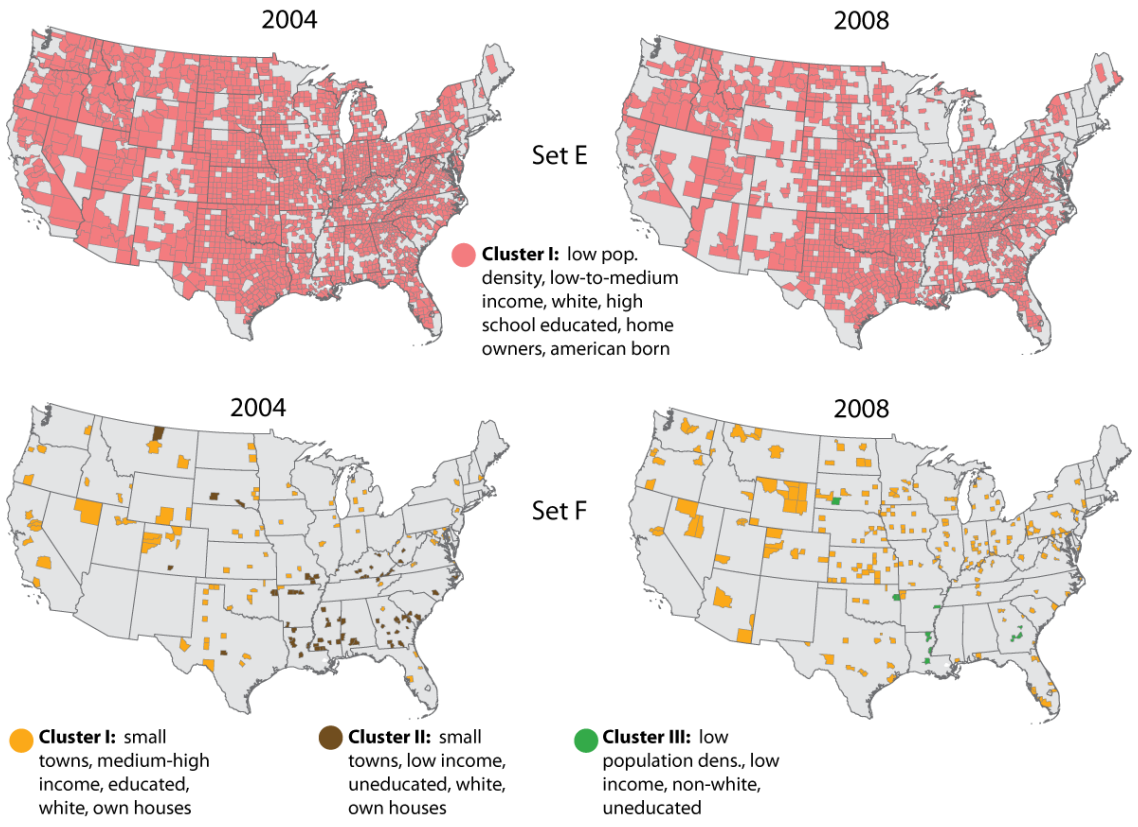


Figure 4. Geographic distribution of socio-economic blocks of counties voting for republicans.

The clusters of similar itemsets are identified as square blocks of red and purple colours located along the diagonal. The advantage of a heat map over other visualization techniques is that it also shows imperfections of agglomerative clustering. These imperfections appear as off-diagonal rectangles of red colours; they indicate existence on additional similarities that may not be picked up by a dendrogram.

By analyzing the two heat maps we offer the following interpretation of our dataset. In 2004 a core support for Dem. (A) can be summarized by four clusters. These are the first four red squares along the diagonal of 2004 map starting from the lower-left corner, two of which are larger and easily seen, the other two much smaller. A non-core democratic support can be summarized by a single cluster (fifth red square counting from the lower-left on 2004 map). In 2008 the core democratic support consolidates to just two clusters (the first two red squares on 2008 map counting from the lower-left) while the non-core support remains concentrated in a single cluster. For Rep. the core support remains steady in 2004-2008 and is concentrated in a single cluster (the second red square on 2004 map counting and the third red square on 2008 map, counting from the top-right corner). Republicans non-core support is a single cluster in 2004 (the first red square on 2008 map counting from the top-right corner) but splits into two clusters (the first two red squares on 2008 map counting from the top-right corner), although only one of them represents continuity of itemsets.

4. CONCLUSIONS

Figs. 3 and 4 show geographic distribution of voting blocks for democrats and republicans, respectively. Temporal change from 2004 to 2008 is inferred by comparing left and right columns of these figures.

Core democratic and republican blocks of support change little during this single election cycle. The different outcome of the Presidential election can be traced to changes in non-core blocks of support. Block B has been present in 2004 but has significantly expanded in 2008. This block consists of counties that have Republican-like profiles but have voted Democratic anyway. Such counties are located mostly in the Midwest and New England. They voted Democratic for reasons that are outside of the socio-economic factors present in our dataset. In addition, Republicans lost the block F-III, which was not compensated by the slight expansion of the block F-I.

REFERENCES

Demsar, U. et al., 2008. Combining geovisual analytics with spatial statistics: the example of geographically weighted regression. *The Cartographic Journal*, 45(3): pp. 182-192.

Stepinski, T.F. et al., 2010a. Discovering spatio-social motifs of electoral support using discriminative pattern mining. In Proceedings of *COM.Geo '10 1st International Conference on Computing for Geospatial Research and Application*, article #39.

Stepinski, T.F. et al., W. 2010b. ESTATE: Strategy for Exploring Labeled Spatial Datasets Using Association Analysis. *Lecture Notes in Computer Science*, 2010, Volume 6332/2010, pp. 326-340.

White D. and Sifneos J.C., 2002, Regression tree cartography. *J. Computational and Graphical Statistics*, 11(3): pp. 600-614.

EMBEDDING AND RETRIEVAL OF WEATHER RADAR SEQUENCES: A DATA MINING APPROACH TO PRECIPITATION NOWCASTING

L. Foresti*, M. Kanevski

Institute of Geomatics and Analysis or Risk, University of Lausanne, 1015 Lausanne, Switzerland - (loris.foresti, mikhail.kanevski)@unil.ch

KEY WORDS: Sequence retrieval, trajectories in eigen-spaces, weather radar, orographic precipitation, short-term forecasting

ABSTRACT:

The goal of this research is to explore and to exploit data archives for short-term forecasting (nowcasting). A large set of weather radar images covering the southern side of the Alps and defined as orographic precipitation events is used for the analysis. Images are first embedded using principal component analysis to construct a space describing the temporal evolution of precipitation. The evidence of similar trajectories (image sequences in the original space) motivates the development of nowcasting models by the method of analogues. Forecasts are generated by retrieving similar radar sequences from the archive and by looking to their subsequent evolution. Preliminary results show that the predictability of the system at a particular lead time can be controlled by retrieving an appropriate number and length of sequences.

1. INTRODUCTION

Sequential data, e.g. data presenting a spatial and/or a temporal ordering, can be found in many applications such as video analysis, gene expression and weather forecasting. Specifically, the trajectories describing the evolution of continuous temporal data streams can be uncovered by finding an appropriate embedding (phase space). Predictability comes from the presence of similar trajectories, also known as naturally occurring analogues in meteorology (Lorenz, 1969; Obled *et al.* 2002; Wilks *et al.* 2009) or collocation episodes in data mining (Cao *et al.* 2006), which can be used to describe the chaotic dynamics of the system. An optimal data embedding, retrieval and combination of trajectories is the basis for data-driven forecasting. Temporal data mining approaches to forecasting are particularly appealing nowadays because of the growing amounts of data which need to be processed efficiently.

Based on a large set of weather radar sequences, we consider the optimal retrieval of trajectories for short-term precipitation forecasting (nowcasting). This study follows the developments of Otsuka *et al.* (2000) and Panziera *et al.* (Submitted in 2010) which studied the precipitation nowcasting problem by sequence retrieval and method of analogues respectively. The issue of selecting an adapted number and length of sequences for forecasting a particular lead time is analyzed and will guide future developments.

2. DATA PREPARATION

Data are obtained from a C-band Doppler radar located at the top of Monte Lema in the southern side of Switzerland. Rainfall intensities with 5 minutes of temporal and 1x1km^2 of spatial resolutions are obtained after pre-processing (Germann *et al.*, 2006). Orographic precipitation events (persistent and intense rainfall on the windward side of the Alps, see Figure 1) are selected using criteria of precipitation duration and spatial extension (Panziera *et al.*, 2010). The dataset is a matrix of 47000 images (instances) and 1614 pixels (dimensions).

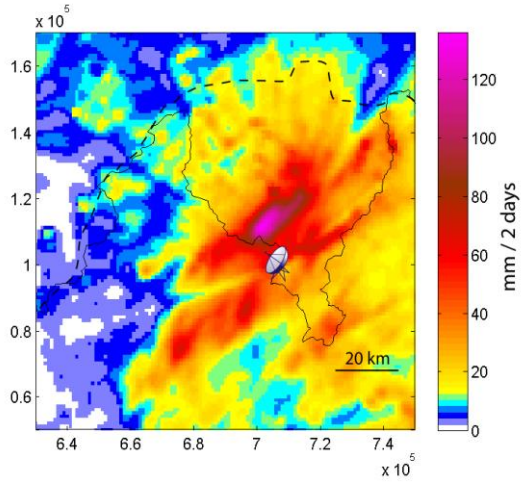


Figure 1. Precipitation totals of 8-9 of October 2009. Antenna sign: radar; continuous line: Swiss border; dashed line: main Alpine ridge.

3. METHODOLOGY AND EXPERIMENTS

Sequence-based nowcasting is performed in the following way:

1. Reduce data dimensionality using PCA. Data trajectories in the eigen-space are shown in Figure 2.
2. Choose an image to nowcast (Figure 3)
3. Retrieve *k* similar (*k*- nearest neighbours, *k*-NN) and temporally independent trajectories from the archive (Figure 3).
4. Forecast is given by averaging the subsequent precipitation fields following the *k* retrieved sequences (Figure 4)

* Corresponding author.

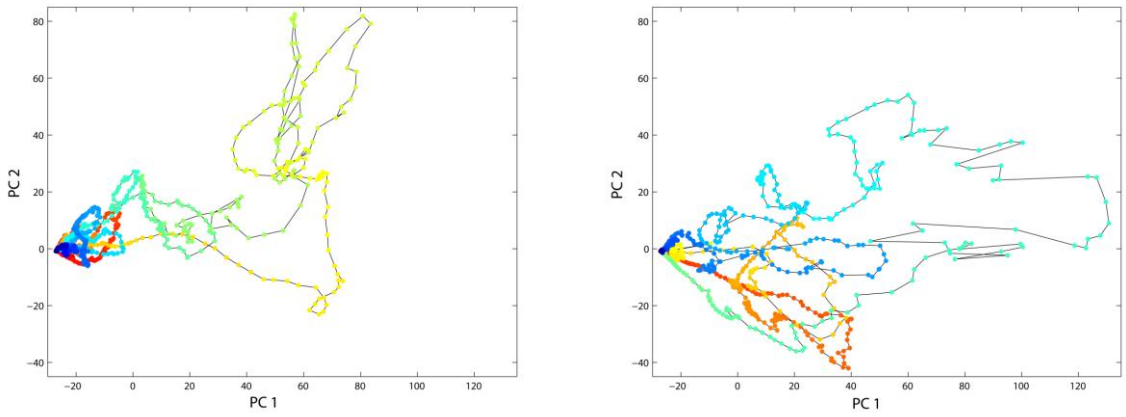


Figure 2. Trajectories of radar images in first two PCA components for two orographic precipitation events in the archive. Left: 14-17 of April 2005; right: 15-18 of May 2005. Colours depict the temporal ordering.

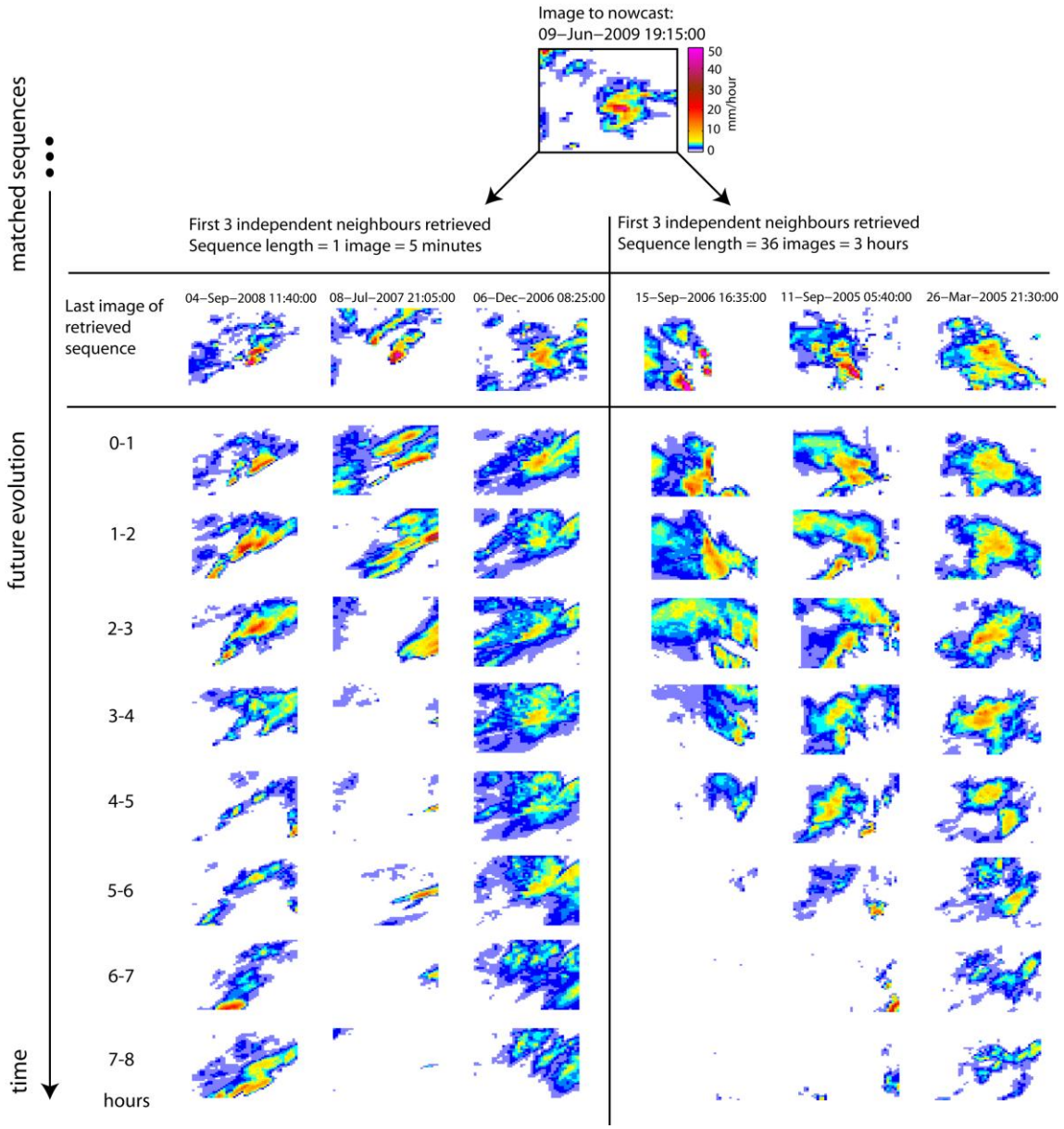


Figure 3. Conceptual illustration of the nowcasting system. Because of the chaotic nature of the atmosphere, the evolution of precipitation patterns in the 8 hours following the retrieved sequences can be quite different. Thus, hourly averages are more robust forecasts. Sequence retrieval (k=36, right) is expected to capture the dynamics of the system yielding to better forecasts compared to simple image retrieval (k=1,left).

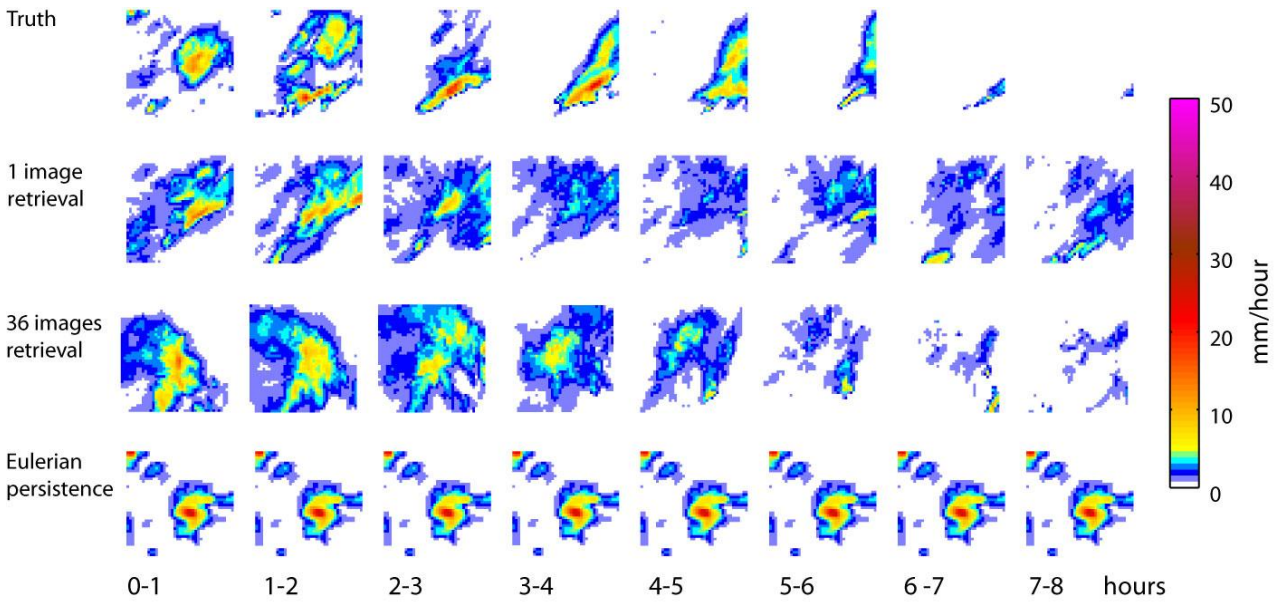


Figure 4. Nowcasting example of hourly rainfall accumulations with 0,...,7 hourly lead times.
1st line: true rainfall rates;
2nd line: forecast (average) given by retrieving 3 sequences of length 1;
3rd line: forecast (average) given by retrieving 3 sequences of length 36;
4th line: forecast given by Eulerian persistence (reference method).

5. Compute a set of skill scores for each lead time (forecast verification)
6. Repeat steps 2 to 5 to compute average skill scores on a representative set of images of orographic precipitation events (Figure 5)
7. Repeat steps 1 to 6 to select best parameters according to the error committed at a particular forecast lead time (Figure 6)

Flexibility of the nowcasting system comes from the choice of the dimensionality reduction method and of the embedding dimension, the length and number of retrieved sequences (neighbours), the similarity measure used for sequence retrieval, the forecasting method (retrieval or time series prediction), the spatial extension and scaling properties of the images to retrieve.

Experiments were performed by varying the number and length of retrieved sequences. Average mean absolute errors (MAE) are computed for each lead time by nowcasting a set of 3104 images. Sequence lengths were set to [1,6,12,24,36,48,60] images (corresponding to 5, 30 minutes and 1, 2, 3, 4, 5 hours). Number of neighbours tried were [1 6 12 21 33]. All other parameters were fixed: PCA with 10 factors (only explains main image patterns without considering small perishable features, e.g. ~45% of the variance), Euclidian distance as similarity measure, 8 hours of temporal independence between neighbours, 24 hours of temporal independence between image to nowcast and neighbours, average between the *k* images (hourly sums) following the matched sequences as forecast.

4. RESULTS

Figure 4 shows an example of forecasts using different settings. Sequence-based forecasts allow considering the dynamics of precipitation compared to Eulerian persistence forecasts which simply keep the most recent image frozen. The lower average MAE in Figure 5 confirm this statement already after 1 h lead time.

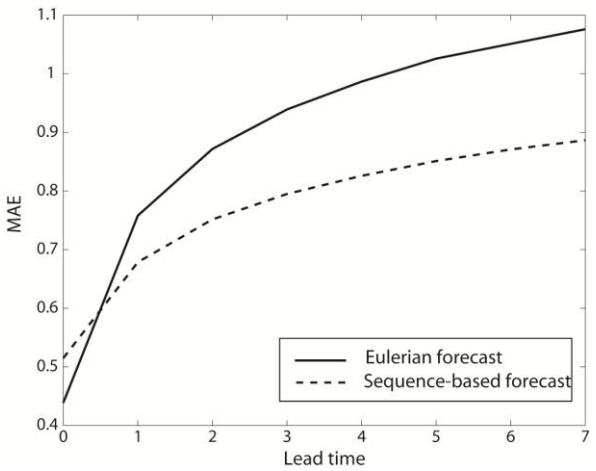


Figure 5. Averaged MAE as a function of lead time for Eulerian and sequence-based nowcasts.
Parameters are optimized for each lead time.

Figure 6 plots average MAE surfaces as a function of the number of neighbours and sequence lengths. The sequence length and neighbours' number needed to minimize forecasting errors increase with increasing lead times. The presence of a minimum in the error surface (Kanevski *et al.*, 2009) for the first two lead times suggests that there is predictability in the system. This lead-time dependence of model parameters can be used to construct nowcasting systems with adaptive retrieval and embedding strategies. However, the selection of parameters highly depends on the skill score used for forecast verification. It is not yet clear if the use of different skill scores still provide a coherent parameter selection according to variable lead times.

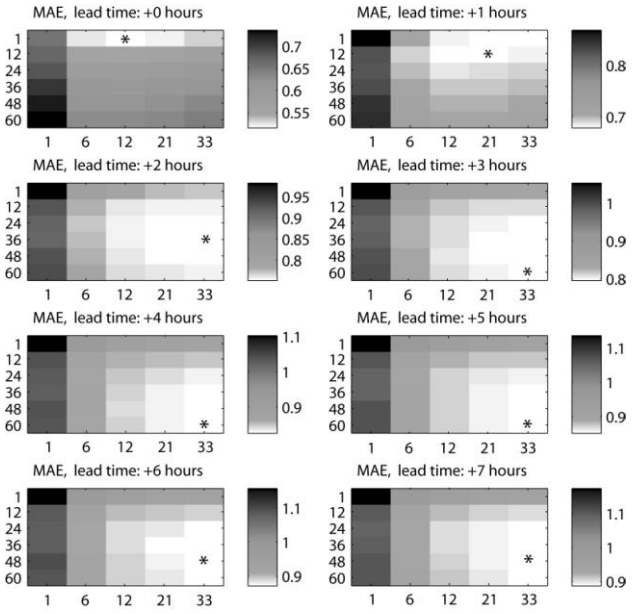


Figure 6. Forecast MAE for each lead time as a function of number of neighbours (x-axis) and of sequence length (y-axis). Asterisk locates the minimum error.

5. CONCLUSIONS

Image sequence retrieval is presented as a method for nowcasting of orographic precipitation from an extensive archive of radar images. Following the ideas of Otsuka *et al.* (2000), radar images are embedded using PCA and the observation of similar trajectories is used for nowcasting purposes. Predictability of precipitation for increasing lead times is assessed by varying the number and length of retrieved sequences.

The most important developments concern the integration of additional variables (wind) as in Panziera *et al.* (Submitted in 2010) and the use of boundary conditions for increasing the predictability of precipitation. Additionally, spatial scale-dependence of precipitation should be accounted to avoid the prediction of unnecessary details and perishable features. Different similarity measures, scaling of variables, retrieval rules (*k*-NN vs. ε-ball) can be tested to optimize the forecasts. However, the most significant improvement in precipitation nowcasting is expected to come from the correct embedding (predictors), the multi-scale characterization of precipitation (Seed, 2003) and the quality of analogues found.

REFERENCES

Cao, H., Mamoulis, N. Cheung, D.W. 2006. Discovery of collocation episodes in spatiotemporal data. *6th IEEE International Conference on Data Mining*, pp. 823-827.

Germann, U., Galli, G., Boscacci, M., Bolliger, M., 2006. Radar precipitation measurement in a mountainous region. *Quarterly Journal of the Royal Meteorological Society*, 132, pp. 1669-1692.

Kanevski, M., Pozdnoukhov, A., Timonin, V., 2009. *Machine Learning for Spatial Environmental Data*. EPFL Press.

Lorenz, E.N., 1969. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric Sciences*, 26, 636-646.

Obled, C., Bontron, G., Garçon, R. 2002. Quantitative precipitation forecasts: a statistical adaptation of model outputs through an analogues sorting approach. *Atmospheric Research*, 63(3-4), pp. 303-324.

Otsuka, K., Horikoshi, T., Kojima, H., & Suzuki, S., 2000. Image sequence retrieval for forecasting weather radar echo pattern. IECE *Transactions Information and Systems.*, E83-D, pp. 1458-1465.

Panziera, L., Germann, U., 2010. The relationship between airflow and orographic precipitation on the southern side of the Alps as revealed by weather radar. *Quarterly Journal of the Royal Meteorological Society*, 136(646), pp. 222-238.

Panziera, L., Germann, U., Gabella, M., Mandapaka, P.V., Submitted in 2010. NORA - Nowcasting of orographic rainfall by means of analogs. *Quarterly Journal of the Royal Meteorological Society*.

Seed, A.W., 2003. A dynamic and spatial scaling approach to advection forecasting. *Journal of Applied Meteorology*, 42(3), pp. 381-388.

Wilks, D.S., Neumann, C.J., Lawrence, M.B. 2009. Statistical Extension of the National Hurricane Center 5-Day Forecasts. *Weather and Forecasting*, 24(4), pp. 1052-1063.

ACKNOWLEDGEMENTS

The research is funded by the Swiss National Science Foundation project *GeoKernels: kernel-based methods for geo- and environmental sciences (Phase II) (No 200020-121835/1)*. This work was possible through the help of the Radar and Satellite team (RASA) at Meteoswiss, in particular Dr. Urs Germann, Dr. Marco Gabella, Dr. Luca Panziera and Dr. Pradeep Mandapaka, which provided the orographic precipitation dataset and created stimulating discussions.

THE INFLUENCE OF SPRAWLED URBAN PATTERNS ON ECOSYSTEM FRAGMENTATION

F. Martellozzo [a, *], K. C. Clarke [b], N. Ramankutty [a]

[a] McGill University, Dept. of Geography, Sherbrooke St. 805 West, H3A 2K6 Montreal QC
f.martellozzo@mail.mcgill.ca – navin.ramankutty@mcgill.ca
[b] University of California - Santa Barbara, Dept. of Geography, 1720 Ellison Hall, Santa Barbara CA 93106-4060
kclarke@geog.ucsb.edu

KEY WORDS: Landcover modelling, Urban Sprawl, SLEUTH, spatial system fragmentation

ABSTRACT:

The monitoring of land cover transitions related to urban development over time is usually to find out the amount and location of land use change for planning purposes. The ability to anticipate a trend in urban sprawl behaviour for a specific region would give planners a useful tool to understand sprawl's long-term impact on a region, or even to take steps to prevent or retard it. The uncontrolled spread of cities into their surrounding rural and natural land is an issue of high interest in modern society and has been widely investigated; however, the so called urban "sprawl", in spite of being well known, remains controversial, hence among scholars there are no universal definitions for causes and variables related to it nor about its dynamics (Ewing et al. 2002). Although various studies have been dedicated to the measurement and monitoring of urban growth (Torrens 2008), they have limitations in providing generalizations of the characteristics of urban sprawl (Heikkila and Hu 2006)). To efficiently measure sprawl we must rely on metrics that consider it a matter of degree (Batty and Longley 1994; Batty 1974). For instance, scattered development and polycentric/multinucleated urban development are very similar; hence the distinction between these two different trends of growth is elusive and leads us to consider sprawl as a matter of resolution and scale as well (Batty 2008). In order to do so we recently developed a simple and useful metric that can detect the degree of dispersal of urban growth trend over time; this metric is simply obtained by merging together the information conveyed by the number of clusters and the average cluster's size of the landcover type under investigation. Furthermore in this study we want to investigate the correlation, as well as the influence of urban growth patterns on the surrounding landscape, in order to do so we applied the same metric to the whole landscape to detect the degree of dispersal, or level of fragmentation, of the whole ecosystem surrounding urbanized areas. A comparison of several cities in diverse geographic regions offered intriguing findings. In general, urban sprawl is associated with a higher level of fragmentation of the whole ecosystem, conversely when the urban pattern is more aggregated also the surrounding landscape is more compact.

1. INTRODUCTION

1.1 Urban sprawl and Landscape fragmentation

The most shared concept about urban sprawl is that it involves the spread of urban areas into the surrounding landscape (Harris and Ullman 1945). This concept is closely associated with the number of distinct clusters in the study area and the clusters' average sizes (Clarke et al. 2006). We describe the spatio-temporal patterns of urban forms focusing on a measure that can detect the degree of urban spatial dispersion over time and can be easily computed and widely applied. It is reasonable to imagine a fragmented region or a sprawled area as characterized by a large number of small disconnected urban land parcels (or clusters), while on the other hand an urban form characterized by a small number of urban cells that is large and connected leads to an urban growth pattern that follows cohesion principles. Imagine an area that is 50% urban. This zone could have one large highly aggregated central cluster, with high degree of spatial autocorrelation, or every other cell could be urban like a chessboard, completely disaggregated and with high negative spatial autocorrelation. The degree to which an urban system is previously clustered represents an initial condition that propagates into future patterns, and can take some time to change in overall structure. One could imagine a ranked size distribution of separate clusters, with one or two large clusters and more and more clusters as the size becomes smaller; this applies to any single landcover type, e.g. the urban form whether our interest is to find out sprawl or compactness, but can also be useful to determine the level of fragmentation of the whole landscape when considered as the contribution of each landcover type to ecosystem dispersion.

1.2 Dataset and case studies area

We compared three geographical regions: Calgary and Edmonton (Canada); Pordenone (Italy). These case studies have considerably different characteristics and patterns, for example: in size, rate of growth, population density, etc. In fact Pordenone is a fairly small but flourishing town in northeast Italy with less than 55.000 people (the area considered is ~ 100.000), while Calgary and Edmonton are both part of one of the fastest growing region in Canada with almost the highest population density in the Country.

The database used is a fusion of thematic archive maps, classified satellite imagery, census data and forecast maps of urban scenarios. The focus of this study is a broad spatial comparison of growth trends across varied urban areas in order to better understand and define sprawled patterns of urban diffusion as well as landscape fragmentation. The time span analysed for each dataset has been computed by merging past

* Corresponding author.

data analyses, and future forecast scenarios from growth modelling. Including forecasts of modelled future scenarios is of dramatic relevance: on the one hand it allows us to extend the time span of our database, on the other hand comparison between observed past data with simulated probable future is interesting. We applied the SLEUTH land use change and urban growth model due to the fact that it has been shown to produce realistic, valid and statistically robust results and has been applied extensively for urban growth planning and prediction for over a hundred cities around the world (Clarke et al., 2006). After SLEUTH was calibrated using the historical data, prediction was performed in order to obtain maps of possible future scenarios of urban spatial extent. Past data for Calgary and Edmonton are from 1988 to 2010 and forecast scenarios are up to 2040, while for Pordenone past data is from 1985 to 2005 and modelled data are up to 2035. These results were then conflated with the classified historical maps to observe spatial form and diffusion patterns over the time considered.

2. METHODOLOGY

2.1 From urban sprawl to landscape fragmentation

The metric we used in this study has been developed recently and usefully applied to determine sprawl and coalescence in northeast Italy. Below we briefly introduce the key concepts it conveys. To quantify this distribution of cluster sizes, we calculated both the average cluster size and the number of clusters and normalized these measures over the time span in order to have more comparable values. The following metric (that for simplicity we abbreviated as NDDI: Normalized Difference Dispersal Index) has then been computed to detect the degree of dispersion at the class level for a single landcover type (in this case urban) for each of the years of the timespan:

$$NDDI\_urb = \frac{(nc - acs)}{(nc + acs)} \quad (1)$$

where:
NDDI_urb is the metric proposed and applied only at the urban form
nc is the number of cluster
acs is the average cluster size

This metric can range from -1 to 1, assuming positive values where the *nc* parameter affects the urban growth more than the *na* parameter, i.e. when it is above the *x* axis urban growth is more influenced by the number of separate clusters rather than the average size of the clusters. Therefore positive values indicate growth characterized by sprawl rather than coalescence. Conversely it assumes negative values when the stronger influence is carried by the average size of clusters. One of the advantages of this metric is the simplicity to compute and that is intuitive to communicate.

As previously stated this study focuses on the assumption that highly fragmented urban patterns influence the entire surrounding ecosystem and increase the degree of dispersal also among the other landcover types within the same region (Irwin and Bockstael 2006). Given that the statistical approach used to determine whether a growth pattern is affected by sprawl or coalescence reflects the degree of spatial dispersion of the phenomenon investigated, it can be applied cartometrically to any single landcover that is present in the landscape object of study. The sum of the dispersion index for each landcover type or class should reflect the total dispersion, or the level of fragmentation of the whole ecosystem. We have to consider that not all the landcover types have the same importance and the same influence on the landscape. For example some classes might be so small that their effect on landscape fragmentation can be minimal, while classes that occupy most of the landscape will have a greater impact on ecosystem fragmentation. To rectify this effect, the contribution of each class to landscape disaggregation has been weighted according with the percentage of total area of each landcover type. Hence the index has been modified as follow:

$$NDDI\_l = \sum_{i->k} \left[\frac{(nc_i - acs_i)}{(nc_i + acs_i)} \cdot \frac{A_i \cdot 100}{A_l} \right] \quad (2)$$

where:
NDDI_l is the metric proposed and applied to the whole ecosystem
k different classes or landcover types form it
nc_i is the number of cluster of the *i*th class
acs_i is the average cluster size of the *i*th class
A_i is the area of the *i*th lancover type
A_l is the total area of the landscape

Equation 2 differs from equation 1 in that it is the sum of the dispersal metric applied singularly to each of the classes and weighted by the percentage of land occupied by each class. The time series obtained was then normalized by the maximum value observed.

Identical methodology was applied to over a 100 thematic maps from both modelling and real data of the three case studies presented. NDDI of the urban class have been performed and at the same time the general fragmentation of the system was computed with the proposed index (missing data have been retrieved by interpolation). In order to better investigate the relation between landscape fragmentation and urban sprawl several other indices have also been computed. These indices have been used to more robustly test our hypothesis. A description of findings and observations is given in the next paragraph.

2.2 Findings and discussion

Our main aim was to investigate the relation between urban growth trend and landscape pattern. Urban sprawled patterns influence the total landscape to be more fragmented, while more compact urban forms influence the whole landscape to be less fragmented. In order to do so, the overall linear correlation of the metrics described in eq. 1 and 2 (fig. 1) has been computed.

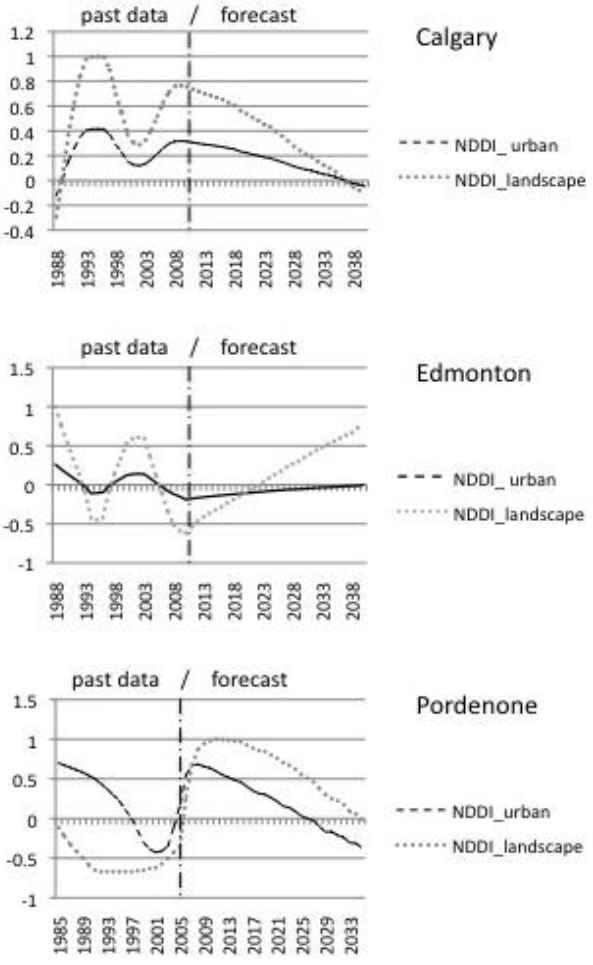


Figure 1. trend of NDDI_landscape and NDDI_urban

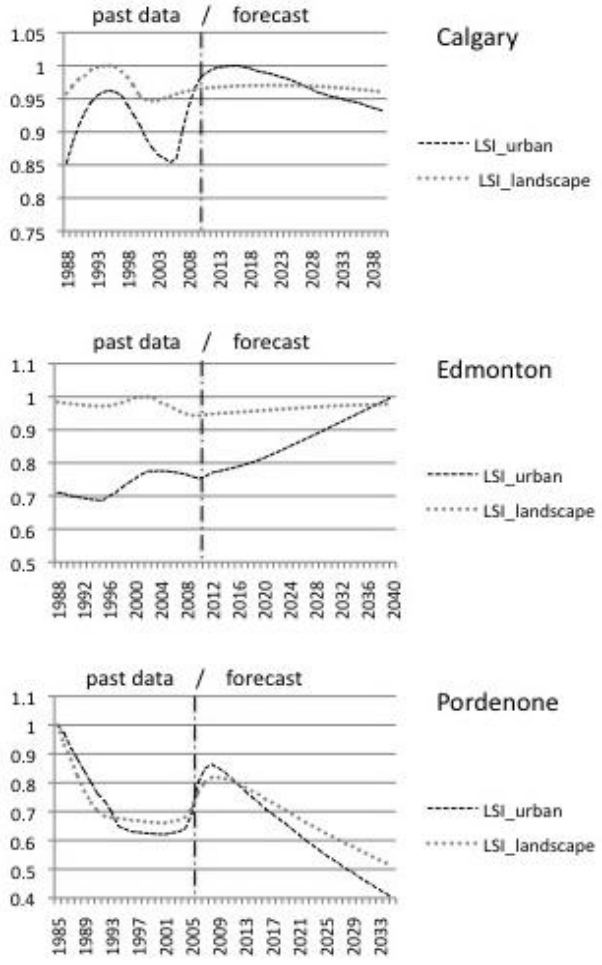


Figure 2. trend of LSI_landscape and LSI_urban

The linear correlation (R-square) between NDDI at the landscape level and the NDDI of the urban class is consistence through all the series with an average value of 0.856 (Calgary 0.91; Edmonton 0.82; Pordenone 0.84). This suggests that the two metrics are strictly correlated, hence the NDDI has been useful to investigate this correlation and we suggest that is urban sprawl to induce landscape fragmentation and not vice versa. Linear correlation can highlight the relation between two variables but not determine dependencies or causality.

In order to determine whether the findings mentioned above are just because of the specificity of the metric we developed, we also computed the Landscape Shape Index (LSI) both at the landscape and at the class level for all the elements of the time series. LSI provides a standardized measure of total edge or edge density that adjusts for the size of the landscape. Therefore it has a direct interpretation: at both the class and the landscape level it can be used to investigate cluster/patch aggregation or disaggregation. Specifically it describes a more disaggregated pattern when it increases, versus more compact patterns when it decreases (Fig. 2).

Linear correlation has also been computed for the LSI at the landscape and at the class level (for the urban landcover type). High linear correlation values are shown, almost identical of what has been observed between the NDDI metrics; the average value of R-square registered is of 0.843 (Calgary 0.81; Edmonton 0.75; Pordenone 0.97).

The fact that correlation between urban pattern dispersion and landscape fragmentation is present in a metric completely independent from the one we previously used, is a reliable support of the fact that the two phenomena are correlated. In order to better assess the robustness of the method proposed, linear correlation between LSI at the landscape level and NDDI at the landscape level has been also performed. In this case linear correlation still has a good score even if it is not as high as what was observed with LSI and NDDI independently; in fact the average R-square was 0.773 (Calgary 0.70; Edmonton 0.76; Pordenone 0.86).

2.3 Direct versus indirect impact

The secondary aim of this study was to determine whether the correlation between urban dispersed patterns and fragmented landscape has to be explained with the fact that when we look at the landscape, the urban pattern is included in it. Therefore we can call it the “direct” effect of urban dynamic on the level of aggregation of landscape. Hence is our convincement that sprawled urban growth has on ecosystem fragmentation both “direct” and “indirect” impacts; if so it would mean that landscape disaggregation is higher when urban sprawl is high because on the one hand urban sprawl is part of it (this is what we found evidences of), on the other hand that urban dispersion influences also other landcover types to be more disaggregated. Hence NDDI at the landscape level without considering the urban form has been also computed, this means considering urban pixels as “holes” in the image. If there is correlation

between NDDI at the landscape level and the same NDDI minus the urban component, is possible to argue that fragmentation or compactness is also influenced indirectly. Preliminary results suggest that the "indirect" impact shows up (we have fairly good level of correlation) only in some of the cases; furthermore correlation seems to be more case specific than pattern specific. Therefore it suggests that "indirect" impact might be influenced by specific regional characteristics, even so this assumption is still object of investigation.

3. CONCLUSIONS

Results show that, in general, a dispersed urban pattern is associated with a higher level of fragmentation of the aggregate system rather than by development mostly characterised by coalescence. Hence higher urban sprawl levels are related to higher fragmentation at the landscape level, conversely when growth takes place without sprawl the whole system is more compact. Our findings support the assumption that scattered urban development is one of the causes of system fragmentation and not vice versa, however this hypothesis is yet to be tested.

Furthermore our aim is also to analyze the similarities and differences between landcover modifications across differently sprawled patterns in different locations. The capability to forecast different patterns of urban growth in relation to the global degree of fragmentation in the system will help on the one hand to better understand the theory of sprawled urban patterns, on the other hand it will provide useful tools to understand, and possibly tackle, the deleterious consequences of urban sprawl as portrayed in the contemporary literature.

REFERENCES

Batty M., 1974. Urban density and entropy functions. *Journal of Cybernetics*, 4, 2, pp. 41-55.

Batty M. and Longley P. A., 1994. *Fractal Cities. A geometry of form and function*. London, Academic Press.

Batty M. 2008 . *The size, scale and shapes of cities*. Science, 319, 769-771.

Clarke K. C., Gazulis N., Dietzel C. K. 2006. A Decade of SLEUTHing: Lessons Learned from Applications of a Cellular Automaton Land Use Change Model. *Twenty Years of the International Journal of Geographic Information Sciences.* T. & F. London 2006, pp. 413-426.

Ewing, R., Pendall, R., & Chen, D. D. T., 2002. *Measuring sprawl and its impact*. Washington D.C.: Smart Growth America.

Harris C. and Ullman E., 1945. The nature of cities. *Annals of the American Academy of Political and Social Science*, 242, 7-17.

Heikkila E. J. and Hu, L., 2006. Adjusting spatial-entropy measures for scale and resolution effects. *Environment and Planning B: Planning and Design*, 33, 845- 861.

Herold M., Scepan J., Clarke, K. C., 2002. The use of remote sensing and landscape metrics to describe structures and changes in urban land uses. *Environment and Planning A*, 34, pp. 1443-1458.

Irwin E. G. and Bockstael N. E., 2007. "The evolution of urban sprawl: evidence of spatial heterogeneity and increasing land fragmentation." Proceedings of the National Academy of Sciences (PNAS). Dec 26, 2007; 104 (52): 20672-20677.Jensen R. J. 1996. *Introductory digital image processing. A remote sensing perspective.* 2nd Ed. Prentice-Hall: Upper Saddle River, NJ.

Torrens, P. M., 2008. A Toolkit for Measuring Sprawl. *Applied Spatial Analysis*, 1, pp. 5–36

ACKNOWLEDGEMENTS

We would like to thank:

Dr. David Price and Dr. Ron Hall (National Forestry Service of Canada) for the precious support and suggestions.
The Government of Canada – Bureau of Foreign Affairs and International Trade for having supported this research through the CBIE Post-doctoral Fellowship Program.
NASA for having supplied remote sensed images.

Author Index