



# CaGIS

Freundschuh, S.M. and D. Sinton (Editors), 2018. Conference Proceedings, AutoCarto/UCGIS 2018.  
The 22nd International Research Symposium on Computer-based Cartography and GIScience,  
Madison, Wisconsin, USA. May 22-24, 2018.

# Frontiers of Geospatial Data Science

Publication of the Cartography and Geographic Information Society (CaGIS)  
and the University Consortium for Geographic Information Science 

## Preface

Since its inception in Reston, VA in 1974, the AutoCarto biennial conference series has brought together leading researchers in cartography and geographic information systems and science to their present cutting-edge research in cartography, geospatial science and visualization.

AutoCarto 2018 was held jointly with the University Consortium for Geographic Information Science (UCGIS) at the Historic Concourse Hotel in Madison, Wisconsin, USA on May 22-24. This is the second time that the UCGIS Symposium and AutoCarto have been coordinated for participants, the first being in 2006 in Vancouver, Washington.

We received 1 position paper, 26 extended abstracts to be considered for both presentation at the conference and publication consideration in the CaGIS Journal, 23 short abstracts to be considered for presentation only, 3 abstracts for lightning talks, 3 abstracts for poster presentations and 14 abstracts for the student poster competition. Each submission was reviewed by at least two Program Committee members.

Of these 70 abstracts, 24 were accepted as full papers for presentation at the conference, 8 papers were accepted for review and publication consideration in the CaGIS Journal, 7 abstracts were accepted as lightning talks, 3 abstracts were accepted as poster presentations, and 8 abstracts were accepted for the student poster competition.

The theme of this conference was to explore the *Frontiers of Geospatial Data Science*. Presentations and discussions centered around connections between **Geospatial Science** and the burgeoning field of **Data Science** seen in newly named academic departments, and in calls for funded research. At this event we explored emerging opportunities and challenges for the geospatial and mapping sciences with an eye on trends in deep learning, data-intensive computing platforms, and visualization, as well as geospatial big data sources and applications such as location-aware social media, autonomous vehicles, and earth-observing micro- and nano-satellites and sensors.

The breadth of the topics in this volume reflects the breadth of the research in this field. Topics include mapping social media data, mobile mapping, 3D mapping tools, interoperability, spatial flow algorithms, volunteered geographic information, geographic ontologies and geographic features, mapping and natural hazards, geovisualization and education, and research questions dealing with large databases.

Organizing a successful conference is not possible without the commitment, additional effort, and diligent help of many people. We would like to thank: the Program Committee for their timely and thorough reviews, and Taylor & Francis Publishers, ESRI and MAPSOL for providing financial support. Finally, would like to thank Enki Yoo for stepping in at the last minute to provide her insightful reviews of submissions.

## **AutoCarto 2018 Program Committee**

Sarah Battersby, *Tableau*

Nicholas Chrisman, *CaGIS Editor*

Dan Cole, *Smithsonian Institution*

Scott M. Freundschuh, *University of New Mexico*

Karen Kemp, *University of Southern California*

Ross Meentemeyer, *North Carolina State University*

Ian Muehlenhaus, *University of Wisconsin Madison*

Diana Sinton, *University Consortium for Geographic Information Science*

Shashi Shekhar, *University of Minnesota Minneapolis*

Eun-Hye Enki Yoo, *State University of New York at Buffalo*

## Table of Contents

<b>Page</b>	<b>Authorship</b>	<b>Title</b>
1-7	Benjamin Acker and May Yuan	<i>Spatiotemporal Modeling of Traffic Accident Using Network Kernel Density Estimation and Space Syntax Analysis: A Case Study in Dallas, Texas, USA</i>
8-9	Ola Ahlqvist	<i>Design-based approaches to identify causation in GIS education research</i>
10	Marc Armstrong	<i>Active Symbolism: Towards a New Theoretical Paradigm for Statistical Cartography</i>
11-17	Marc Armstrong, Shaowen Wang and Zhe Zhang	<i>The Internet of Things and Fast Data Streams: Prospects for Geospatial Data Science in Emerging Information Ecosystems</i>
18-25	Anne Berres, Rajasekar Karthik, Alexandre Sorokine, Philip Nugent, Melissa Allen, Ryan McManamay, Varun Chandola, Arshad Zaidi and Jibonananda Sanyal	<i>EWN-KDF: A Knowledge Discovery Framework to Understand the Energy Water Nexus</i>
26-35	Aileen Buckley and Kevin Butler	<i>Analysis of the Adoption of Story Maps</i>
36-42	T. Edwin Chow	<i>A Crowdsourcing-Geocomputational Framework of Mobile Crowd Estimation</i>
43	Pranab K. Roy Chowdhury, Jeanette Weaver, Eric Weber, Dalton Lunga, St. Thomas LeDoux, Amy Rose and Budhendra L. Bhaduri	<i>Understanding of Intra-city Electricity Consumption Patterns Through Settlement Characterization</i>
44-49	Kevin Curtin	<i>Unmanned Aerial Vehicle Logistics Modeling and Performance: A Demonstration of Integrative Data Science</i>
50-51	Brent Dell and May Yuan	<i>A 3D spatial optimization problem for determining optimal locations for Bluetooth beacon placement</i>
52-57	Somayeh Dodge	<i>Embracing visualization as a key element in computational movement analytics</i>
58	Weiwei Duan and Yao-Yi Chiang	<i>Automatic Alignment of Geographic Features in Contemporary Vector Data and Georeferenced Historical Maps Using Reinforcement Learning</i>
59-63	Weiwei Duan, Yao-Yi Chiang, Craig A. Knoblock, Johannes H. Uhl and Stefan Leyk	<i>Automatic Generation of Precisely Delineated Geographic Features from Georeferenced Historical Maps Using Deep Learning</i>
64	Xin Feng, Shaohua Wang, Alan T. Murray, Yuanpei Cao, Song Gao	<i>Optimizing Activity Locations in GIS using a Multi-Objective Trajectory Approach</i>
65	Amy Frazier and Benjamin Hemingway	<i>Unmanned Aircraft Systems and the Atmospheric Boundary Layer: A New Frontier for Geospatial Data Science?</i>
66	Lee Hachadoorian	<i>Gerrymandering and Geospatial Analysis of Redistricting Plans</i>
67	Yanghui Kang and Mutlu Özdoğan	<i>Forecasting County-Level Maize Yield with Deep Learning and Satellite Remote Sensing</i>
68-73	Caglar Koçlu, Bryce Dietrich and Ryan Larson	<i>Geovisual text analytics for exploring public discourse on Twitter: A case study of immigration tweets before and after the January 27, 2017 Travel Ban</i>
74-75	Barry Kronenfeld, Larry Stanislawski, Tyler Brockmeyer and Barbara Buttenfield	<i>Area-Preserving Simplification of Polygon Features (research paper)</i>
76-89	Irma Kveladze and Niels Agerholm	<i>GeoVisual Analytics for the Exploration of Complex Movement Patterns on Arterial Roads</i>
90	Dan Lee, Nobbir Ahmed and Iffat Chowdhury	<i>Incorporating Changes in Multi-scale Databases</i>
91	Samuel Levin and May Yuan	<i>Viewshed Analysis for UAS Flight Planning</i>
92	Xiaojiang Li and Carlo Ratti	<i>Using deep learning and Google Street View to quantify the shade provision of street trees in Boston, Massachusetts</i>
93-99	Xiao Li, Daniel Goldberg, Tracy Hammond and Xingchen Chen	<i>Geospatial Machine Learning: Predicting Accident-Prone Road Segments Using GIS and Data Mining</i>
100	Qingmin Meng	<i>Deep leaning for geospatial big data analytics: Terrestrial ecological systems recognition and classification assessment</i>
101-102	Amy Moore	<i>Development of a GIS-Based Model to Examine Alternative Scenarios for Last-Mile Freight Delivery</i>

103-105	Thomas Pingel, Matthew Mendez and Earle Isibue	<i>From Point Clouds to Tactile Maps: How Lidar and Photogrammetry Can Improve Maps for People with Visual Impairments</i>
106-108	Anthony Robinson	<i>Position Paper: Understanding the Analytical Affordances of Absence in Spatial Data Science</i>
109	Robert Roth, Carl Sack, Meghan Kelly, Nick Lally and Kristen Vincent	<i>MapStudy: An Open Source Survey Tool for Researching Interactive Web Maps</i>
110	Ethan Shavers and Lawrence Stanislawski	<i>Aerial Imaging and Lidar Point Cloud Fusion for Low-Order Stream Identification</i>
111-119	Lawrence Stanislawski, Barry Kronenfeld, Barbara Buttenfield and Tyler Brockmeyer	<i>Generalizing Linear Stream Features to Preserve Sinuosity for Analysis and Display: A Pilot Study in Multi-Scale Data Science</i>
120-121	Hoda Tahami, Bo Zhao and David Wrathall	<i>Visualizing Sea Level Rise Induced Migration Using Hexagonal Grids</i>
122	Ross Thorn and Shane Loeffler	<i>Flyover Country: Mobil Visualization of Geoscience Data</i>
123-124	Johannes Uhl, Stefan Leyk, Yao-Yi Chiang, Weiwei Duan and Craig Knoblock	<i>Exploring the potential of deep learning for settlement symbol extraction from historical map documents</i>
125	E. Lynn Usery and Dalia Varanka	<i>The Evolution of Cartography in the Digital Age: From Digitizing Vertices to Intelligent Maps</i>
126-129	Dalia Varanka, William Baumer and Logan Powell	<i>Maps as Graphs: An Implementation for Cartographic Retrieval of Linked Geographical Data</i>
130-131	Nancy Wiegand	<i>Web-Based Demo to Show Ontology Matches</i>
132-134	Xin Yu, Yuanbin Cheng, Yijun Lin, Fan Pan, Yao-Yi Chiang and Dimitris Stripelis	<i>MAPINS: An intra-city PM2.5 modeling web application using a scalable data management and analysis system integrating public multi-source data</i>
135-145	Roberto Zagal-Flores, Miguel-Felix Mata-Rivera, Christophe Claramunt and Edgar Armando Catalán-Salgado	<i>Identifying crime patterns in Mexico using geo-social mining and clustering</i>
146	Qunshan Zhao, Heather Fischer, Wei Luo and Elizabeth Wentz	<i>Community resilience in Maricopa County, Arizona, USA: the analysis of indoor heat-related death and urban thermal environment</i>
147-155	Xiran Zhou and Jun Liu	<i>Deep Convolutional Neural Networks for Map Type Classification</i>

# Spatiotemporal Modeling of Traffic Accidents Using Network Kernel Density Estimation and Space Syntax Analysis: A Case Study in Dallas, Texas, USA

Benjamin Acker and May Yuan

**ABSTRACT:** This paper seeks to understand and model the effect of space syntax, and site and temporal characteristics on accident likelihood. Network kernel density estimation is used to inspect how accident varies over time and place. Logistic regression and random forest classification are used to predict accident likelihood on 100m segments at 1-hour intervals in Dallas, TX, using explanatory variables including time, the space syntax variables of integration and choice, and other site and dynamic characteristics.

**KEYWORDS:** traffic accident, network spatial analysis, kernel density estimation, space syntax analysis, spatiotemporal modeling

## Introduction

### *Research Questions and Objective*

This research sought to develop a model that will classify the likelihood of traffic accidents occurring along individual street segments using both static and dynamic data, including road characteristics, space syntax analysis variables, and nearby accidents. Thus, the main research questions are as follows: (a) How do road characteristics, such as functional class, influence accidents? (b) How does the space syntax of a road network influence traffic accidents? (c) How does a traffic accident near a road segment influence that segment's risk of having a new accident occur? (d) Are accidents more likely to occur at certain times or days? This research posits that traffic accidents are influenced by site characteristics (e.g. spatial structure of roads) and situational characteristics (e.g. function, usage, and interactions). The four research questions attempt to examine both site and situational characteristics using static and dynamic data to gain new insights into the spatiotemporal distribution of traffic accidents on street networks. While Dallas is the study area for this research, the methodology is generic and applicable to other cities.

## Literature Review and Background

Previous research has used kernel density estimation (KDE) to identify areas of high and low risk for traffic accidents. Anderson (2009) used KDE as part of a broader attempt to profile accidents, but used KDE on a 2D surface, which is not appropriate for events occurring on linear networks. Xie and Yan (2008) attempted to do network KDE of accidents, but problems were identified in their research by Okabe et al. (2009), which developed modified version of kernel density estimation for networks that is mathematically sound. An additional version of network kernel density estimation was developed in Baddeley et al. (2015).

Space syntax has also been used to model accident risk in Obeidat et al. (2017) and general traffic flow Serra et al. (2015). However, these models did not include a temporal component

(i.e. how the time of day/week influence traffic), nor did they investigate the second-order effect that nearby accidents have on accident likelihood.

## Methodology

### ***Network Kernel Density Estimation (KDE)***

KDE is a common method to identify hotspots of accidents. This research adds temporality to KDE to discriminate between persistent hotspots and hotspots that only occur at certain times of day and/or days of the week. Since networks are essentially a quasi-1D surface, a special form of KDE needs to be used, such as that developed by Okabe et al. (2009), visualized in Figure 1. There are software implementations of such methods in the R library Spatstat, which implements Baddeley et al. (2015) and the software SANET, which implements Okabe and Sugihara (2012). SANET was used in this research to calculate the accident density for two-hour blocks of time and a script was developed to identify persistent and temporary hotspots.

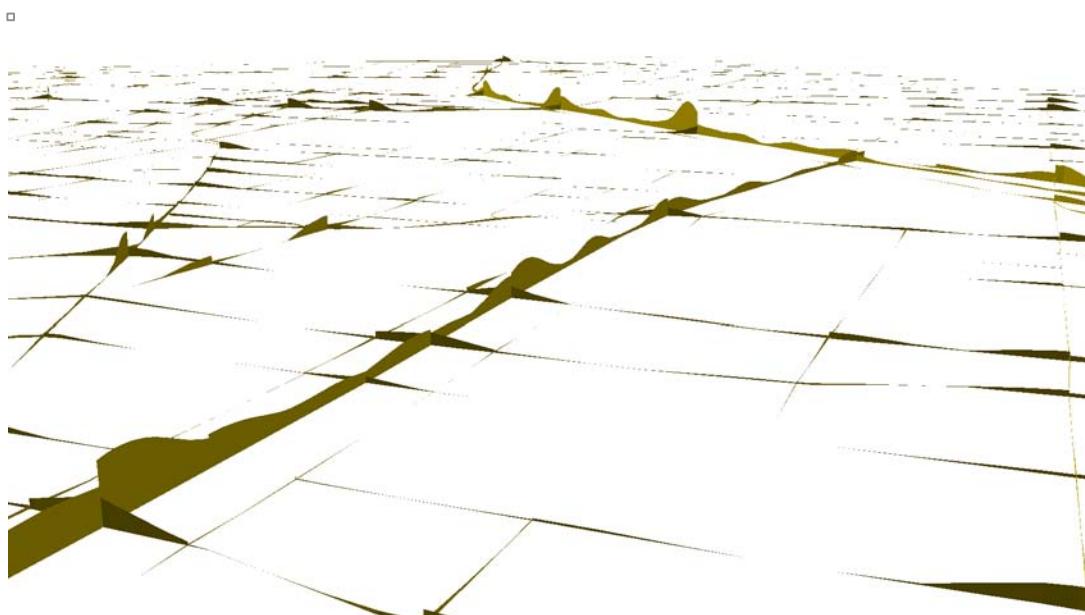


Figure 1: Accident KDE in Dallas.

### ***Predictive Model and Space Syntax Analysis***

This research developed predictive models of accident likelihood in Dallas using logistic regression and random forest classification. The dependent variable is a binary value indicating whether an accident occurred on a 100m segment for a given hour. Logistic regression and random forest classification were chosen as they output the probabilities associated with each classification. Independent variables include space syntax measures, road class, proximity to major intersection, time, weather, and nearby accidents.

The research conducts space syntax analysis using angular segment measures on the centerline road network of the Dallas. The output includes measures of choice and of integration for each road segment of the network. Turner (2007) demonstrated that angular segment analysis is on par

with axial analysis, the traditional form of space syntax analysis, while being computationally simpler.

This research incorporates the second-order effects caused by nearby accidents, postulating that if an accident is nearby a road segment in both space and time, that accident will increase the likelihood of an accident occurring soon after on that road segment, through clustering, as drivers are distracted by the nearby accident.

## Results

### *Network KDE*

For the sake of computational efficiency, the equal-split discontinuous kernel density function was chosen over the equal-split continuous kernel density function from SANET. Figure 2 shows the output of this analysis with bandwidth of 300ft and cell size of 300ft for a 2-hour interval. All 2-hour interval outputs were classified into hotspots and subsequently, temporary and persistent hotspots were identified.

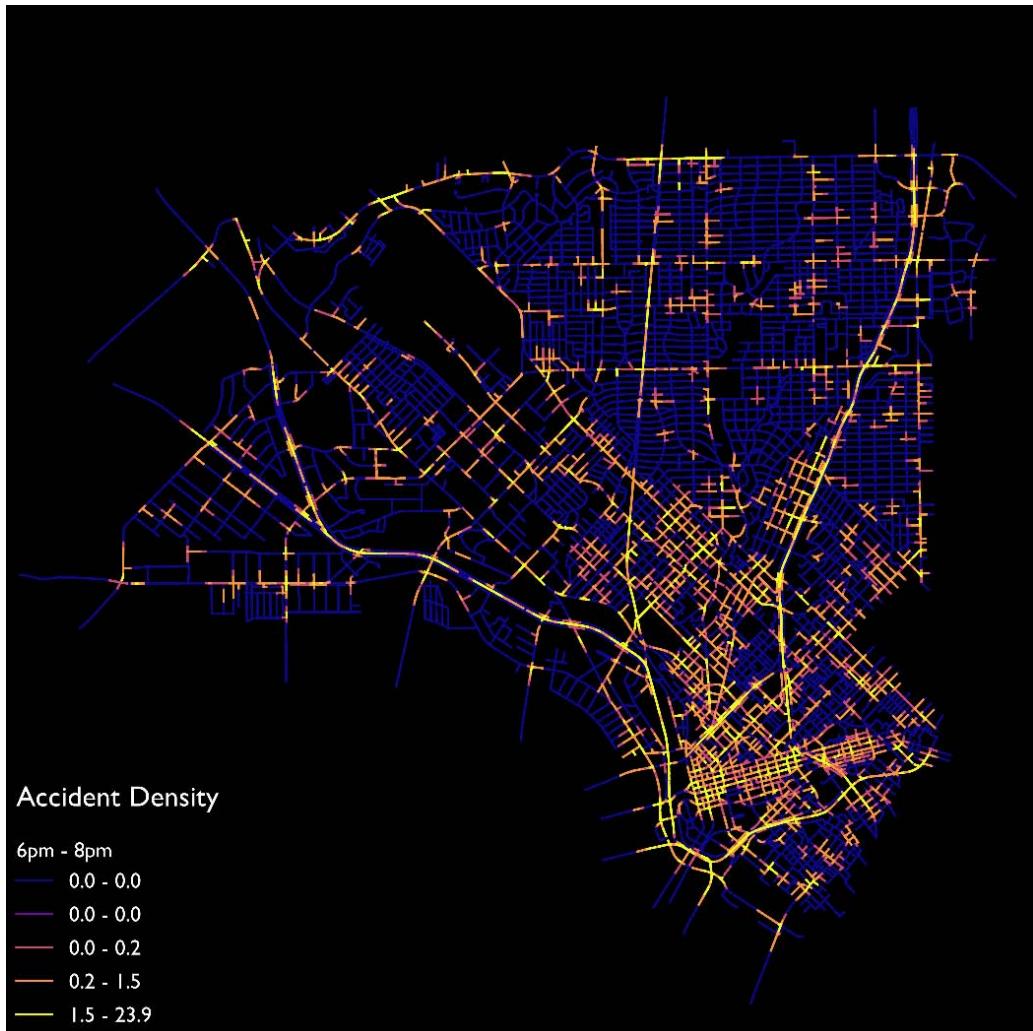


Figure 2: KDE identifies high accident density at road intersections during 6-8 pm.

## Predictive Models and Space Syntax Analysis

Prior to constructing this model angular segment analysis, was run on the roads within 10 kilometers of Dallas. This network is larger than the study area and was chosen to mitigate edge effects, which space syntax analysis is vulnerable to. Length was used as a weight, and the range was specified as 10 kilometers. Figure 3 shows the choice measure from the analysis.

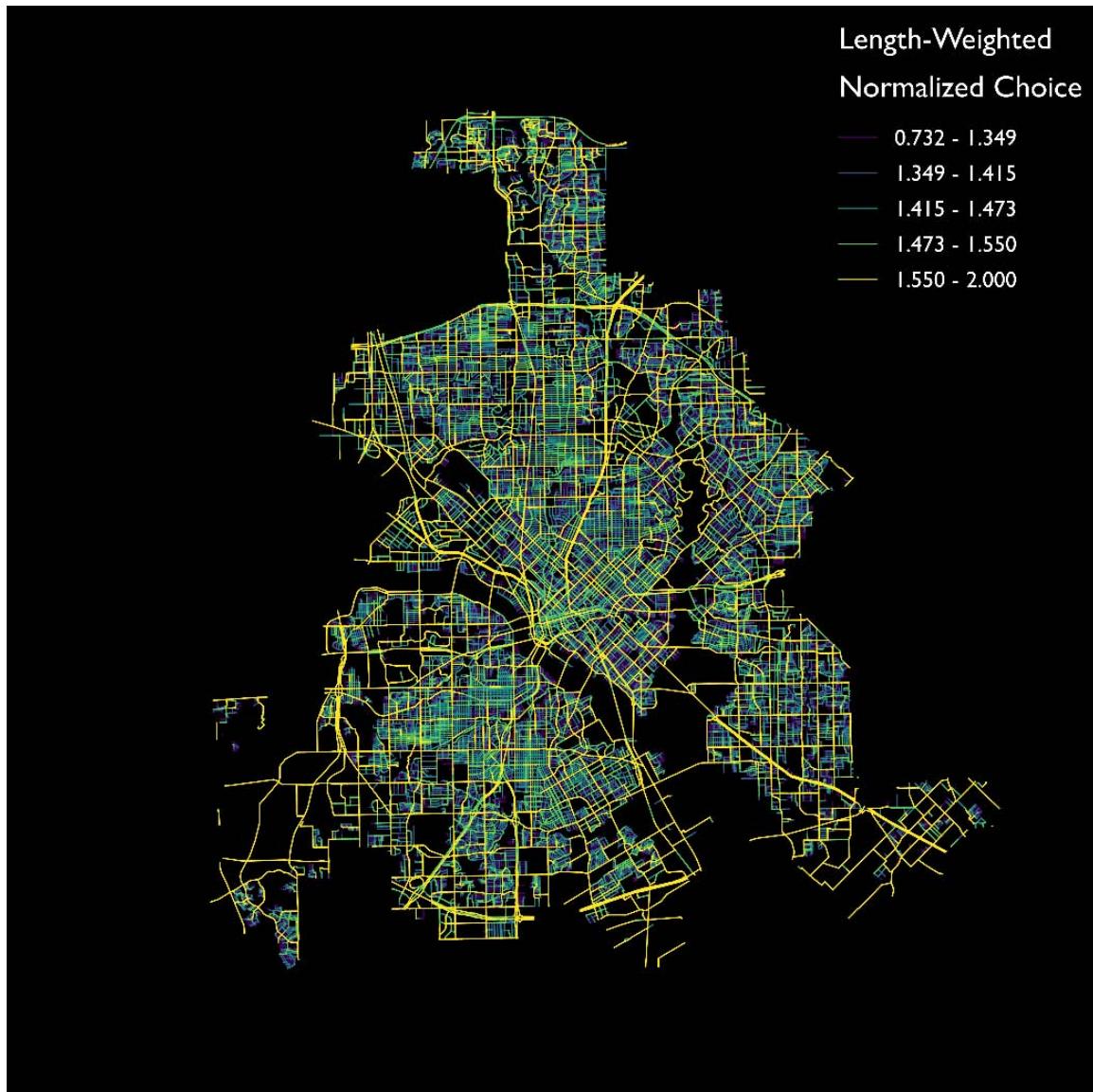


Figure 3: The relative measures of a road segment chosen as part of a shortest path in Dallas.

Accident data from 2015 to 2016 were used. Training and testing datasets were constructed by separating odd and even days. The effect of time on accident density was modeled with separate Fourier transformations for weekdays and weekends, with r-squared values of 0.95 and 0.80, respectively.

Both a logistic regression model and a random forest model were constructed with all variables. The explanatory performance of each variable was considered relative to one another. For the logistic regression model, change in area under the receiver operating characteristic (ROC) curve was used, as the training data set was too large for statistical significance to be meaningful. For the random forest model mean decrease in accuracy and mean decrease in Gini coefficient were used. Variables that performed poorly were excluded from the final models.

Variables that provided strong explanatory power included time, both space syntax variables, road type, proximity to a major intersection, and the cascading effect of traffic accidents. The performance of both models was comparable, with accuracies of 84.11% and 85.42% and RMSE of 0.3350 and 0.3272, for the logistic regression and random forest models, respectively. Accident likelihood was broken into five categories, based on the quantiles of each model's output and is shown mapped with the logistic regression model for 21:00 on a Friday in Figure 4.

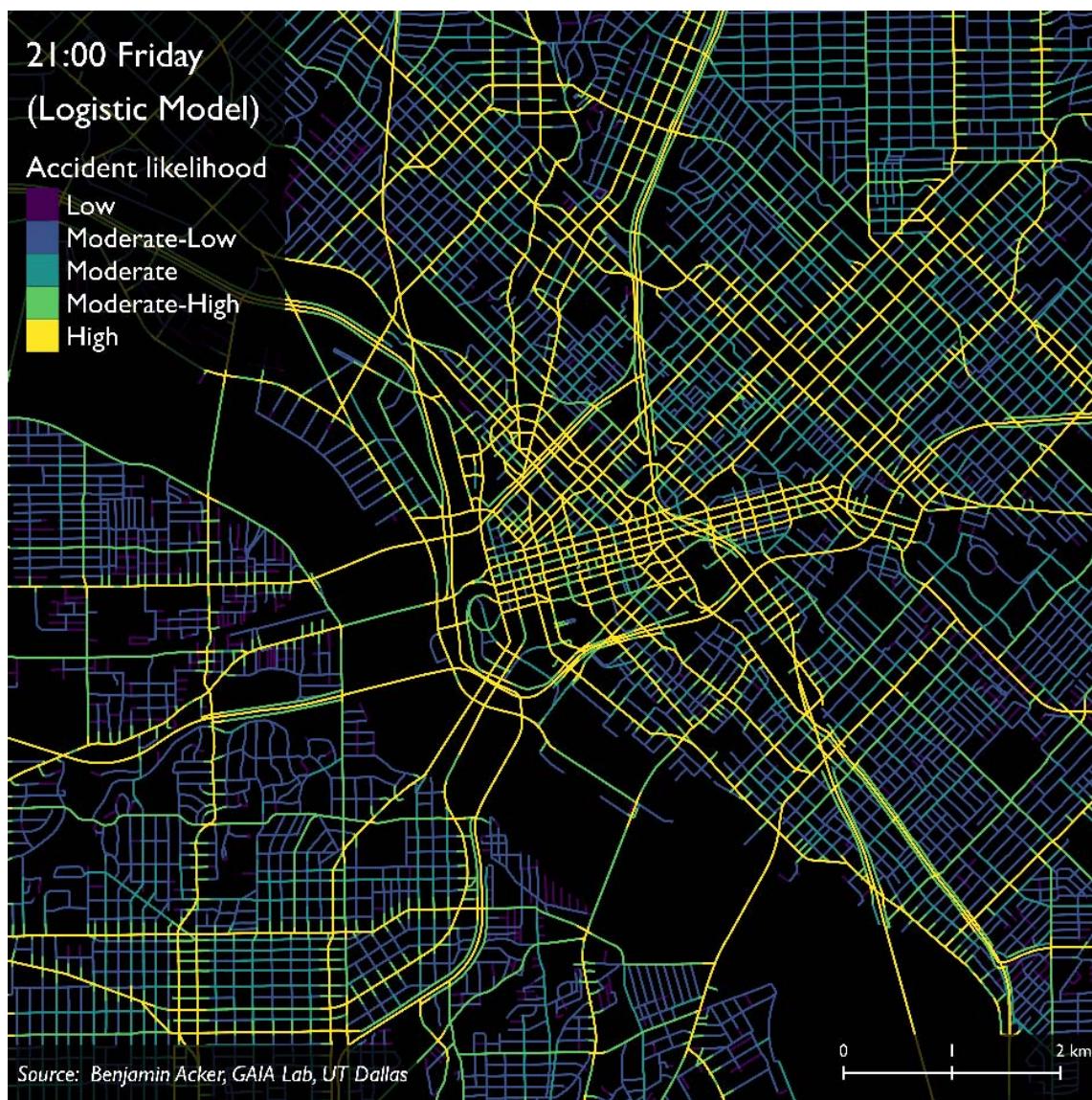


Figure 4: Accident risks on Dallas streets on 9 pm Fridays.

## Conclusions

Network KDE clearly highlights roads and intersection with a high density of accidents and with the addition of a temporal component, discriminated between persistent and temporary hotspots. Both logistic regression and random forest modeling have demonstrated that space syntax analysis is a meaningful predictor of traffic accidents, alongside other site and dynamic characteristics. By integrating these analyses together with temporal and second-order effects, this model is able to classify road segments in Dallas based on their relative risk of accident occurrences. All methods used and developed in this project are site-independent, so the methods are applicable to other cities.

## References

- Anderson, T. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis and Prevention*, 41, 3, pp. 359-364.
- Baddeley, A., Rubak, E., and Turner, R. (2016). *Spatial point patterns : Methodology and applications with R* (Interdisciplinary statistics).
- Obeidat, B., Alhashimi, I. and Tawalbeh, S. (2017). Urban Grid and Traffic Safety: Using space syntax as an assessment tool. In: Heitor, T., Miguel, S., Silva, J. P., Bacharel, M. and da Silva, L. C. (eds.) *Proceedings of the 11<sup>th</sup> Space Syntax Symposium, 11<sup>th</sup> International Space Syntax Symposium, 03-07 July 2017, Lisbon, Portugal*. Instituto Superior Tecnico, University of Lisbon. pp. 99.1-99.9.
- Okabe, A. and Sugihara, K. (2012) *Spatial Analysis Along Networks: Statistical and Computational Methods* (Statistics in practice). Chichester, West Sussex: Wiley.
- Okabe, A., Satoh, T., and Sugihara, K. (2009). A kernel density estimation method for networks, its computational method and a GIS-based tool. *International Journal of Geographical Information Science*, 23, 1, pp. 7-32.
- Serra, M., Hillier, B. and Karimi, K. (2015). Exploring countrywide spatial systems: Spatio-structural correlates at the regional and national scales. In: Karimi, K., Vaughan, L, Sailer, K., Palaiologou, G. and Bolton, T. (eds.) *Proceedings of the 10<sup>th</sup> International Space Syntax Symposium, 10<sup>th</sup> Space Syntax Symposium, 13-17 July 2015, London, UK*. Space Syntax Laboratory, The Bartlett School of Architecture, University College London. pp. 84:1-84:18.
- Turner, A. (2007). From Axial to Road-Centre Lines: A New Representation for Space Syntax and a New Model of Route Choice for Transport Network Analysis. *Environment and Planning B: Planning and Design*, 34, 3, pp. 539-555.
- Xie, & Yan. (2008). Kernel Density Estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems*, 32, 5, pp. 396-406.

**Benjamin Acker**, Master's Student, School of Economic, Political and Policy Sciences, University of Texas at Dallas, Dallas, TX

**May Yuan**, Ashbel Smith Professor, School of Economic, Political and Policy Sciences, University of Texas at Dallas, Dallas, TX

# Design-based Approaches to Identify Causation in GIS Education Research

Ola Ahlqvist

**ABSTRACT:** The field of Geography education research has been described as an area of inquiry "...where research in geography and in education overlap" (S. Bednarz, 2000). Since each of those fields can be subdivided into specific areas (e.g. spatial analysis, ecological analysis, or regional complex analysis in Geography and learning theory, teacher education, or applied topics in Education) any intersection of these sub-domains forms distinct areas for geographic education research. In this paper we are primarily concerned with the intersection of GIS&T with educational theory and application. With GIS&T being a relative newcomer in the geographic domain, the track record of GIS education research is clearly much shorter than that of geography education research. It has slowly emerged as GIS&T itself evolved to a distinct discipline in the early 90's with a few important milestones related to curriculum development (e.g. the NCGIA Core Curriculum in 1990) and research progress in the following decades (T. G. Baker, Kerski, Huynh, Viehrig, & Bednarz, 2012). Throughout this time there has been a growing assumption that GIS&T can help in teaching a range of skills related to spatial thinking (National Research Council, 2005). But GIS education research has still not managed to provide foundational and guiding evidence of the effectiveness in GIS classrooms. In an effort to consolidate efforts by outlining a coherent research agenda, Baker et al. (2015) identified four broad foci for continued research: (1) connections between geospatial technologies (GST) and geospatial thinking; (2) learning GST; (3) learning other content matter with or through GST; and (4) educators' professional development with GST. The nature of the proposed research questions also signaled a shift in emphasis from asking if students learn or what they learn, to asking how learning happens and why? Example questions could be "How do students learn GST?", "How does the learning environment impact learning?", "What specific aspects of a geospatial task or analytical process are helping or hindering learners?"

These types of questions a research design focused on causation and the actual learning process, rather than description and correlation, something we all know is a lot harder, if not impossible in some cases, save for purely laboratory, experimental settings. In this work I seek to build on their efforts and draw attention to the designs and approaches of current research on GIS&T education. Based on a review of papers published 2007-2017 in three leading journals I will discuss the slow shift from descriptive, smaller-scale efficacy studies papers that start to address the important causal mechanisms and processes behind GIS&T learning. I also introduce Design-Based Research (DBR) as an approach that can provide a much needed supplement to the existing methods employed by our community, particularly in order to address questions about learning mechanisms and processes. I will give an overview of DBR approaches and give examples from the existing literature and our own research of how design-based research on GIS education can be conducted in the undergraduate geography classroom setting.

**KEYWORDS:** GIS education, design-based research

## References

- Baker, T. G., Kerski, J. J., Huynh, N. T., Viehrig, K., & Bednarz, S. W. (2012). Call for an agenda and Center for GIS education research. *Review of International Geographical Education Online (RIGEO)*, 2(3), 254.
- Baker, T. R., Battersby, S., Bednarz, S. W., Bodzin, A. M., Kolvoord, B., Moore, S., Uttal, D. (2015). A Research Agenda for Geospatial Technologies and Learning. *Journal of Geography*, 114(3), 118–130.
- Bednarz, S. (2000). Geography Education Research in the Journal of Geography 1988–1997. *International Research in Geographical and Environmental Education*, 9(2), 128–140.

**Ola Ahlqvist**, Professor, Department of Geography, The Ohio State University, Columbus, OH 43210

# Active Symbolism: Towards a New Theoretical Paradigm for Statistical Cartography

**Marc P. Armstrong**

**ABSTRACT:** Conventional statistical cartography practices situate a mapmaker in a computer environment in which choices are made about a limited number of map characteristics. A map is then produced using these choices, some of which may be defaults. A resulting map often will be judged inadequate and an alternative will be generated, typically by changing some parameter value (*e.g.*,  $n$  choropleth classes). This is backwards. The purpose of this paper is to initiate a paradigm change in statistical cartography, one in which the production of maps switches from a sequence of actions taken by the map maker to a process of specifying criteria, and then selecting, and possibly modifying, a solution that satisfies these elicited criteria and others derived from cartographic theory and praxis. The cartographer's role in this paradigm shifts from involvement in an ill-defined collection of low-level software mediated tasks to a higher level. This Active Symbolism approach is described and then illustrated using the production of dot maps to illustrate how collections of intelligent design agents are used to generate emergent alternatives that can be evaluated against design criteria.

Three kinds of stored knowledge guide the production of statistical maps.

- Geometrical knowledge consists of feature descriptions of absolute and relative locations (*e.g.*, dot locations, administrative boundaries).
- Structural knowledge is a representation of expertise that may be derived from the cartographic literature (*e.g.*, guidance about dot coalescence) and from knowledge engineering practices.
- Procedural knowledge selects and deploys operators that perform statistical mapping tasks.

These three kinds of knowledge must be accessible to individual, neighborhood and global design agents. Individual symbols (*e.g.*, dots) are active agents that are able to, for example, position, resize or color themselves according to governance rules (using structural and procedural knowledge). The assumption of a particular state by an agent, however, may move it into a local optimum, one that is optimal for, say, a pair of dots, but sub-optimal when viewed at a different scale. Overcoming premature convergence into local optima requires a higher level of supervisory control. To achieve this goal, local supervisors operate in neighborhoods defined either by distance (*e.g.*, within a radius) or by topology (adjustable triangulation of network relations among geometric objects). Supervisory design agents are able to mediate among individual agents and, possibly across administrative borders, to yield a higher state of optimality than is achieved by individual agents operating independently. Finally, global operators consider complete alternative solutions and evaluate them according to global criteria that measure the success of a design. It is at this point that humans enter into the picture, using their highly developed pattern recognition capabilities to support the evaluation of alternatives and to make trade-offs among competing objectives.

**KEYWORDS:** statistical cartography, cartographic symbolism, software agents

**Marc P. Armstrong**, Professor, Department of Geographical and Sustainability Sciences, and Associate Dean, College of Liberal Arts and Sciences, The University of Iowa, Iowa City, IA 52242

**The Internet of Things and Fast Data Streams:  
Prospects for Geospatial Data Science in Emerging Information Ecosystems**  
**Marc P. Armstrong, Shaowen Wang and Zhe Zhang**

**ABSTRACT:** This paper surveys the rapid development of the Internet of Things (IoT), the massive data streams that are only now beginning to be generated from it, and the resulting opportunities and challenges that these data streams bring to geographic information analysis. These challenges arise because streaming data volumes cannot be subjected to analysis using the standard repertoire of methods that have been designed to analyze static geospatial datasets. New approaches are needed, not to supplant, but to supplement, these existing tools. A particular focus is placed on the concept of data velocity (fast data) and its effects on sampling and inference. Innovative data ingestion strategies based on principles related to reservoir sampling and sketching are described. Dynamic temporal data flows present significant challenges to load balancing in distributed (e.g., cloud) parallel environments, even at exascale levels of performance. Further advances in the exploitation of data locality based on geographical concepts, as well as advanced processing methods based on edge and approximate computing, require further elucidation. Concepts are illustrated using a database compiled from a distributed sensor network of mobile radioactivity detectors.

**KEYWORDS:** Internet of things, fast data, streaming data

## **Introduction**

We are now at the cusp of a new paradigm of data acquisition and analysis characterized by the collection and transmission of massive streaming data volumes generated by the Internet of Things (NRC, 2013). This poses immense challenges to researchers interested in knowledge discovery processes driven by these massive streams of geospatial data. The purpose of this paper is to address the challenges encountered when geospatial innovations are applied to data streams to gain knowledge and insight from them. A particular focus is placed on the concept of data velocity and its effects on geospatial sampling and analysis.

## **The Emerging Information Ecosystem**

The widespread availability of inexpensive sensors with radios has created a digital ecosystem in which digital components are incorporated into a vast array of “things”. This new ecosystem has now placed data velocity in the spotlight and geographic information streamed from sensors has brought us into an era of fast data. Jarr (2015:20) provides an example in which 53 million electric meters stream usage information several times each second to monitor changes in demand and provide feedback to systems that reduce demand peaks. In what may be a burst of commercial hyperbole, a technology CEO has projected that there will be 500 billion connected devices by 2030. No matter the number, sensors are able to stream fine resolution data at high sampling rates and are

wirelessly communicating as edge components of the rapidly expanding ecosystem. While terrestrial communication coverage continues to improve, existing gaps are being filled by satellite constellations. For example, Eutelsat's 39 geostationary satellite constellation provides Ka-band broadband services to European and North American (de Selding, 2015) and other companies are planning global coverage (OneWeb, LeoSat, SpaceX) using large constellations (100s) of satellites.

## Data Velocity Effects on Scientific Research Paradigms and Geographic Analytics

Devices streaming massive amounts of data contribute to what is now called the fourth paradigm of scientific research: data-intensive discovery (Hey, Tansley and Tolle, 2009). New methods of spatial analysis are required to analyze fast data streams with evolving stationarity (Cao, et al. 2015). Figure 1 shows how fast data streams are ingested and subjected to fast space-time (FaST) analytics that require real-time (or near-real-time) response (IEEE, 1990:61). This requires high performance computing as well as new sampling methods that have lightweight computational intensity (Wang and Armstrong, 2009). The observations are then exported to a big data store where traditional methods of analysis are applied. A data value chain can help develop this conceptualization a bit more. Each data element is differentiated according to its “age” and whether it is an aggregated or singleton value (Figure 2). Fresh data has its greatest value when as an individual, and its value declines as it is replaced by new streaming observations. Observations regain value as they age and are aggregated with other values to yield synoptic insights.

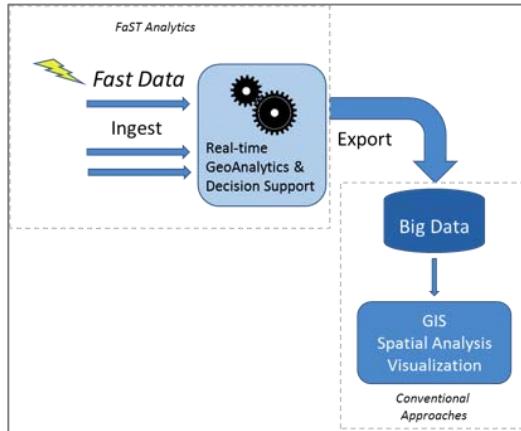


Figure 1. Fast space-time analytics (FaST) operate on fast data immediately upon ingestion and pass them to a data repository for conventional analyses (modified from Jarr, 2015).

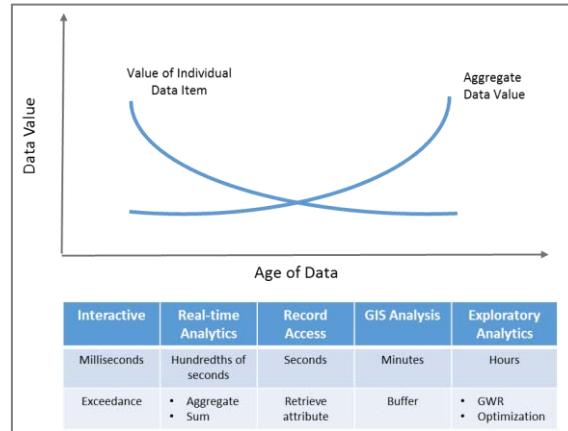


Figure 2. Changes in the form of a streaming data value chain (modified from Jarr, 2015).

## **Geo-streaming Analytics**

The following sections briefly review concepts in cognate areas of data reduction and computing that can be applied to improve the performance of geospatial analysis of fast data streams.

### ***Reservoir Sampling***

Load shedding, discarding observations, may be used to reduce stream-processing requirements, but since this can introduce bias, reservoir sampling (Vitter, 1985) is used to reduce the magnitude of load shedding bias; new data has greater salience.

### ***Windowing and Filtering***

In its most basic form, a sliding window can be applied to assess recently arriving data. In other cases, it may be useful to filter and smooth the stream.

### ***Sketching***

Sketching creates synopses and reduces the dimensionality of streaming data (Aggarwal and Yu, 2007; Cormode, et al. 2012). For each new element, a sketch determines set membership, cardinality (how many different types are in the stream?) and frequency. Ellis (2014: 331) states that sketch algorithms have three desirable features:

- Data updates are performed in constant time;
- Storage space is independent of stream size; and
- Queries are performed in linear time for the worst case.

### ***Compositing***

Composite sampling (Dorfman, 1943) takes a sample from each individual that will permit two tests to be performed; rather than testing n samples, they are composited into groups. Each group is tested for a signal and if there is none, then all members of that group are negative. If a signal is detected, each individual is tested to determine the sample(s) providing a positive signal.

### ***Sublinear Time Algorithms***

Algorithms that execute in linear time represent a “holy grail” of efficiency, though polynomial time is usually considered acceptable. Sublinear time algorithms make assumptions about data distributions to yield answers that are imprecise. Rubinfeld and Shapira (2011:1562) suggest that “there are many situations in which a fast approximate solution is more useful than a slower exact solution.”

### ***Approximate Computing***

Approximate computing asserts that processing must become energy efficient by changing the way that computation takes place: shortcuts reduce energy-consuming cycles (Kugler, 2015). According to Moreau, Sampson and Ceze (2015: 12, emphasis added) “Approximate computing is especially relevant in mobile environments...

[because] ... most applications in mobile devices are inherently approximable—including ... sensor data collection and summarization.”

### ***Hierarchical Decomposition and Sensor Organization***

Many strategies “divide and conquer” space in order to reduce search, and to efficiently store information (Morton, 1966; Samet, 1984; Wang and Armstrong, 2003). DaCosta (2013) describes a three-tiered structure in which end nodes (leaves) pass data (“chirps”) to propagator nodes that pass the data to integrators that perform higher levels of analysis and control. Chirps are lightweight and do not require the sensor to run the full IP communication stack. Propagator nodes eliminate redundancy and provide additional context information, such as location.

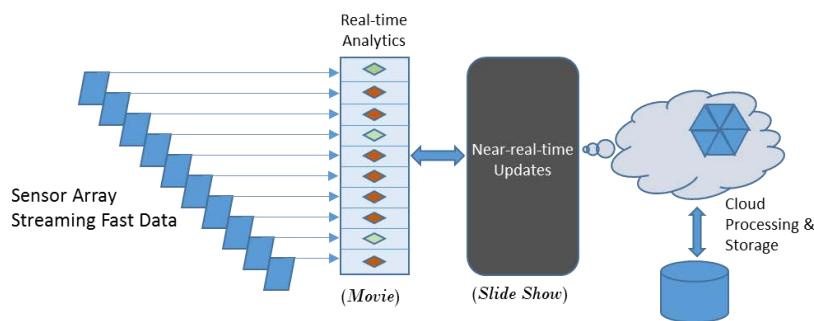


Figure 3. A staged stream processing architecture.

### ***Processing Requirements***

Performance is critical when processing fast data streams. As future computing systems are expected to have billion-way concurrency, areas of concern are related to load balancing, latency, and the development of a failure tolerant and locality aware programming model (Reed and Dongarra 2015). Data locality, which may reduce communication needs, is often based on an abstract partitioning of geographic space (Wang and Armstrong, 2003).

### ***Conceptual Illustration***

We illustrate concepts using an example cyberGIS workflow designed to detect anomalous radioactive sources in streaming data anonymized from Safecast (<https://blog.safecast.org>), a global volunteer-centered citizen science project. The data were streamed into and stored in a cyberGIS computing environment.

CyberGIS is defined as GIS based on advanced computing and cyberinfrastructure (Wang 2010; Wang 2016). In this case, a cyberGIS workflow is designed to generate an alarm if an anomalous radioactive source is found. Since ionizing radiation occurs naturally (cosmic rays), and is emitted by rocks, soil and building materials, non-zero radiation levels are detected without an anomalous source present. The challenge is to detect a source with a low signal-to-noise ratio, where the source is the signal and the background radiation is the ambient noise, in the presence of confounding factors (GPS accuracy, detector motion, shielding and weather conditions). In this study, two radiation source

types are considered, naturally occurring background and anomalous sources, which might come from dirty bombs, radioactive waste, or the precursors to such threats.

The cyberGIS workflow defines two types of radiation level estimates (Figure 4). The first is a background radiation level estimate (BRL), which refers to the radiation level without an anomalous radioactive source. This estimate is based on historical radiation data. The second estimate refers to the current measured radiation level (CRL) at a location. For example, CRL can be estimated by using the current measured radiation value averaged over the last n seconds. An alarm will trigger if there is a large difference ( $\Delta$ ) between the BRL and CRL, meaning an anomalous radioactive source is present near the current detector location. A k-nearest neighbor (KNN) algorithm combined with historical radiation data is used to estimate BRL. The algorithm is adopted from Keller et al. (1985) and applied to the data streams in which observations are analyzed to predict their classification, the interactive category in Figure 2.

The cyberGIS workflow developed in this research is required to resolve the computational intensity of KNN search using high-performance distributed computing (Wang and Armstrong 2003; Wang and Armstrong 2009). Without cyberGIS, this type of scientific workflow would not be possible (Wang 2017; Wright and Wang 2011). Furthermore, as geospatial data are accumulated in a streaming fashion, cyberGIS analytics must adapt to dynamic changes (e.g., background). A failure of the system to adapt would significantly reduce its ability to detect threats (Wang et al. 2014).

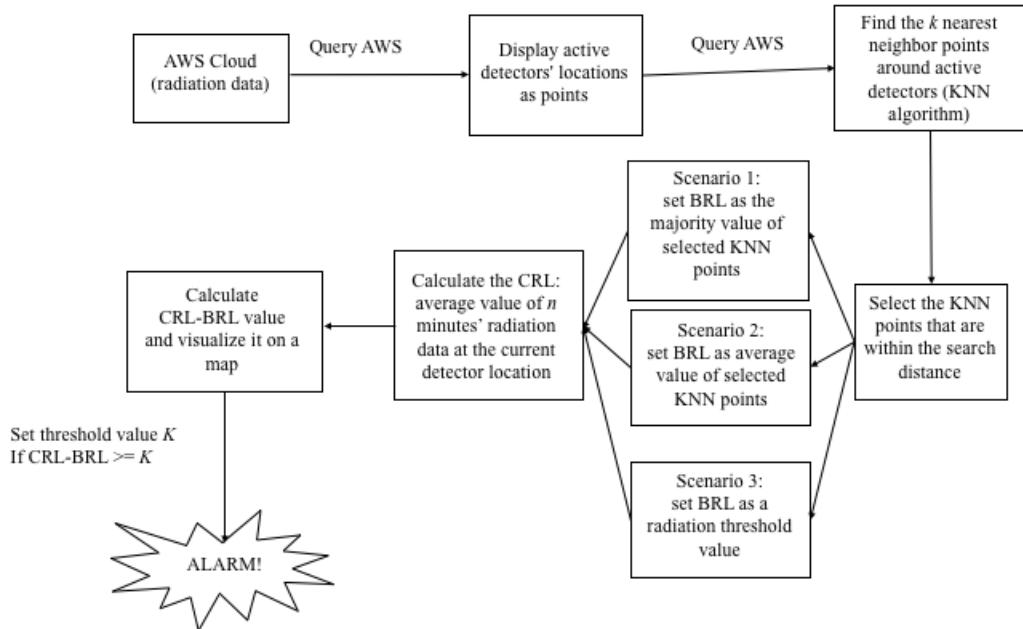


Figure 4. CyberGIS workflow.

## References

- Aggarwal, C. and Yu, P. (2007) A Survey of Synopsis Construction in Data Streams. In C. Aggarwal (ed.) *Data Streams: Models and Algorithms*. New York, NY: Springer, pp. 169-207.
- Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., and Soltani, K. (2015) A Scalable Framework for Spatiotemporal Analysis of Location-based Social Media Data. *Computers, Environment and Urban Systems*, 51, pp. 70-82.
- Cormode, G., Garofalakis, M., Haas, P., and Jermaine C. (2012) *Synposes for Massive Data: Samples, Histograms, Wavelets and Sketches*. Boston, MA: now Publishers.
- daCosta, F. (2013) *Rethinking the Internet of Things*. New York, NY: Apress/Springer.
- de Selding, P.B. (2015) Facebook-Eutelsat Internet Deal Leaves Industry Awaiting Encore. *SpaceNews*, 26, 36, pp. 1, 5.
- Dorfman, R. (1943) The Detection of Defective Members of Large Populations. *The Annals of Mathematical Statistics*, 14, 4, pp. 436-440.
- Ellis, B. (2014) *Real-Time Analytics*. Indianapolis, IN: John Wiley & Sons.
- Hey, T., Tansley, S., and Tolle, K. (eds.) (2009) *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research.
- IEEE. (1990) *IEEE Standard Glossary of Software Engineering Terminology*, Std 610.121990. New York, NY: The Institute of Electrical and Electronics Engineers.
- Jarr, S. (2015) *Fast Data and the New Enterprise Data Architecture*. Sebastopol, CA: O'Reilly Media, Inc.
- Keller J., Gray M., and Givens J. (1985) A Fuzzy K-Nearest Neighbor Algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, 15, 4, pp. 580–585.
- Kugler, L. (2015) Is ‘good enough’ Computing Good Enough? *Communications of the Association for Computing Machinery*, 58, 5, pp. 12-14.
- Moreau, T., Sampson, A., and Ceze, L. (2015) Approximate Computing: Making Mobile Systems More Efficient. *IEEE Pervasive Computing*, 14, 2, pp. 9 – 13.
- Morton, G.M. (1966) *A Computer Oriented Geodetic Data Base; and a New Technique in File Sequencing*. Ottawa, CA: International Business Machines, Ltd. <http://domino.research.ibm.com/library/cyberdig.nsf/0/0dabf9473b9c86d48525779800566a39?OpenDocument>.

NRC (National Research Council). 2013. *Frontiers in Massive Data Analysis*. Washington, DC: The National Academies Press.

Reed, D.A. and Dongarra, J. (2015) Exascale Computing and Big Data. *Communications of the Association for Computing Machinery*, 58, 7, pp. 56-68.

Rubinfeld, R. and Shapira, A. (2011) Sublinear Time Algorithms. *SIAM Journal of Discrete Mathematics*, 25, 4, pp. 1562-1588.

Samet, H. (1984) The Quadtree and Related Hierarchical Data Structures. *Computing Surveys*, 16, 2, pp. 187-260.

Vitter, J.S. (1985) Random Sampling with a Reservoir. *ACM Transactions on Mathematical Software*, 11, 1, pp. 37-57.

Wang, S. (2017) CyberGIS. In: *The International Encyclopedia of Geography: People, the Earth, Environment, and Technology* edited by D. Richardson, N. Castree, M. F. Goodchild, A. L. Kobayashi, W. Liu, and R. Marston, Wiley-Blackwell and the Association of American Geographers, DOI: 10.1002/9781118786352.wbieg0931.

Wang, S. (2016) CyberGIS and Spatial Data Science. *GeoJournal*, 81, 6, pp. 965-968.

Wang, S. (2010) A CyberGIS Framework for the Synthesis of Cyberinfrastructure, GIS, and Spatial Analysis. *Annals of the Association of American Geographers*, 100, 3, pp. 535-557.

Wang, S., and Armstrong, M. P. (2009) A Theoretical Approach to the Use of Cyberinfrastructure in Geographical Analysis. *International Journal of Geographical Information Science*, 23, 2, pp. 169-193.

Wang, S., and Armstrong, M. P. (2003) A Quadtree Approach to Domain Decomposition for Spatial Interpolation in Grid Computing Environments. *Parallel Computing*, 29, 10, pp. 1481-1504.

Wang, S., Hu, H., Lin, T., Liu, Y., Padmanabhan, A., and Soltani, K. (2014) CyberGIS for Data-Intensive Knowledge Discovery. *ACM SIGSPATIAL Newsletter*, 6, 2, pp. 26-33.

Wright, D. J. and Wang, S. (2011) The Emergence of Spatial Cyberinfrastructure. *Proceedings of the National Academy of Sciences*, 108, 14, pp. 5488-5491.

**Marc P. Armstrong**, Department of Geographical and Sustainability Sciences, The University of Iowa, Iowa City, IA 52241

**Shaowen Wang**, Department of Geography and Geographic information Science, University of Illinois, Urbana, IL, 61801

**Zhe Zhang**, Department of Geography and Geographic information Science, University of Illinois, Urbana, IL, 61801

# EWN-KDF: A Knowledge Discovery Framework to Understand the Energy Water Nexus

**Anne S. Berres, Rajasekar Karthik, Alexandre Sorokine, Philip J. Nugent, Melissa R. Allen, Ryan A. McManamay, Varun Chandola, Syed Mohammed Arshad Zaidi, Jibonananda Sanyal and Budhendra Bhaduri**

**ABSTRACT:** The Energy Water Nexus Knowledge Discovery Framework (EWN-KDF) is a web-based tool that provides easy and seamless access to wide variety of data from disparate and distributed datasets such as those containing physiographic, socioeconomic, and critical infrastructure information. It facilitates collaboration among analysts through shared workspaces and a growing toolbox of analytic tools and visualization options.

**KEYWORDS:** web-based analytics, web-based visualization, energy water nexus, integrated analysis, data fusion

## Motivation

Recent studies have highlighted the intense water demands associated with U.S. electricity production (Averyt et al., 2011; Cooley et al., 2011; EPRI, 2011) and the potential implications of increasing competition for water among the energy, agricultural, and residential/commercial sectors (EPRI2011). Such competition may be exacerbated by climate change and its impacts at local, regional, national, and global scales. Yet understanding the future consequences of climate change for the Energy-Water Nexus (EWN) is dependent upon understanding future regional trajectories for climate, population growth, land use, economic activity, and energy technologies and how they scale over space and time (Sovacool and Sovacool, 2009a, 2009b; Parish, 2012); as well as potential innovations in technology, adaptation, and resilience options that may be deployed on the U.S. landscape.

Analyzing the Energy Water Nexus involves addressing some of the greatest challenges in big data analysis: large volume of data, the speed at which new data is generated, and a great variety of data sources, formats, and time frames. These challenges make it increasingly difficult for analysts to keep an overview of available data, store it locally on their personal machines, and share it with colleagues.

Data fusion and integrated analysis are paramount to supporting analysts in deriving greater value from the whole, than they could gain from the individual components. Facilitating quick and easy access to datasets, data sharing, and enabling seamless integration of data analytics and visualization of relevant data within a single system is a vital step towards this goal. The EWN-KDF grants analysts access to integrated, user-guided, and exploratory analysis and knowledge discovery.

## System Architecture

The Energy Water Nexus Knowledge Discovery Framework (EWN-KDF) is a web-based tool that is designed to empower analysts to access and interact with data in their own browsers, analyze and visualize it, and share results with their collaborators. Its architecture separates the concerns of presentation, data access, and data storage. Additionally, the EWN-KDF utilizes several services and components built specifically for the requirements of the tool. The EWN-KDF backend is based on WSTAMP (Stewart et al., 2015), with an underlying database storing a large portion of the data.

A dedicated middleware component has been created to manage and delegate computationally intensive tasks requested by users to several other components that run the analysis. A multitude of data preparation functions is supported. For example, a user may request to aggregate a raster dataset to a predefined boundary or link comma separated data to attributed geospatial data to derive a new dataset. The requested tasks are inserted into a distributed task queue and are processed asynchronously as resources become available. This allows long running, computationally intensive tasks to run without impeding the user experience on the client. Additionally, tasks can be chained together to create analytic workflows with data in the system or data derived from a previous analytic workflow. User authentication is achieved using Globus Authentication (Foster, 2011; Allen, 2012), which makes it easy to connect new users to the system.

The frontend fuses and extends WSTAMP's time series analysis capabilities and D3.js-based charts (a javascript-based data manipulation library) with a more flexible spatial visualization that uses WebWorldWind (Bell et al., 2007), a 3D virtual globe API for HTML5 and JavaScript. Data, analysis, and visualization are organized in workspaces that analysts can share with their collaborators.

## Data

Some of the greatest challenges in big data analysis are the volume of data, the speed at which new data are generated, and the variety of data formats. To address the first two challenges, the EWN-KDF data catalog service is compliant with the OGC Catalog Services 3.0 standard and provides search and metadata filtering capabilities to users to enable analysts to search for the data they need for their applications within nearly 12,000 available datasets. Selected datasets can be added to a workspace, that manages the selected data, to any data operation, to analytics, and to visualization. Furthermore, the EWN-KDF provides workspaces with preloaded data for a number of use cases.

The datasets registered with the catalog are accessible either through web services (Web Map, Feature, and Coverage Services or WMS, WFS, and WCS respectively) or as Globus uploads (Foster et al., 2011; Allen et al., 2012). LandScan (Rose et al., 2014) and LandCast (McKee et al., 2015) are two such datasets. Finally, users can also add their own data from Globus endpoints or through the EWN-KDF tool's web interface.

To address the challenge of diversity of data formats, the EWN-KDF supports a variety of different formats and data models. Vector can be provided as shape or geojson files and linked with tabular data in XLS or CSV formats for rendering and analysis. Raster data can be ingested into the system through Web Coverage Service (LandScan/Landcast) or in netCDF format (for atmospheric datasets). Socio-economic data is linked to geographic units through FIPS code and are stored in the SQL database for efficient access and analysis. Most datasets require cleaning and reformatting to be compatible with the EWN-KDF. This process is facilitated through data fusion modules for a variety of common data formats. Once data is in the system, it is ready for analysis and visualization.

## Analytics

Data aggregation is a fundamental component of many data analysis workflows. Incoming data may be in an unsuitable format for display or analysis (e.g. rasterized climate data when the analysis operates on states), or it may be at a too coarse or too granular level for the intended use. To address these cases, aggregation functionalities are offered at different levels of granularity, from state-level to census block-level. In addition to cumulative aggregation, the aggregation module also offers descriptive statistics and correlation analysis. Other analytic tools include analysis of data completeness within the workspace, data repetition, and clustering.

One example of correlation analysis is a comparison of electricity customer consumption and temperature (Figure 1). Some areas have high electricity use when it is hot, others have high electricity consumption when it is cold, yet others do not have a clear correlation between temperature and electricity consumption. An analyst could draw conclusions about, e.g., building insulation differences in areas with similar climate.

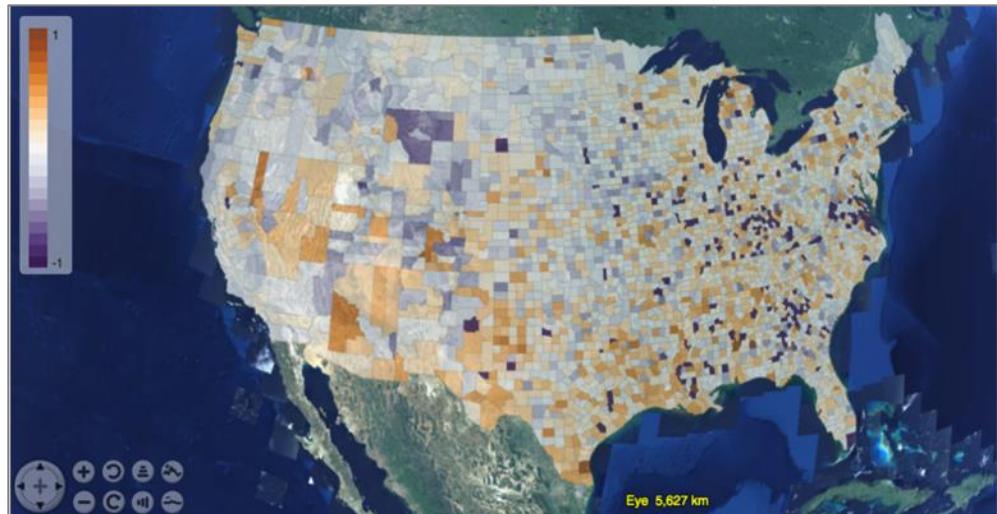


Figure 1: Correlation between customer electricity consumption and temperature. Orange denotes positive correlation whereas purple denotes inverse correlation. White denotes no correlation.

Another analytics tool of note is Dynamic Time Warping (DTW) (Müller, 2007), which examines similarities among temporal patterns by developing a non-linear “warped” dimension from which similarities or distances are measured. These distances are transformed into a distance matrix assessing similarities amongst entities (e.g., counties). DTW can include or ignore the effects of magnitude. The distance matrix is then used in hierarchical agglomerative clustering or any distance-based clustering procedure. Figure 2 shows space-time trends in thermoelectric water use.

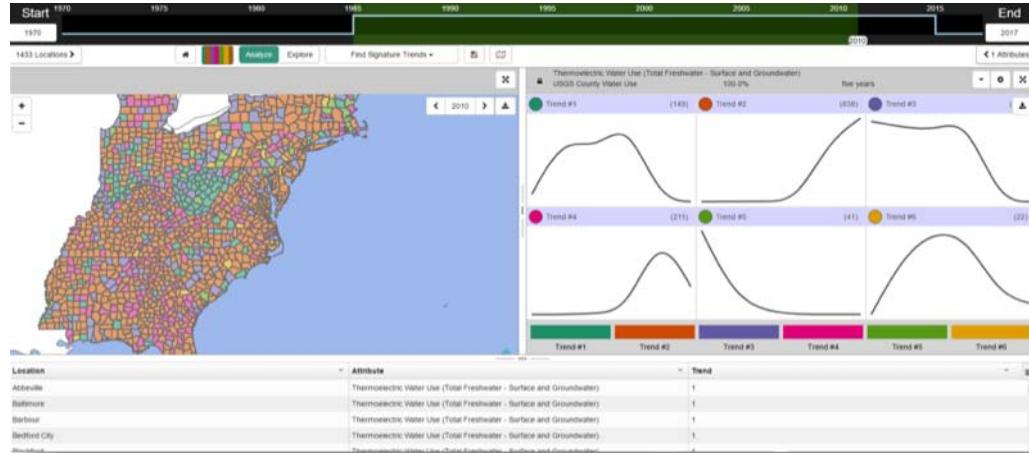


Figure 2: Dynamic Time Warping clusters time series data into groups with similar behavioral patterns allowing users to explore possible reasons for the observed groupings and to examine spatial clusters, trends, and anomalies that generate new hypotheses and guide scientific inquiry. Here we show the analysis with county-by-county thermoelectric water use.

## Visualization

The EWN-KDF visualization suite consists of two major visualization components: a map for spatial visualization, and charts for non-spatial representations. Spatial visualization is conducted with WebWorldWind, a flexible, javascript-based virtual globe library. For the EWN-KDF, WebWorldWind was augmented with color mapping capabilities in order to support choropleth maps. Analysts can choose between a variety of different color maps, including color scales (similar color, dark to light), divergent color maps (white for neutral, different colors on either end of the scale), and color maps with a gradient of different colors (Samsel, 2016; Harrower and Brewer, 2003). To account for varying distribution of values within different datasets, different transfer functions are offered to adapt colors: linear, square root, and logarithmic. In addition, analysts can adjust transparency as needed to see the underlying base map or other data layers. Time-varying data can be animated, or a single time step can be viewed individually.

Charts are generated with D3.js. This component provides a variety of different chart formats, including lines, scatterplots, histograms, box plots, and table heat maps.

## User Interface

The user interface consists of three main parts (Figure 3). On the left, the analyst has access to all data in their workspace. Each dataset can be added separately to the map.

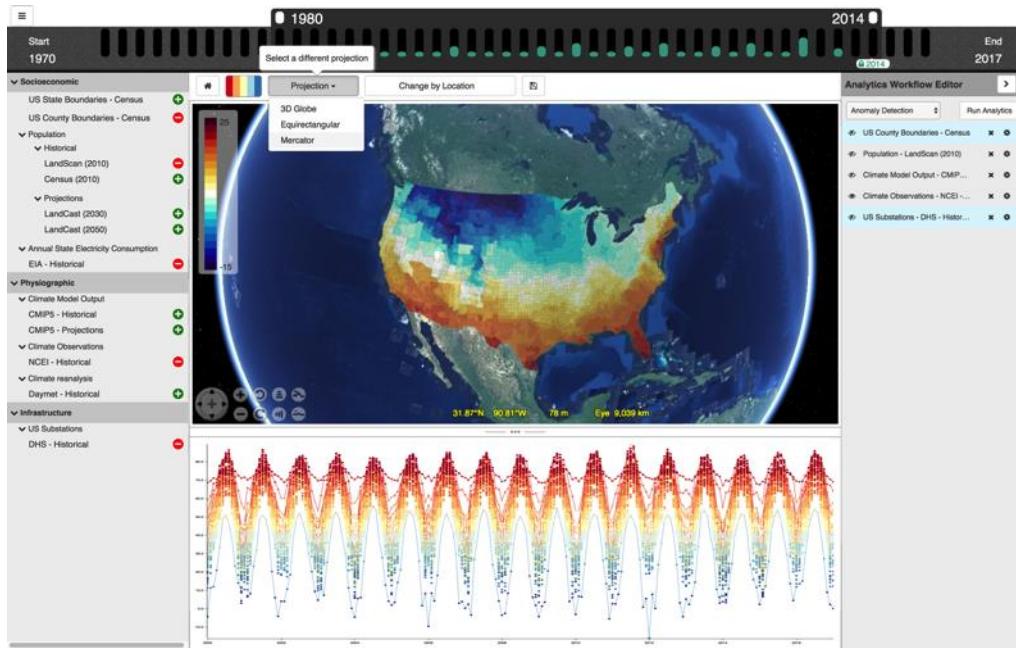


Figure 3: User interface for the EWN-KDF showing NCEI climate observations (temperature in Celsius).

On the right, the layer manager lets the analyst choose which layers should be visible, and which ones are on top. They can also pick one or several layers to run analytics on. In the center, visualization is provided with the two visualization components, map and charts. These two views are linked. Within each layer, data can be selected for further analysis by clicking on polygons in regions of interest, or on elements in the charts. Both views are updated to highlight the current selection. Different user interface elements provide access to available analytics, color map choices, and other functions such as switching between different map projections.

## Conclusion

The EWN-KDF supports analysts in their task of studying the energy-water nexus by lending them the tools to discover and load data from various sources into the system. They can gain new knowledge using the analytics tools provided for individual datasets, and for finding correlations between multiple datasets. Single time-step data can be displayed alongside time-variant data, and analysts can examine trends over time.

## Acknowledgements

This material is based upon work supported by the US Department of Energy, Office of Science, Office of Science, under contract number DE-AC05-00OR22725.

## References

- Allen, B., Bresnahan, J., Childers, L., Foster, I., Kandaswamy, G., Kettimuthu, R., ... & Tuecke, S. (2012). Software as a service for data scientists. *Communications of the ACM*, 55(2), 81-88.
- Averyt, K., Fisher, J., Huber-Lee, A., Lewis, A., Macknick, J., Madden, N., ... & Tellinghuisen, S. (2011). Freshwater use by US power plants: Electricity's thirst for a precious resource.
- Bell, D. G., Kuehnle, F., Maxwell, C., Kim, R., Kasraie, K., Gaskins, T., ... & Coughlan, J. (2007). NASA World Wind: Opensource GIS for mission operations. In *Aerospace Conference, 2007 IEEE* (pp. 1-9). IEEE.
- Chandola, V., Vatsavai, R. R., & Bhaduri, B. (2011, May). iGlobe: an interactive visualization and analysis framework for geospatial data. In *Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications* (p. 21). ACM.
- Cooley, H., Fulton, J., Gleick, P. H., Ross, N., & Luu, P. (2011). Water for energy: Future water needs for electricity in the intermountain West. *Pacific Institute, Oakland, USA*.
- EPRI (2011.) Water Use for Electricity Generation and Other Sectors: Recent Changes (1985-2005) and Future Projections (2005-2030). 1023676 (2011).
- Foster, I. (2011). Globus Online: Accelerating and democratizing science through cloud-based services. *IEEE Internet Computing*, 15(3), 70-73.
- Harrower, M., & Brewer, C. A. (2003). ColorBrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1), 27-37.
- McKee, J. J., Rose, A. N., Bright, E. A., Huynh, T., & Bhaduri, B. L. (2015). Locally adaptive, spatially explicit projection of US population for 2030 and 2050. *Proceedings of the National Academy of Sciences*, 112(5), 1344-1349.
- Müller, M. (2007). Dynamic time warping. *Information retrieval for music and motion*, 69-84.

Parish, E. S., Kodra, E., Steinhaeuser, K., & Ganguly, A. R. (2012). Estimating future global per capita water availability based on changes in climate and population. *Computers & Geosciences*, 42, 79-86.

Rose, A. N., & Bright, E. A. (2014). *The LandScan Global Population Distribution Project: current state of the art and prospective innovation*. Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States).

Samsel, F., Klaassen, S., Petersen, M., Turton, T. L., Abram, G., Rogers, D. H., & Ahrens, J. (2016). Interactive colormapping: Enabling multiple data range and detailed views of ocean salinity. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 700-709). ACM.

Sovacool, B. K., & Sovacool, K. E. (2009). Identifying future electricity–water tradeoffs in the United States. *Energy Policy*, 37(7), 2763-2773.

Sovacool, B. K., & Sovacool, K. E. (2009). Preventing national electricity–water crisis areas in the United States. *Colum. J. Envtl. L.*, 34, 333.

Stewart, R., Piburn, J., Sorokine, A., Myers, A., Moehl, J., & White, D. (2015). World Spatiotemporal Analytics and Mapping Project (wstamp): Discovering, Exploring, and Mapping Spatiotemporal Patterns across the World's Largest Open Source Data Sets. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2(4), 95.

**Anne S. Berres**, Postdoctoral Research Associate, Computational Science and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831

**Rajasekar Karthik**, Research Scientist, Computational Science and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831

**Alexandre Sorokine**, R&D Staff Member, Computational Science and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831

**Philip J. Nugent**, Research Scientist, Computational Science and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831

**Melissa R. Allen**, Research Scientist, Computational Science and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831

**Ryan A. McManamay**, Team Lead, Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831

**Varun Chandola**, Assistant Professor, Department of Computer Science and Engineering,

**Syed Mohammed Arshad Zaidi**, Ph.D. Student, Department of Computer Science and Engineering, University at Buffalo, Buffalo NY, 14260

**Jibonananda Sanyal**, Team Lead, Computational Science and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831

**Budhendra Bhaduri**, Group Lead, Computational Science and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831

# Analysis of the Adoption of Esri Story Map Apps

Aileen R. Buckley and Kevin A. Butler

**ABSTRACT:** Story maps are a sequence of maps that narrate a story. While these have existed in various forms for some time, they have become more popular in recent years due to use of the web to integrate maps with text, photos, and video and to provide functionality, such as panning and zooming, pop-ups, magnifying glasses, swipe tools, and time sliders—all of which may help users to better understand the story. Esri story maps were first introduced in 2012 through a single basic web application (app) which allowed users to create their own story maps. The suite of Esri story map apps has since grown, and their variety and functionality have increased. All of them share the advantages of users being able to create story maps without having to code, use special software, or store content on their own computers. In recent years, the number of story maps created has skyrocketed, arguably because of adoption of the Esri apps. In this study, we analyzed data for the creation of story maps using nine Esri apps from the date of their introduction to present. To better understand the patterns of adoption of these innovations, we performed a changepoint analysis to identify takeoff points in the uptake of the apps.

**KEYWORDS:** story map, Esri, web app, takeoff point, changepoint analysis

## Introduction

A broad definition of a story map is a series of maps presented in sequence to narrate a story. Historical examples of these include atlases, such as the *Atlas of Oregon*, which aims to “capture in maps the essential nature of Oregon (p. xvii, Loy, *et al.*, 2001). Other examples include printed page compilations such as many of the insets distributed with the National Geographic magazine or visitor maps distributed by the National Park Service (Figure 1). Myriad other examples also exist.

In recent years, story maps have come to mean something different—specifically web-based maps integrated with ancillary information, such as text, photos, and video, in apps that provide interactive functionality, such as pop-ups, swipe tools, and time sliders. While the map stories are basically linear in nature, their contents can also be explored in a nonlinear fashion by interacting with the map and using the various navigation controls. It can be argued that this new conception of story map is at least partly due to the introduction of Esri story maps apps (Figure 2). Since the introduction of this technology about five years ago (2012), Esri story maps have been adopted by many people, used for diverse purposes, and shared with large numbers of consumers.

The popularity of this new generation of story maps could also be due to the advantages that the Esri apps provide. People can quickly and easily create online story maps without having to know GIS or web programming. The apps incorporate builder functions that

## Exploring Theodore Roosevelt National Park

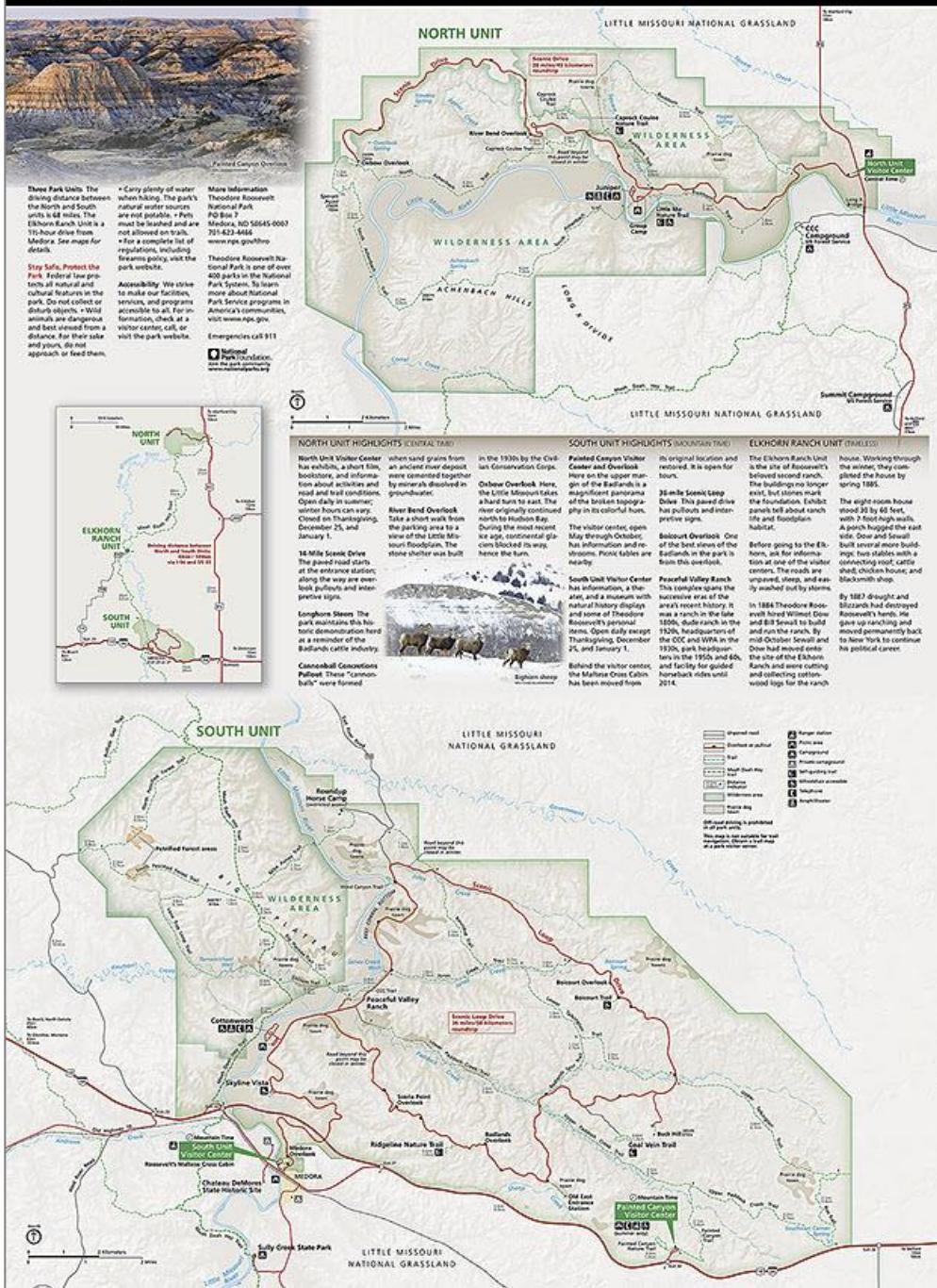


Figure 1: A story map in the form of a single printed page. Source: <https://www.nps.gov/thro/planyourvisit/maps.htm>.

step creators through the story map creation process. The apps are also open source so they can be downloaded and customized. Additional resources, such as tutorials, galleries of examples, and FAQs, are available to aid creators. Because the apps can be built and are stored in the cloud, creators do not have to store content on their own computers and their story maps can be shared with anyone. The apps can be viewed on a variety of devices, with screen sizes as small as smart phones and as large as personal computers.

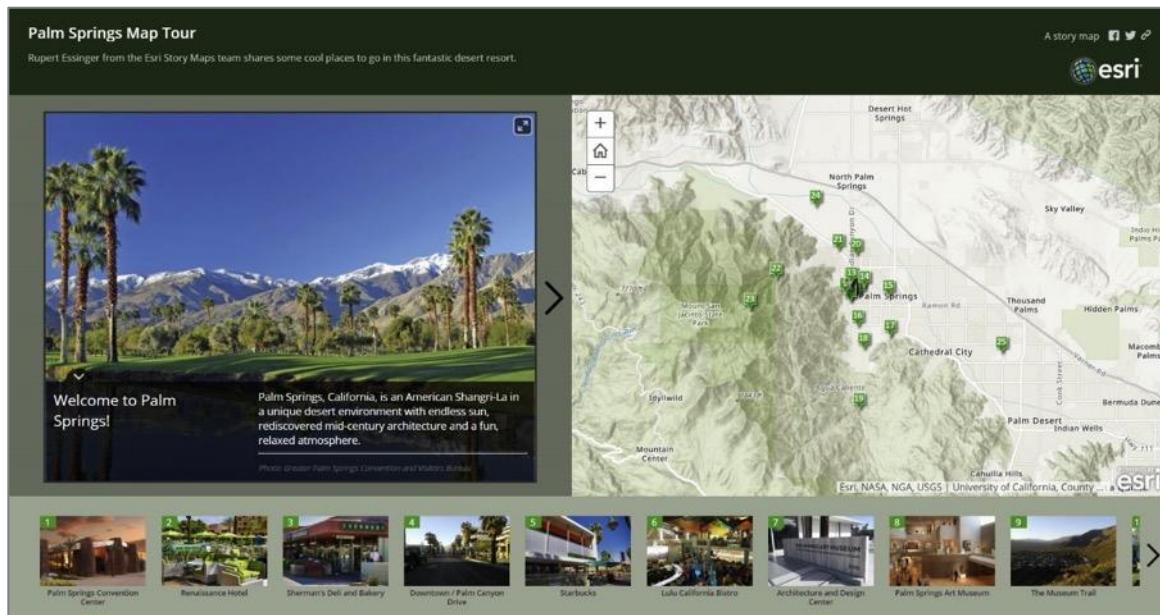


Figure 2: A story map in the form of an online app that includes a map and images. Source: [http://storymaps.esri.com/stories/demo/map\\_tour/?webmap=7190edafe7464cb19c1caf1360cd6ee5](http://storymaps.esri.com/stories/demo/map_tour/?webmap=7190edafe7464cb19c1caf1360cd6ee5).

While it is difficult to determine how many people have *used* these story maps, we do have information about how many story maps have been *created* using Esri's story map apps. With this information, we can learn about the patterns of uptake of these apps and predict adoption of future apps. These findings can be used in decisions relating to current and future product development and support.

### ***Esri story map apps***

There are currently eleven story map apps which can be categorized into seven groups depending on the primary intent of the story map (Table 1). There are two story map apps that are currently in beta release: Story Map Shortlist (beta) and Story Map Crowdsource (beta). These beta apps are stable and supported by Esri, and they can be used in production scenarios. However, new features will not be added to these apps, and only major bugs will be addressed.

Table 1. The uses and layouts of Esri story map apps.

<b>Presenting one map</b>	
<b>Story Map Basics<sup>SM</sup></b>	Present a map via a very simple user interface. Apart from the title bar and an optional legend, the map fills the screen. 
<b>Providing a rich, multimedia narrative</b>	
<b>Story Map Cascade<sup>SM</sup></b>	Create a visually and editorially engaging full-screen scrolling experience for the audience blending narrative text, maps, 3D scenes, images, videos, etc. Sections containing text and in-line media can be interspersed with "immersive" sections that fill the screen, including map animations and transition effects. 
<b>Story Map Journal<sup>SM</sup></b>	Create an in-depth narrative organized into sections presented in a scrolling side panel. As users scroll through the sections in the Map Journal, they see the content associated with each section, such as a map, 3D scene, image, video, etc. 
<b>Presenting a dynamic collection of crowdsourced photos</b>	
<b>Story Map Crowdsource BETA</b>	Publish and manage a crowdsourced story to which anyone can contribute photos with captions. Use Crowdsource to engage the audience and collect their photos and experiences, thoughts or memories on the subject of interest, all linked to a map. A vetting function lets the creator review and approve contributions. 
<b>Presenting a series of maps</b>	
<b>Story Map Series - Bulleted Layout</b>	Present a series of maps via numbered bullets, one map per bullet. This is a good choice when there are a large number of maps or locations to present. There is an optional description panel for presenting text and other content associated with each map. 
<b>Story Map Series - Side Accordion Layout</b>	Present a series of maps and accompanying text and other content for each map in an expandable panel. Clicking a title selects the map and expands the panel to reveal the text. 
<b>Story Map Series - Tabbed Layout</b>	Present a series of maps via a set of tabs. There is an optional description panel for presenting text and other content associated with each map. 

### **Presenting a curated set of places of interest**

<b>Story Map Shortlist BETA</b>	Present a large number of places organized into tabs based on themes, for example, restaurants, hotels, and attractions. As users navigate around the map, the tabs update to show them only the places within the current map extent.	
-------------------------------------	--	---

### **Presenting one map**

<b>Story Map Tour</b>	Present a set of photos or videos, along with captions, linked to an interactive map.	
-----------------------	---	---

### **Comparing two maps**

<b>Story Map Swipe</b>	Let users compare two separate web maps or two layers of a single map by sliding a swipe tool back and forth.	
<b>Story Map Spyglass</b>	Similar to Swipe but enables users to peer through one map to another with a spyglass tool.	

## **Methods**

For this study, we analyzed data from Esri for each of the nine story map apps both visually and statistically. A description of the data and the analyses follows.

### **Data**

The data from contains information about the date that the app was used to create a story map, as well as the type of app that was used, and the total number of apps that were created on that date. A snippet of the dataset is shown in Table 2.

Table 2. A portion of the story maps data table.

<b>date</b>	<b>app</b>	<b>total</b>
2/3/2018	Story Map Basic	7
2/2/2018	Story Map Basic	25
2/1/2018	Story Map Basic	29
1/31/2018	Story Map Basic	33
1/30/2018	Story Map Basic	37
1/29/2018	Story Map Basic	34
1/28/2018	Story Map Basic	16
1/27/2018	Story Map Basic	7
1/26/2018	Story Map Basic	45

The interpretation of a row is “On [date], users created [total] story maps of type [app]”. If a date is missing in the sequence, then no story map with that type of app was created

on that day. For this study, data for some of the apps were combined. The Story Map Series data includes counts for all flavors of the app, including Bulleted Layout, Side Accordion Layout, and Tabbed Layout. Data for the Story Map Swipe and Spyglass apps were also combined. The result is information about nine types of Esri story map apps: Basic, Cascade, Journal, Crowdsource, Series, Shortlist, Tour, and Swipe/Spyglass. The data were summarized by month as the daily volatility was confounding the analyses. Data was excluded from December 2017 because data for only a partial month were available; as a result, a false downward trend for all the story map apps was being reported.

## Visual Analysis

For this study, we started with visual exploration of the data, then we used changepoint analysis to statistically analyze the time series data. Visual analysis included the number of weeks that each app has been available (Figure 3). We also visualized the total usage by app type (Figure 4).

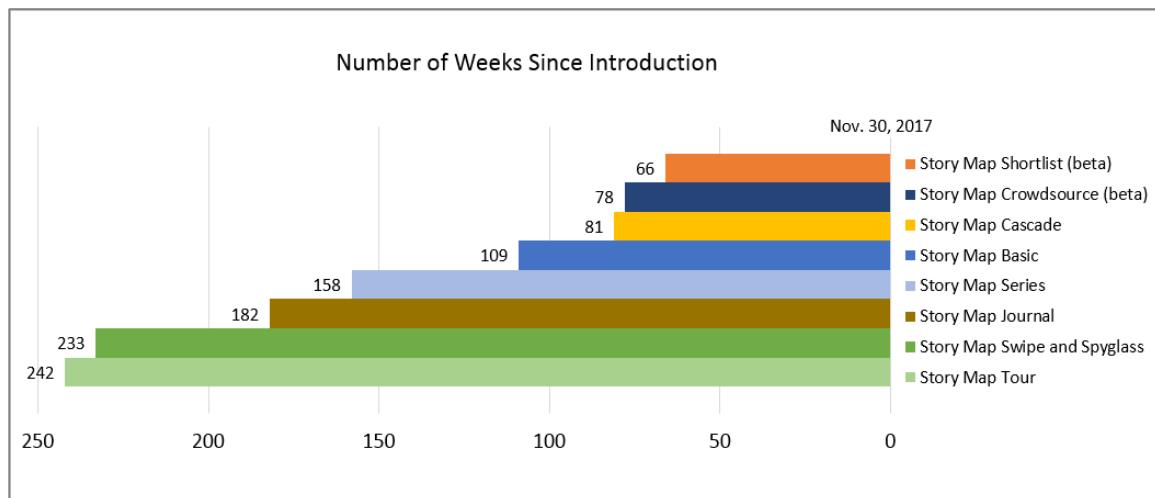


Figure 3. The number of weeks since each app was released (to November 30, 2017), that is, the total number of times that a story map was created using each type of app.

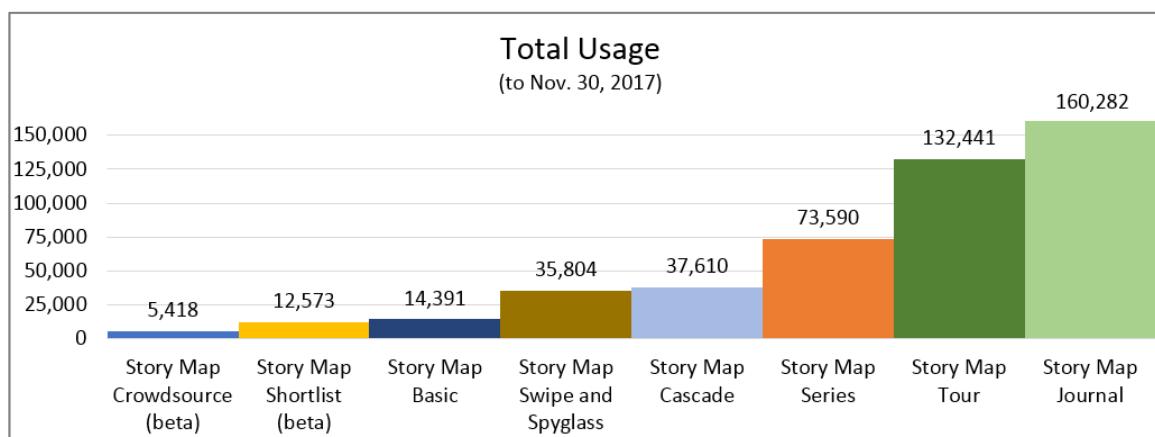


Figure 4. The number of times each app was used to create a story map.

Various aspects about usage were summarized for each type of app, including the average number of story maps created each day, the most apps created on any one day, and the number of days that an app was not used to create a story map (Figure 5).

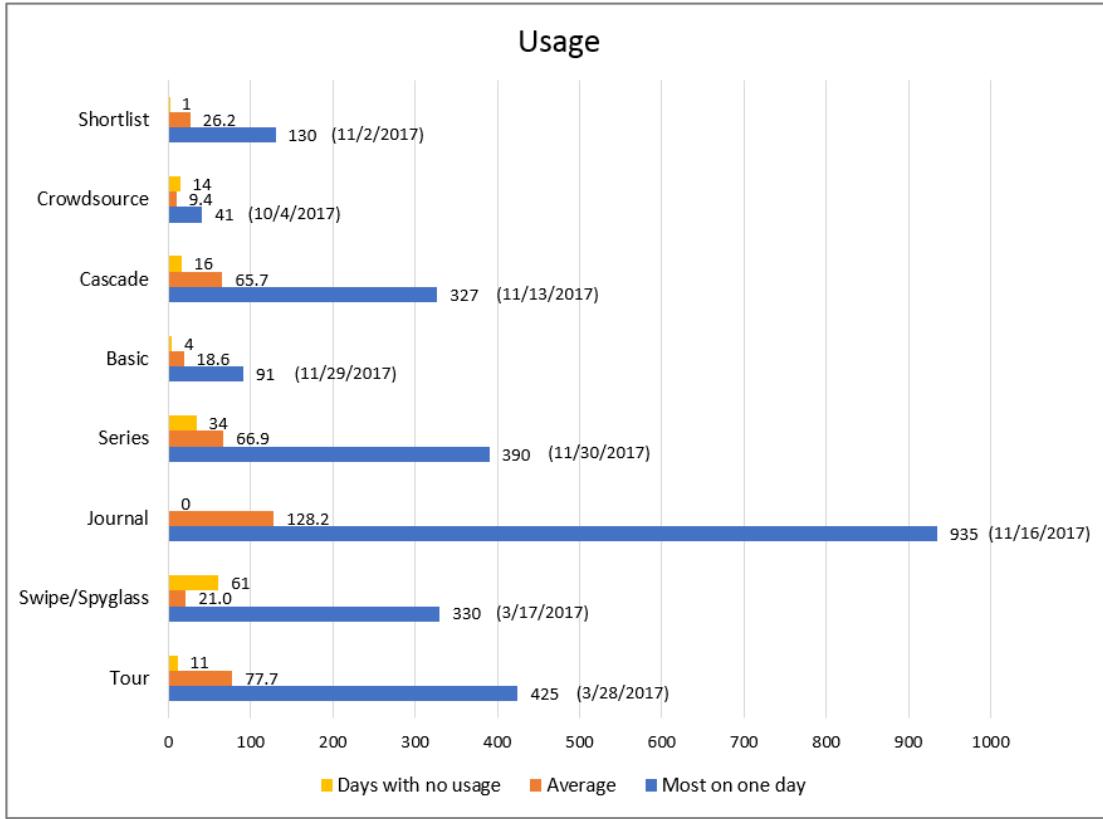


Figure 5. A variety of usage aspects were charted for each type of app.

Finally, we charted the monthly usage of each type of app (Figure 6).

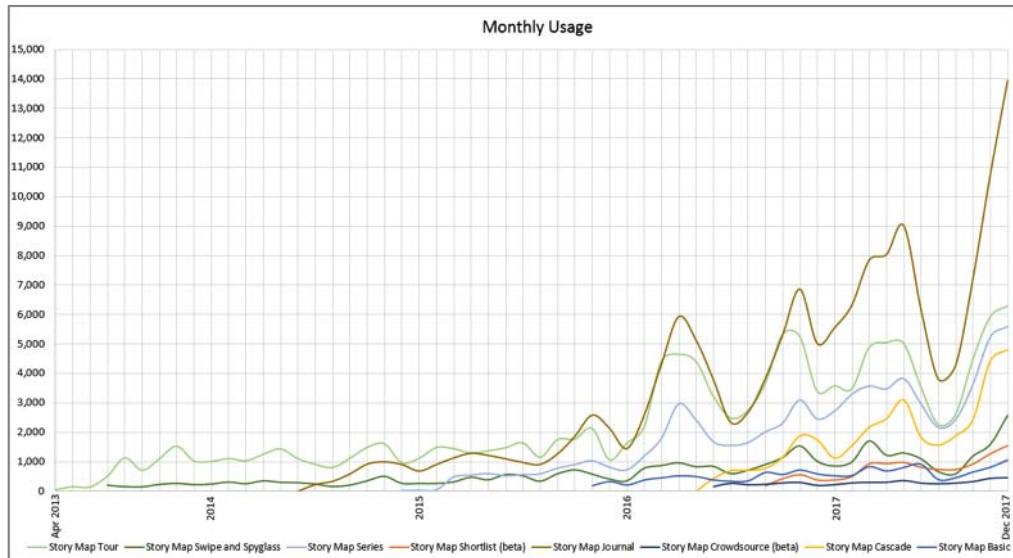


Figure 6. The monthly usage of each app to create story maps.

## **Changepoint Analysis**

Changepoint analysis identifies if and when a change has taken place in a time series (Figure 7 a). “In its simplest form, changepoint detection is the name given to the problem of estimating the point at which the statistical properties of a sequence of observations change” (Killick & Eckley, 2014). Knowing when a change has occurred can help to identify the cause, plan a response, and predict future change. This type of analysis has been used in a variety of application areas, such as climatology, bioinformatic applications, finance, oceanography, and medical imaging (see <http://www.changepoint.info/> for more on changepoint analysis and its uses). Changepoints also appear under a variety of synonyms in literature such as segmentation, structural breaks, break points, regime switching, and detecting disorder.

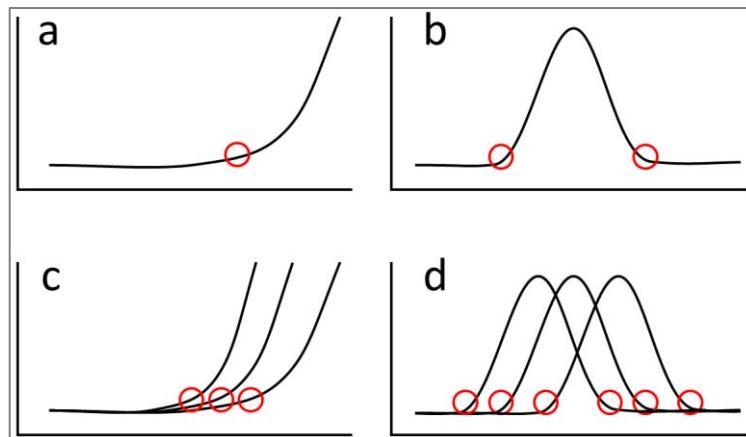


Figure 7. Changepoint analysis can be used to detect a single change in a single time series (a), multiple changes in a single time series (b), single changes in multiple time series (c), and multiple changes in multiple time series (d).

The first article about changepoints was published in 1954 by E.S. Page in relation to a quality control setting in manufacturing. This early analysis tested for a potential single changepoint for data from a common parametric distribution. Since then, changepoint analysis has developed to consider multiple changepoints (Figure 7 b, c, and d), different types of data, and other assumptions.

Our study involved a changepoint analysis to detect changes within a multiple time series (Figure 7 c), that is, the takeoff points of adoption for each of the story map apps in our dataset. For our study of the adoption of story map apps. The changepoints can be referred to as takeoff points, after which the usage of the apps increased significantly. To find these points, we used the changepoint package in R which has been developed to provide a choice of multiple changepoint search methods for use in conjunction with a given changepoint method (that is, changes in mean and/or variance using distributional or distribution-free assumptions).

We created trend lines (polynomial of order 3) to “smooth out” the data and see broader trends of adoption (Figure 8). The takeoff points were plotted on the trend lines to better see the relative timing of takeoff for each app.

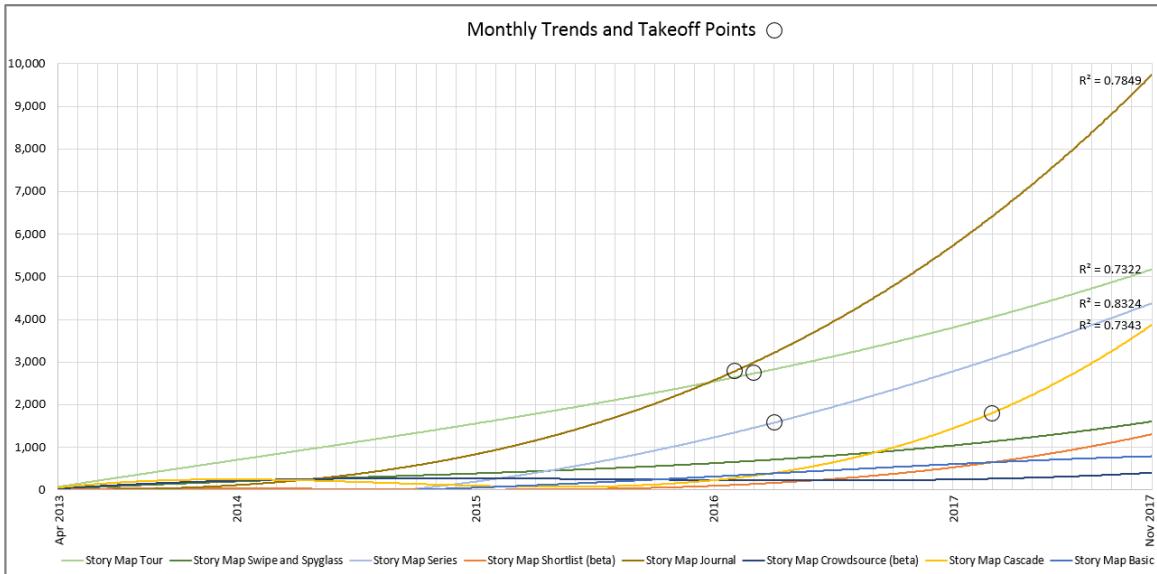


Figure 8. Monthly trends for all apps, and takeoff points for the four full-feature apps.

## Results

In order, Journal, Map Tour, Map Series, and Cascade have high adoption, and all four types appear to be trending up at a pretty good rate (Figure 8). Journal, Tour, Map Series and Cascade seem to be in a group (all four are full-feature apps), and Swipe/Spyglass, Shortlist, Story Map Basic, and Crowdsource are more in a second group (these are special purpose apps with limited functionality).

For the four full-featured apps, there is a clear, slow, introductory period and then a “takeoff point”. This fits with the observation by Trellis et al. (2003): “New products do not grow into maturity at a steady rate. Rather, their sales pattern is marked by a long introduction period when sales linger at low levels. At a certain point in time, the new product breaks into rapid growth, associated with a huge jump in sales. Academic literature and the business press refer to this point as the takeoff in sales. It is the point of transition between the introduction and growth stage of a new product. The time-to-takeoff is the duration of the introductory stage or the period from the introduction to the takeoff”. For our study, the time-to-takeoff was determined using changepoint detection.

Table 2. Results of Changepoint Analysis

<i>Story Map App</i>	<i>Date of introduction</i>	<i>Date of takeoff</i>	<i>Number of months from introduction to takeoff</i>
Tour	4/2013	3/2016	35
Journal	6/2014	2/2016	20
Series	12/2014	4/2016	16
Cascade	5/2016	3/2017	10

Sorting this table in order of when each app was released, it becomes apparent that each new generation of a full-service story map app is taking less time to reach takeoff. For example, the app most recently released app (Cascade) had a time to take off that was one-third the time of the original full-service app (Tour). However, from this analysis, we do not have the ability to say why these trends are happening. Nonetheless, that was not intent of this study (being primarily exploratory in nature).

## Conclusions

There is a number of directions in which this analysis could be extended. An obvious next step would be to identify specific questions related to why the times to takeoff differ. For example, is there an effect on the timing of takeoff points of events such as the Esri User Conference or the Esri Federal GIS Conference; the publication of an article, blog post, or Tweet; or other happenings? Also, there is a lot of volatility in the data, even when it is summarized monthly. What's causing this? What would an analysis of cumulative adoptions reveal? In this study, we barely scratched the surface of an analysis of this dataset; much can still be learned from these data.

## References

Killick, R., and Eckley, I. (2014). Changepoint: An R Package for Changepoint Analysis. *Journal of Statistical Software*, 58, 3, pp. 1-19.

Loy, W. G., Allan, S., Meacham, J. E., & Buckley, A. R. (2001). *Atlas of Oregon*. University of Oregon.

Tellis, G.J., Stremersch, S., and Yin, E. (2003). The International Takeoff of New Products: The Role of Economics, Culture, and Country Innovativeness. *Marketing Science*, 22, 2, pp. 188-208.

**Aileen R. Buckley**, Research Cartographer, Esri, Inc., 380 New York Street, Redlands, CA, 92373

**Kevin A. Butler**, Spatial Statistics Product Engineer, Esri, Inc., 380 New York Street, Redlands, CA, 92373

# A Crowdsourcing-Geocomputational Framework of Mobile Crowd Estimation

T. Edwin Chow

**ABSTRACT:** Population estimation typically involves counting a group of individuals at a specific place and time, with the assumption that the crowd does not move. Estimating people on the move, i.e. a mobile population, is indeed an intellectual challenge both in theory and in practice. The research objective is to develop a framework to better understand the dynamics of a mobile population by leveraging crowdsourced data. The proposed framework includes the following steps: 1) collect and analyze small area population movement (i.e. locations over time) from crowdsourced data, 2) reconstruct the mobile population dynamics during the rally (e.g. identifying the starting time, varying crowd density, walking velocities, and entry/departure gateways, etc.) using a Geographic Information System, 3) simulate the mobile population by using agent-based modeling. In the proposed framework, each individual is encoded as an agent with associated rules to define their crowd behaviors in walking, stopping and personal spacing. For simplicity, a distance and direction grid is computed from the origin (Victoria Park in Causeway Bay) to destination (Statue Square in Central), so that each agent moves towards the cell with next lower distance if the occupying capacity of the cell area has not exceeded the crowd density of a given region at that time. The intellectual merits and research findings sheds useful insights to improve mobile population estimation, and leverage alternative data source to support related scientific applications.

**KEYWORDS:** Small area population geography, human dynamics, population count, agent-based modeling, movement simulation

## Background

Crowd estimation typically involves counting a group of individuals at a specific place and time. Despite its long history, crowd estimation is often not an exact science – “The number of those who ate was about five thousand men, besides women and children” (Matthew 14:21). While the goal is simple, counting a crowd can be difficult in reality depending on the crowd size, weather, site, mobility, and event nature, etc. In fact, the uncertainty, and sometimes controversies, of crowd counting is often intensified by politics and public relations (Robertson and Farley, 2017).

In crowd estimation, it is often assumed that the mass crowd do not move (i.e. static population). Estimating people on the move, i.e. a mobile crowd such as those in protest, is indeed an intellectual challenge both in theory and in practice. In recent years, the need for estimating a mobile crowd in near real-time has become more acute due to practical needs for better estimation of event attendance, crowd management, emergency response, etc. (Watson and Yip, 2011).

Conventional crowd estimation methods typically involve field surveying. Pioneered by Jacobs (1967), the basic premise of this field method is to sample population density to estimate the crowd size by multiplying the average density with the total area that the crowd occupied. Seidler et al., (1976) improved this density approach by segregating the crowd into zones and sectors. This simple method works well for a static crowd but can be subjected to uncertainty for mobile crowd, because its dynamic nature would introduce high uncertainty in the sampling bias of crowd density over time and space.

Besides density-areal calculation, social scientists also used the head-counting approach. To estimate the mobile crowd attending a demonstration, the Hong Kong University Public Opinion Programme (HKUPOP) deployed a field crew to count the number of protestors at a specific location *A* along the rally route at a regular fixed time interval (e.g. count for 2 minutes and rest for 2 minutes). To compensate for the protestors who might have left before or joined after that location *A*, which would underestimate the crowd size, the HKUPOP team would follow up with a random telephone poll to the population at large to see if they participated the demonstration and had passed location *A* to adjust the crowd size (HKUPOP, 2017). To avoid the sampling bias of a random telephone bias, Yip et al. (2010) proposed a “double count and spot-checking” method to conduct headcount at two locations (*A* and *B*), where *A* is near the start and *B* is near the end. Besides head counting, the “double count and spot-checking” method also involves an in-situ survey at location *B* and asks if the participants had passed point *A* to adjust the overall count. These field methods are theoretically sound, efficient and quick, but the fieldwork can be labor-intensive, costly and subjected to human errors.

The third approach utilize various products of remote sensing to estimate population and associated parameters. A common approach in remote sensing is to extract a proxy of human settlement (e.g. count the number of houses) to estimate population through areal-interpolation or statistical modeling (Lo, 1986; 1995; Wu et al., 2005). Using nighttime imagery in remote sensing such as Defense Meteorological Satellite Program – Operational Linescan System (DMSP-OLS), early work also attempted to correlate artificial lighting with population density to estimate population (Sutton et al., 1997; Lo 2001). With emerging availability of fine-spatial-resolution remote sensing and light detection and ranging (lidar), it is possible to extract building footprints, building heights to differentiate dwelling types and refine the associated population density to estimate resident population (Silván-Cárdenas et al., 2010). While remote sensing has been very useful in extracting human settlement, most remote sensing applications associate the proxies of human settlement with the static nighttime population at their residence. LandScan is a global population distribution dataset that leverage dasymetric modeling to interpolate “ambient population” (i.e. daily average that accounts for daytime movement) based on multiple ancillary data, such as census, administrative boundary, land cover, slope, roads, urban areas, and fine-spatial-resolution imagery (Bhaduri et al., 2007). LandScan products, however, are limited in spatial resolution (global data at 30 arc seconds, approximately 1 km x 1 km near equator) and temporal resolution (daily) for mobile crowd estimation. Although LandScan USA dataset has finer resolutions at 90 m x 90 m in both daytime and nighttime, it remains limited due to the dynamic nature of mobile crowd estimation in both space and time.

## Methodology

In this paper, the research objective is to develop a framework to better estimate crowd size and model their movement. The proposed framework leveraged various crowdsourced data and geocomputation techniques. This research applied the crowdsourced-geocomputation (CG) approach to estimate the crowd attending the July 1<sup>st</sup> rally in Hong Kong to answer these questions:

- How does the crowd characteristics (e.g. density, velocity) vary over space and time during a rally event?
- How does the estimated crowd size and simulated movement of the crowdsourced-geocomputation framework sensitive to the model parameters?
- Is there any significant difference(s) of crowd size across the mobile crowd estimation approaches?

The proposed framework includes the following steps: 1) collect various crowdsourced data indicative of crowd characteristics, 2) analyze and reconstruct the mobile crowd characteristics during the rally (e.g. identifying the starting time, varying crowd density, walking velocities, and gateways of late arrival/early departure, etc.), 3) simulate the mobile crowd in an agent-based model based on the crowdsourced characteristics.

In this research, two types of crowdsourced data were collected – 1) low-altitude and ground photography of the crowd, and 2) volunteered personal trajectory during the rally. The first type of data would be identified and downloaded from an exhaustive search of keywords related to the July 1<sup>st</sup> rally from various search engines, public and social media. The pictures can either contain spatial reference (i.e. geotagged) or have been taken with identifiable landmark(s) along the rally route as the backdrop. On the other hand, this research also launched a crowdsourcing project titled “I go there I count” to recruit volunteers to gather personal trajectory during the rally. A website was created for volunteers to sign up and learn more about downloading and installing a mobile application to record personal trajectory (<https://chowte.wixsite.com/dynamicpop>). Several blog tutorials were created to educate volunteers about the research project and associated protocols in location privacy, data collection and sharing. The personal trajectories were crowdsourced to describe the movement of the mobile crowd during the rally.

The crowdsourced pictures would be geotagged by referring to its spatial reference or matched against the Google StreetView pictures along the route. From the low-altitude or ground photographs, geographic landmarks would be identified to digitize a bounding polygon of a crowd subset in GIS (Figure 1a). Each individual within the bounding polygon of the photography will be labeled and counted to derive the crowd density at a specific timestamp (Figure 1b). In particular, pictures about the head crowd (indicated by the rally banner in Figure 1) will be cross-referenced with the timeline documented from public media to infer and reconstruct the starting and ending time of the rally. Moreover, walking velocity could also be derived from crowd density based on Weidmann’s equation (Figure 2). Similar analytics were applied to all crowdsourced photographs to infer crowd density and velocity over time and space throughout the rally. On the other

hand, it was relatively straightforward to extract walking velocity and infer crowd density based on the volunteered personal trajectories. Moreover, in-situ videos were taken at the gateways of early departure and late arrival to derive the associated model parameters empirically.



Figure 1: Reconstruction of crowd density and event timeline based on crowdsourced photographs.

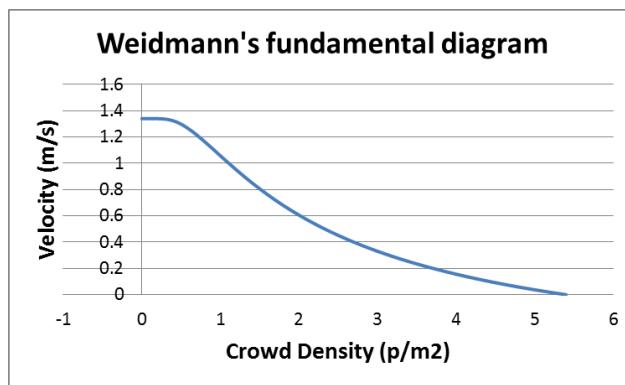


Figure 2: The relationship between crowd density and walking velocity (Bruno and Venuti, 2008).

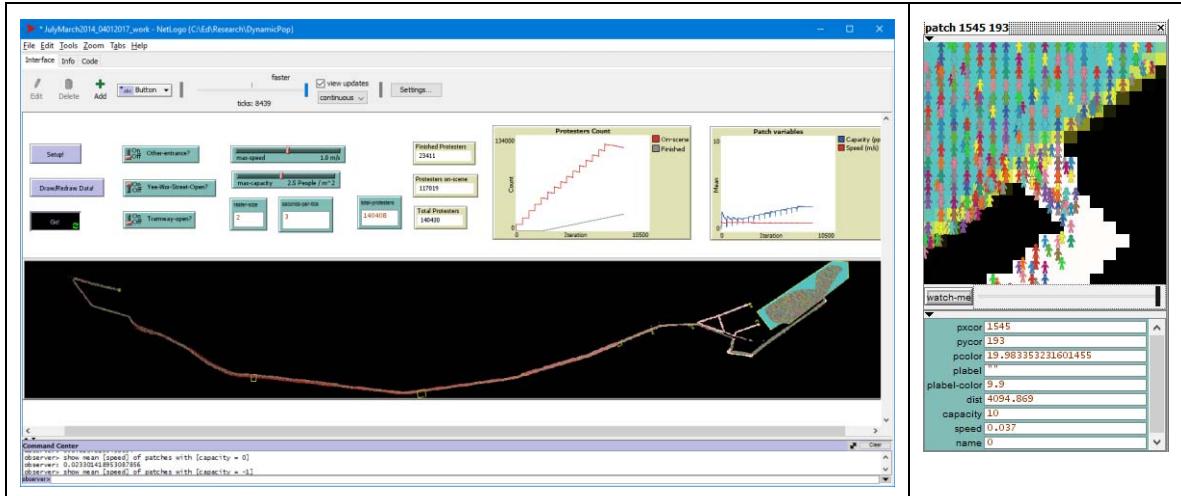


Figure 3: The interface of ABM simulating the mobile crowd of July 1st rally in HK.

The above crowdsourced data provided important model parameters, namely the crowd velocity and density, to simulate crowd behaviors during the rally. Using other ancillary geospatial data (e.g. road, sidewalk, etc.), an agent-based model (ABM) was implemented using NetLogo to simulate the mobile crowd (Figure 3). Within the ABM, each individual is encoded as an agent with associated rules to define their crowd behaviors in walking, stopping and personal spacing constrained by the crowd density. For simplicity, a distance and direction grid is computed from the origin (Victoria Park in Causeway Bay) to destination (Statue Square in Central), so that each agent moves towards the cell with next lower distance if the occupying capacity of the cell area has not exceeded the crowd density of a given region at that time.

## Results and Conclusions

This research developed and calibrated the proposed CG model with the data crowdsourced from 2017 July 1st rally in HK. To answer the research questions, crowd behaviors were observed throughout the simulated rally. It was found that the crowd density varied over space and are very sensitive to the associated rules implemented in ABM (Figure 4a). Moreover, crowd density decreased a little at the beginning as the head crowd started to depart the Victoria Park (VP). The simulated crowd density then increased steadily with fluctuation at the beginning of rally because the larger crowd slowly crawled through the narrow streets in HK and built up congestion. The crowd density eventually decreased slowly as congestion improved. Despite the average velocity remained relatively stable over time, it was observed that velocity varies over space. In general, walking velocity increases with increasing distance from VP as well as over time. There were more and more protestors arriving to the scene as fluxes of late arrivals joined the rally (Figure 4b). Nevertheless, the crowd on-scene decreased after the peak crowd appeared about two hours after the event started. Within the time constrain of rally event, the simulation had about 140,000 to 180,000 people within tested range of model parameters.

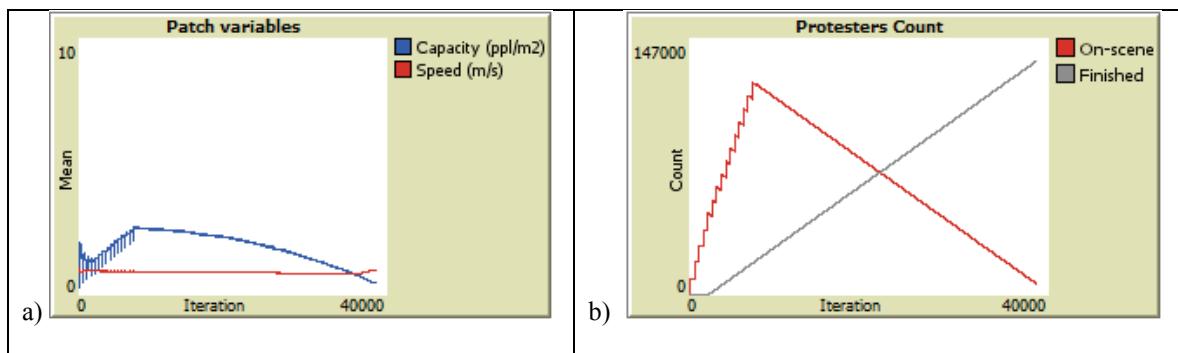


Figure 4: The plots illustrating model parameters of crowd density, velocity, and size over time.

By answering these research questions, the intellectual merits and research findings shed useful insights to 1) improve mobile crowd estimation in near real-time, and 2) leverage crowdsourcing to support related scientific applications. For example, knowing the

spatiotemporal distribution of human population during and after a hazardous event is vital to better manage emergency response, evacuation, search-and-rescue, and resource allocation.

## References

- Bhaduri, B., Bright, E., Coleman, P. and Urban, M. (2007) LandScan USA: A High Resolution Geospatial and Temporal Modeling Approach for Population Distribution and Dynamics. *GeoJournal*. 69, 1-2, pp. 103-117.
- Bruno, L. and Venuti, F. (2008) The pedestrian speed-density relation: modeling and application. Proceedings of 3rd International Conference Footbridge 2008. [http://staff.polito.it/luca.bruno/2008\\_footbridge\\_bruno\\_venuti.pdf](http://staff.polito.it/luca.bruno/2008_footbridge_bruno_venuti.pdf)
- HKUPOP (2017) July 1 Rally.  
<https://www.hkupop.hku.hk/english/features/july1/headcount/2017/index.html> Last visited 1/30/2018.
- Jacobs, H. (1967) To count a crowd. *Columbia Journalism Review* 6, pp. 36-40.
- Lo, C.P. (1986) Accuracy of population estimation from medium-scale aerial photography. *Photogrammetric Engineering and Remote Sensing*. 52, pp. 1859-1869.
- Lo, C.P. (1995) Automated population and dwelling unit estimation from high-resolution satellite images: a GIS approach. *International Journal of Remote Sensing*. 16, pp. 17-34.
- Lo, C.P. (2001) Modeling the population of China using DMSP operational linescan system nighttime data. *Photogrammetric Engineering & Remote Sensing*. 67, pp. 1037-1047.
- Robertson, L. and Farley, R. (2017) The facts on crowd size.  
<https://www.factcheck.org/2017/01/the-facts-on-crowd-size/> last visited 1/28/2018.
- Silván-Cárdenas, J.L., Wang, L., Rogerson, P., Wu, C., Feng, T. and Kamphaus, B.D. (2010) Assessing fine-spatial-resolution remote sensing for small-area population estimation. *International Journal of Remote Sensing*. 31, 21, pp. 5605-5634.
- Sutton, P.C. (2003) Estimation of human population parameters using nighttime satellite imagery. *Remotely Sensed Cities*, ed. Mesev, V. London: Taylor & Francis, pp. 301-334.
- Sutton, P.C., Roberts, C., Elvidge, C and Meij, H. (1997) A comparison of nighttime satellite imagery and population density for the continental united states. *Photogrammetric Engineering and Remote Sensing*. 63, 11, pp. 1303-1313.

Watson, R. and Yip, P. (2011) How many were there when it mattered? Estimating the sizes of crowds. *Significance*. 8, 3, pp. 104-107.

Wu, S., Qiu, X. and Wang, L. (2005) Population estimation methods in GIS and remote sensing: a review. *GIScience and Remote Sensing*. 42, pp. 80-96.

Yip, P.S.F., Watson, R., Chan, K. S., Lau, E. H. Y., Chen, F., Xu, Y., Xi, L., Cheung, D.Y.T., Ip, B., Y.T. and Liu, D. (2010) Estimation of the number of people in a demonstration. *Australian & New Zealand Journal of Statistics*. 52, 1, pp. 17-26.

**T. Edwin Chow**, Associate Professor, Department of Geography, Texas State University, San Marcos, TX 78666

# **Understanding of Intra-city Electricity Consumption Patterns Through Settlement Characterization**

**Pranab K. Roy Chowdhury, Jeanette Weaver, Eric Weber, Dalton Lunga, St. Thomas LeDoux, Amy Rose and Budhendra L. Bhaduri**

**ABSTRACT:** Urban areas consume around 75% of global primary energy, which is only expected to steadily increase in the near future due to population growth and unabated urbanization. Increasing urban energy consumption poses a serious threat to urban and environmental sustainability as well as energy security. However, scientific studies aimed at addressing these consequences of urbanization are severely hampered due to the lack of energy data at urban scales. The aggregated national or regional level electricity consumption data fail to capture the granularity required for urban scale analysis. This problem is more exacerbated in developing and under-developed countries, where such datasets are virtually non-existent. As per the current projections, more than half of the new global population growth will occur in these data starved regions of the world. Hence, to fill the data void, research methods for monitoring and understanding urban energy utilization patterns is urgently required. In this work, we adopt a data-driven approach to understand local level electricity consumption patterns in urban areas. We first characterize urban settlements into different typologies based on their formality in Ndola, Zambia; Sana'a, Yemen; and Johannesburg, South Africa. Using nighttime lights data as a surrogate for electricity consumption, we show a significant association between the derived settlement typed and nighttime lights emission. This study presents a simple yet scalable solution to fill the data scarcity and help in understanding intra-city electricity consumption patterns. Our approach could be helpful for urban planners and policymakers to gather insights on urban energy consumption and aid in sustainable city planning.

**KEYWORDS:** Settlement Characterization; Cities; Urban Electricity Consumption; VIIRS Nighttime lights; Developing Countries

**Pranab K. Roy Chowdhury**, Ph.D. Candidate, Bredesen Center for Interdiscip

**Jeanette Weaver**, Urban Dynamics Institute & Geographic Information Science and Technology Group, Oak Ridge National Laboratory, Oak Ridge, TN 37831

**Eric Weber**, Urban Dynamics Institute & Geographic Information Science and Technology Group, Oak Ridge National Laboratory, Oak Ridge, TN 37831

**Dalton Lunga**, Urban Dynamics Institute & Geographic Information Science and Technology Group, Oak Ridge National Laboratory, Oak Ridge, TN 37831

**St. Thomas Le Doux**, Urban Dynamics Institute & Geographic Information Science and Technology Group, Oak Ridge National Laboratory, Oak Ridge, TN 37831

**Amy Rose**, Urban Dynamics Institute & Geographic Information Science and Technology Group, Oak Ridge National Laboratory, Oak Ridge, TN 37831

**Budhendra L. Bhaduri**, Urban Dynamics Institute & Geographic Information Science and Technology Group, Oak Ridge National Laboratory, Oak Ridge, TN 37831

# Unmanned Aerial Vehicle Logistics Modeling and Performance: A Demonstration of Integrative Data Science

Kevin M. Curtin

**ABSTRACT:** Some of the most profound advances in science occur when the intellectual resources from one discipline are brought to bear on the unresolved problems of another discipline. There is emerging recognition of the potential for data science to act as the catalyst for this interaction. The research presented here is intended as both a demonstration of potential of integrative data science more broadly, and as a contribution to the research areas of logistics and unmanned and autonomous vehicle operations. This research presents an integration of the spatial analytic and visualization techniques of Geographic Information Science, with the modeling and solution procedures for highly combinatorially complex problems from Location Science and Operations Research. The context for the problems solved is the operation of unmanned aerial vehicle platforms, and the goals are to both broaden the understanding of these vehicles in logistics operations, and to be able to discern the mix of platform types that perform most efficiently.

**KEYWORDS:** unmanned aerial vehicles, autonomous logistics, location science, integrative data science

## Integrating GIScience and Location Science

Some of the most profound advances in science occur when intellectual resources from one discipline are brought to bear on the unresolved problems of another. Geographers have long recognized that some of the most dramatic mixing of ideas, objects, or cultures happens at boundaries. This is as true of the boundaries between disciplines as it is of the boundaries between nations. The advance of data science as a discipline is, in some measure, a recognition that it is the commonalities among scientific approaches, and the places where disciplines can come together that offer the most fruitful ground for basic scientific advance.

This research describes one effort to exploit the opportunity for integrative data science. Two areas of research are examined for their integrative potential, Location Science (a sub-discipline of geography with extensive ties to Operations Research (OR)) and GIScience (GISci). Both research areas address problems that can be highly data intensive: GISci has long addressed problems with massive datasets (e.g. global remote sensing data, detailed census enumerations), and the problems addressed in Location Science are often highly combinatorially complex, requiring the management of extraordinarily large datasets in the solution process. The integration of these research areas is examined in the context of unmanned aerial vehicles (UAV) – where the data demands and modeling structures are still in their infancy. More specifically, this research demonstrates how spatial analytic techniques can be employed to develop spatially aware scenarios, how those scenarios can be used as the basis for solving UAV

logistics problems, and how the data resulting from this integration can be used to understand the performance of UAVs in a logistics operation.

## **Background**

The background for this research lies primarily in two areas: the ties between the Operations Research and Geography, and the data and modeling needs for UAVs, particularly in the context of logistics.

### ***Location Science: Integrating GIScience and OR***

Although perhaps an oversimplification, the genesis of the field of Location Science lay in the realization 1) among OR practitioners that a significant subset of their problems are fundamentally spatial in nature, and 2) among geographers that many of their spatial problems require optimization techniques in order to generate the best solutions. This mutual recognition of the benefits of collaboration, and a dependency on each other's expertise is the foundation of Location Science (Hale & Moberg, 2003). However, it has been shown that, while some GIS platforms offer the means to model and solve location science problems, the methods used to provide these solutions are generally based on heuristics that often produce sub-optimal results (Curtin, Voicu, Rice, & Stefanidis, 2014). In contrast these same problems can be solved optimally if the spatial data is integrated with OR optimal solution procedures (Curtin, Hayslett-McCall, & Qiu, 2010).

It may be that the key to further advances in the integration of GISci and Location Science lies in identifying the data structures and models on which both depend and where the best methods from each research area can be applied. It is possible that this search for integration will lead to the ability to address research questions that neither discipline could sufficiently solve in isolation.

### ***Optimization Models for Unmanned Aerial Vehicles***

Transportation researchers recognize that technological advances in transport modes can revolutionize travel, and in turn have profound influence on cultural and economic conditions. Over centuries, advances in shipping, rail transport, air travel, and personal automobiles have been the stimuli for these transportation revolutions. It is possible – although as yet undetermined – that the advent of UAVs may contribute to the next revolution in transportation. If there is even a small chance that this is the case, the potential for positive change or for catastrophic negative consequences demands that the scientific community put significant effort into understanding these changes.

The development of UAV technology is happening at a rapid pace. There is increasing research interest in the operation of UAVs including their navigation abilities, measures to increase safety, and their impact on transportation planning. Of particular interest here is the modeling of optimal logistics behavior; that is, how can unmanned vehicles be directed to most efficiently deliver materiel given constraints on time, cost, and access. This research seeks to contribute by demonstrating how the integration of methods can lead to greater understanding of the performance of UAVs in logistics operations.

## Methods: ALOFT – the Autonomous Logistics Optimization Family of Tools

This research effort has produced an integrated set of models and tools to address unmanned and autonomous logistics. The data needs and scenario development, modeling and solution procedures, and methods for visualizing model outputs have been combined into a testbed environment to analyze logistics operations with a focus on the unmanned and autonomous elements of those systems.

### *Scenario Development Process*

To address the autonomous platform mix problem, a comprehensive data collection process is critical. This effort must include platform data collection, knowledge of the facilities from which the platforms will operate and any delivery locations, and data regarding the materiel to be delivered.

The set of logistics scenarios should be realistic approximations of the environments under which the platforms for delivery will need to operate. When possible the pedigree of the scenario should be checked against the expert knowledge of those working in the system; e.g. logistics experts who plan for increased integration of UAVs in their operations. Figure 1 illustrates a scenario where military assets are employed for a rescue and recovery mission with both sea based and land based facilities.

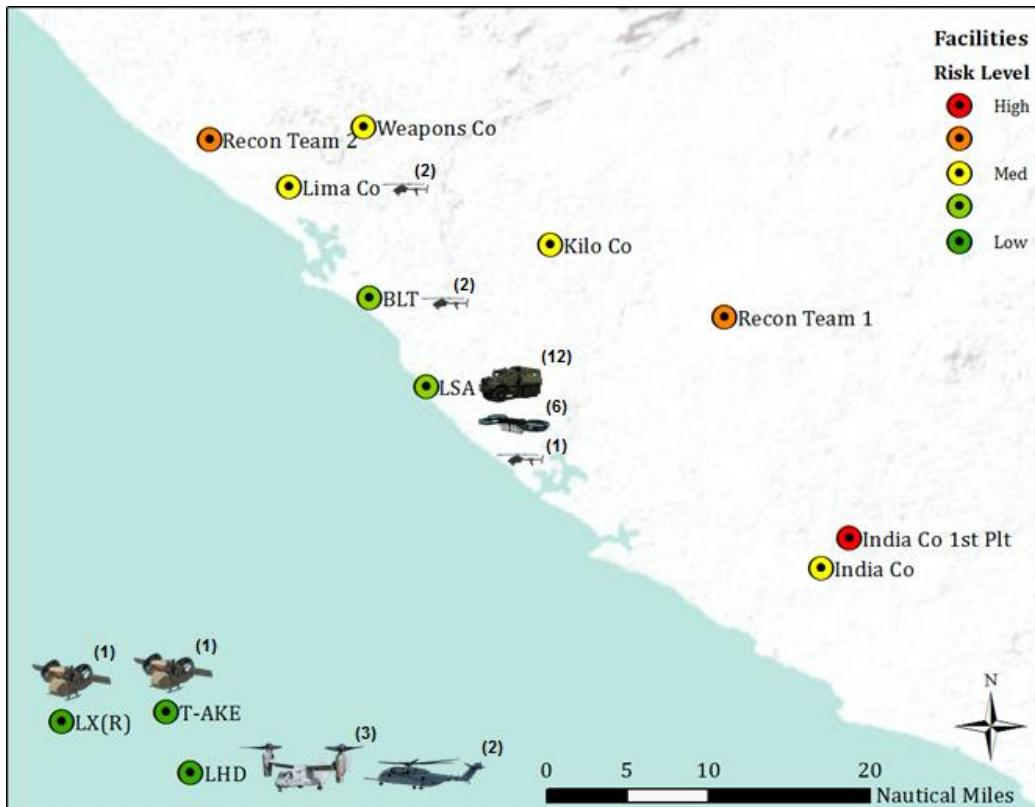


Figure 1: An example logistics scenario with unmanned aerial vehicle platforms assigned to some facilities. This scenario represents a relief and rescue mission in coastal West Africa.

## **Optimization Modeling and Solution**

While the focus of this research article is the integration of GISci and Location Science methods, a significant element is the set of optimization models that permit solutions to problem instances generated in the scenario development process. A detailed formulation will appear in a companion article but is summarized here. This formulation includes three objectives (minimum cost, minimum risk, minimum unmet demand) and a composite objective. Constraints include platform range limits, platform capacity limits, and restrictions on platform access to facilities.

## **Integration and Generation of a Testbed Environment**

Tools for scenario development, methods for executing optimal solution procedures, and means of visualizing and querying the results of those models are combined into the ALOFT testbed environment. For this research ArcGIS is used for spatial scenario development, supporting database management, and visualization tools. Gurobi is used for linear programming solution procedures. Python libraries provide the bridge between these sets of tools.

## **Results**

The results generated by integrating GIScience and Location Science in the ALOFT testbed environment fall into three categories. First, the testbed environment itself represents an advance, given the novel integration of spatial analytic methods, optimization modeling and solution procedures, and visualization methods. Figure 2 shows an example of the output from this integration. The scenario results, including the flights of the individual unmanned aerial systems can be viewed and queried providing information about the facilities visited, the amount and type of materiel delivered, and the cost and distance parameters of the flight.

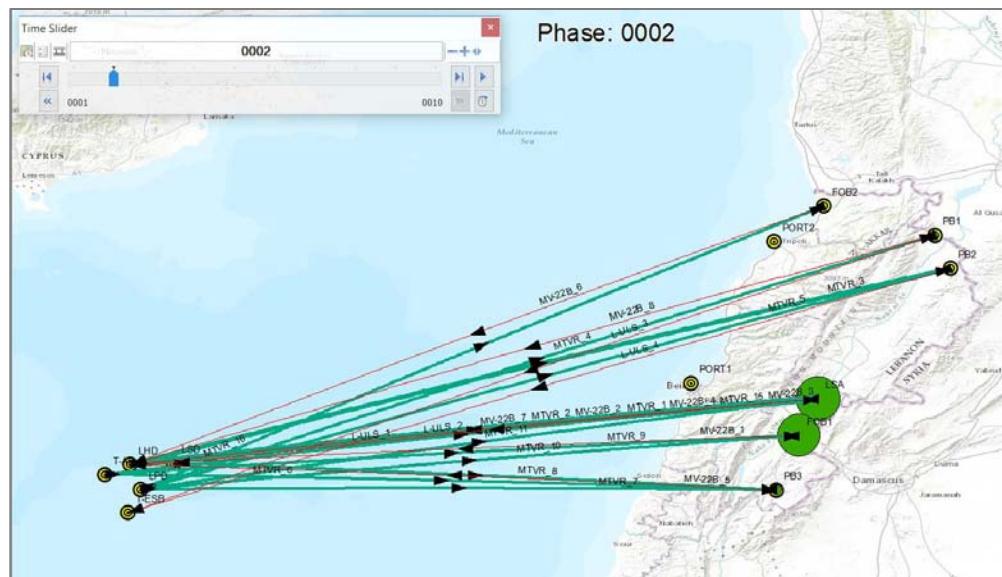


Figure 2: Detail of the results of a logistics operation with unmanned aerial vehicle platforms. Facility, platform, and delivery parameters can be visualized and queried.

Second, statistical results regarding platform performance, and comparisons across platform mixes are provided from the scenarios and models. As one element of these results, Figure 3 shows Pareto tradeoff curves that can indicate to decision-makers how platform mixes perform with regard to a series of objective function values, and how this performance is associated with the cost to acquire those platforms. Consider that not only are the optimization problems being solved highly combinatorially complex, but when many variant problem instances are solved across many logistics scenarios, the resulting data for analysis is multiplied, requiring more sophisticated data scientific storage and manipulation processes, and the increased involvement of statistical analysis.

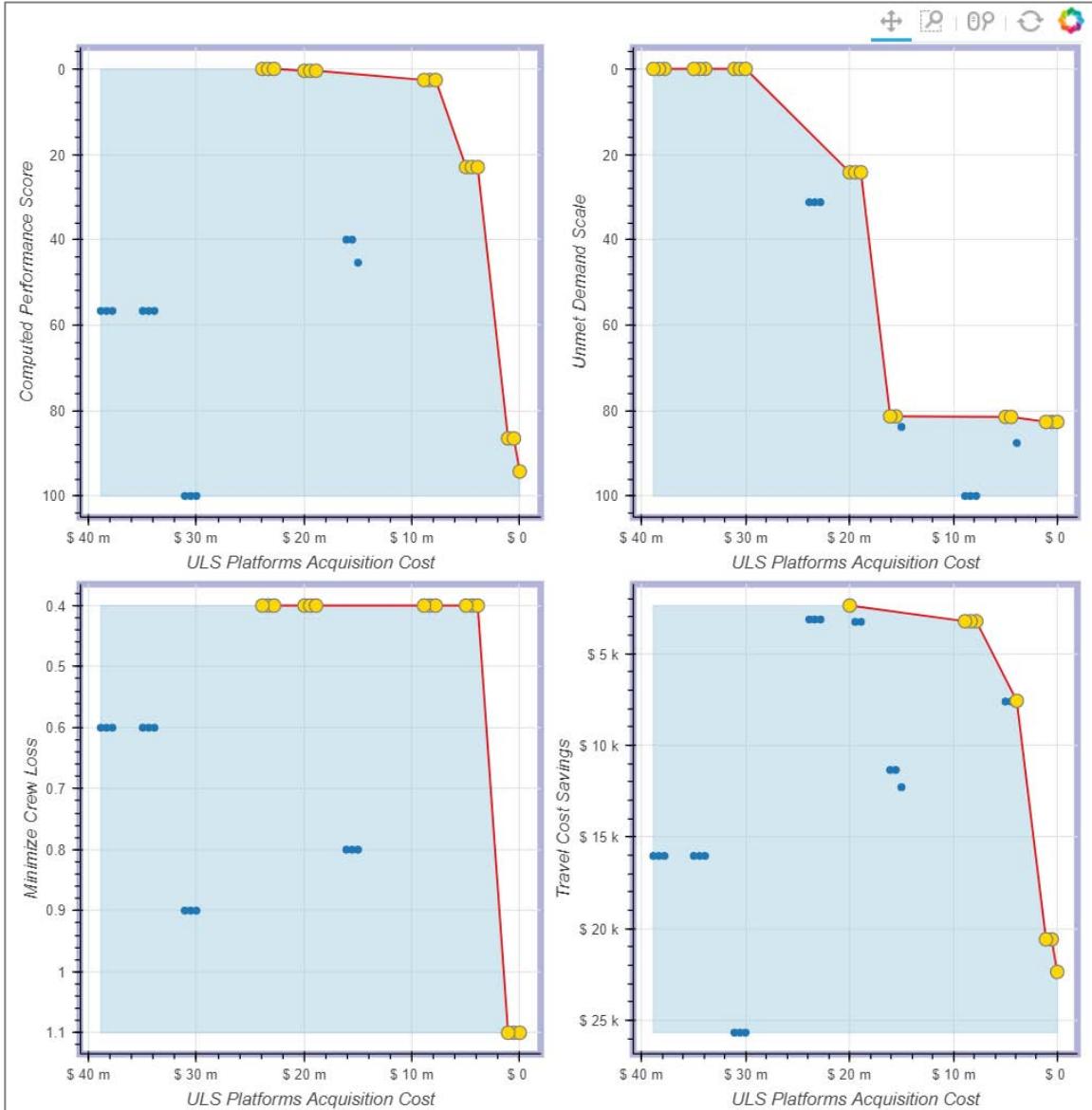


Figure 3: Pareto tradeoff curves.

But third – and perhaps most importantly – this integration across disciplines allows for ongoing experimentation into the nature of logistics operations involving unmanned

systems. The testbed environment is not intended to dictate the logistics decisions that should be made using UAVs, but rather as a means of understanding the nature of those operations under different conditions. The most significant result of this exercise is the generation of additional knowledge about how UAVs can be employed, and where the idiosyncrasies of those systems are not well modeled with current capabilities.

## Conclusions

The research presented here attempts to make contributions at several levels. At a practical level the ALOFT system and its testbed environment provides a means of generating geographic scenarios, implementing optimal logistics formulations, and displaying and analyzing the results of those implementations to evaluate unmanned vehicle performance. However, while those results may be useful for developing particular logistics plans and vehicle acquisition plans, it is the greater understanding of UAVs through the process of experimenting with scenarios and optimal logistics models that represents the value in this integrative process.

## References

- Curtin, K. M., Hayslett-McCall, K., & Qiu, F. (2010). Determining Optimal Police Patrol Areas with Maximal Covering and Backup Covering Location Models. *Networks and Spatial Economics*, 10(1), 125–145. <https://doi.org/10.1007/s11067-007-9035-6>
- Curtin, K. M., Voicu, G., Rice, M. T., & Stefanidis, A. (2014). A Comparative Analysis of Traveling Salesman Solutions from Geographic Information Systems. *Transactions in GIS*, 18(2), 286–301. <https://doi.org/10.1111/tgis.12045>
- Hale, T. S., & Moberg, C. R. (2003). Location Science Research: A Review. *Annals of Operations Research*, 123(1–4), 21–35. <https://doi.org/10.1023/A:1026110926707>

**Kevin M. Curtin**, Professor, Department of Geography, Director – Laboratory for Location Science, University of Alabama, Tuscaloosa, AL 35401

# A 3D Spatial Optimization Problem for Determining Optimal Locations for Bluetooth Beacon Placement

Brent Dell and May Yuan

**ABSTRACT:** Navigation in GIS environment requires traversal on a network with location inputs for origin and destination. We typically geolocate the origin through use of global positioning systems (GPS). This works generally well for outdoor navigation, where receivers have full line of sight to multiple GPS satellites. Within an indoor environment, features, such as walls and floors, block, scramble or reflect most GPS signals.

One solution to indoor positioning is through Bluetooth connection between mobile devices and beacons. These beacons have a fixed location ( $x$ ,  $y$ ,  $z$ ) and send packeted information over Bluetooth wireless signal. Triangulation of known beacon locations and distances estimated from Bluetooth signals to mobile devices can determine indoor locations. However, existing studies have not yet fully examined factors to determine how many beacons are necessary, where they should be placed, and what influence the positioning uncertainty in a complex indoor environment. This research aims to systematically study these questions in the context of indoor navigation.

We develop an algorithm to maximize signal coverage of Bluetooth beacons in a multi-level building structure. Pre-assumed constraints are applied to the model that reflect the working environment. The first constraint is that the beacons need to be placed in an un-obstructable location within hallways. The location must also be fixed in order for triangulation to be reliable. The second constraint requires that every location within the building can derive distance to at least four beacons. This is necessary to triangulate an individual within three-dimensional space.

Ji et al. (2015) discuss the prevalence of Bluetooth low energy (BLE) devices for various purposes. Through their research in a simple spatial layout, they demonstrated that as the number of deployed beacons increases, the error rate in geolocation decreases. Similarly, as beacon intervals increased, so did the error rate in meters. This data will factor in to spatial optimization of beacon placement. Following Ji et al. (2015) we test the capabilities of Estimote and Eddystone beacons for use in indoor navigation networks. Eddystone beacons are preferred over ibeacons due to rich sets of libraries and diverse applicable platforms. In addition, Estimote beacons come with software development kits and key spatial information, such as proximity.

Afghantoloe et al. (2014) discuss coverage estimation within 3D vector environments. The typical coverage estimation approach is to use a direct surface model, however, Afghantoloe et al. (2014) argue that vector data models can provide more accurate results in Wireless geoSensor Network (WSN) optimization. Expanding from their studies, we utilize a 3D polygon vector environment to model optimal locations for beacon placement.

We chose the Cecil B. Green Hall on the University of Texas at Dallas campus as the test site. The Green Hall is a four-story building with an opening cut across multiple floors in the middle of the building with multiple hallways and pillars throughout. Our familiarity and access to the building provide certain advantages in testing beacon placements and functionality. This study is also part of a larger smart campus initiative, which objectives include an integrated indoor/outdoor navigation system.

**KEYWORDS:** Optimal location, Bluetooth beacon, indoor, location

## **References:**

Afghantoloe, A., S. Doodman, F. Karimipour, and M. A. Mostafavi. 2014. "COEVRADE ESTIMATION OF GEOSENSOR IN 3D VECTOR ENVIRONMENTS." *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XL-2/W3 (October): 1–6. <https://doi.org/10.5194/isprarchives-XL-2-W3-1-2014>.

Ji, Myungin, Jooyoung Kim, Juil Jeon, and Youngsu Cho. 2015. "Analysis of Positioning Accuracy Corresponding to the Number of BLE Beacons in Indoor Positioning System." In *Advanced Communication Technology (ICACT), 2015 17th International Conference On*, 92–95. IEEE.

**Brent Dell**, Ph.D. Student, Department of Geospatial Information Sciences, University of Texas at Dallas, TX 75080

**May Yuan**, Ashbel Smith Professor of Geospatial Information Sciences, Department of Geospatial Information Sciences, University of Texas at Dallas, TX 75080

# Embracing Visualization as a Key Element in Computational Movement Analytics

## Somayeh Dodge

**ABSTRACT:** Visualization is a fundamental element in computational movement analytics. It is used to unveil hidden patterns of movement and communicate the results of computational methods. It can facilitate hypothesis generation and evaluation of algorithms. This paper aims to highlight the important role of geographic visualization in computational movement analytics, and discuss opportunities and challenges associated with visualization of movement.

**KEYWORDS:** Movement; Visualization; Visual Analytics; Movement Analytics; Trajectory; Space-time Cube; Animation; Cartography.

## Introduction

Movement is key to many social, natural, and ecological systems. Movement data are spatiotemporal signals of real-world entities that carry important information about movement behavior of individuals, their social interactions, and their relationships to their environment. The increasing availability of movement data sets and recent advances in computational movement analytics (Laube, 2014) have led to the emergence of a multidisciplinary research area to solve the complex problems associated with analyzing and modeling movement and to advance the understanding of dynamic natural and human systems.

Movement is realized in a multidimensional space — encapsulating geography, context (e.g. social and environmental variables), and time. Because of its complex multidimensionality nature, effective representation of movement and its patterns remains as an ongoing challenge. With the advances in sensor technologies (e.g. GPS collars, RFID tags, transit data, social media) and abundance of movement data sets, together with the availability of large data sets of context variables (e.g. remote sensing data of environment), the Geographic Visualization community is positioned well to undertake this challenge to advance visualization methods and effective visual analytic techniques for capturing movement in relationships to its environment.

Visualization provides a powerful means for data exploration and discovery of hidden patterns by giving a cognitively plausible visual structure to complex data sets through aggregation, generalization, and cartographic processes. Dodge (2016) introduced the continuum of movement research (figure 1), which emphasizes its reliance on effective visualization approaches for data inspection, exploratory data analysis, illustration of patterns, communication and interpretation of results, and validation of algorithms and analysis outcomes. In fact, visualization is envisioned to encircle the entire continuum as an essential element that facilitates and supports all other components (figure 1 and figure 2). Research has shown visualization can boost multidisciplinary collaboration by providing a common language for visual

communication, and it can serve as a powerful tool for hypothesis generation by uncovering unknowns and discovery of unexpected structures in data sets (Dodge 2016; Dodge et al., 2017).

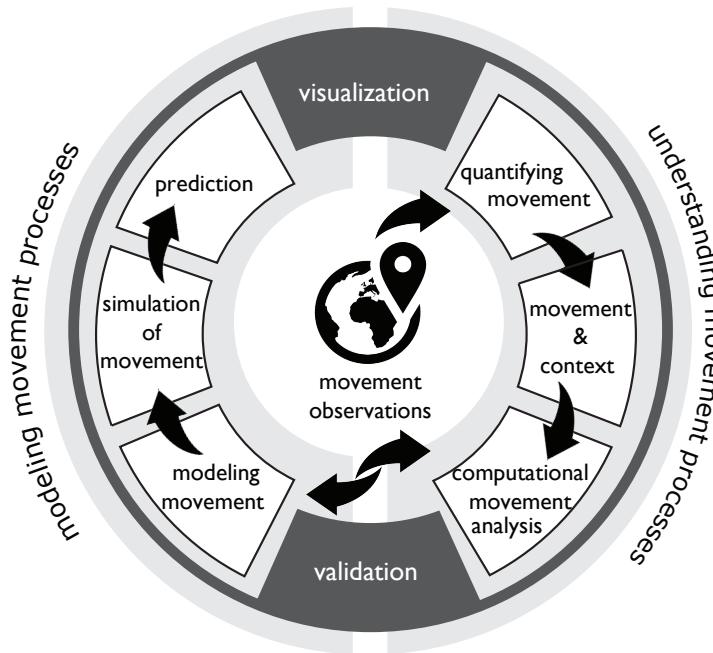


Figure 1. Visualization is a fundamental element in the continuum of movement research and facilitates all other analytical and modeling components. (Dodge, 2016)

While computational movement analytics have been advancing rapidly, embracing the increasing access to movement data sets, there has not been a significant breakthrough in geographic visualization and visual analytic approaches for movement. Existing movement visualization techniques offer mostly complex representations of moving phenomena that are not intuitive and are often static (Andrienko et al., 2013). These approaches are often based on the three dimensional Hägerstrand's space-time cube representation (Kraak, 2003; Demšar and Virrantaus, 2010), snapshots of movement tracks, complex flow maps and origin–destination matrices (Andrienko et al., 2013), hierarchically structured tree-maps (Slingsby et al., 2008), and two-dimensional time plots to summarize patterns in time (Andrienko et al., 2013; Song and Miller, 2012). Although these static representations can be effective for a small number of trajectories or movement in smaller geographic areas, they can be cognitively very complex, particularly, when a large number of long trajectories are involved and long-term moving processes are depicted (Lautenschütz, 2011). Song and Miller (2012) proposed a space-time data cube that provides a multidimensional representation of movement data in geographic and attributes space and time. However, because of the complexity of visualization in a 3D space, they decompose the cube to multiple 2D time plots. Animation seems to provide a promising and intuitive medium, especially when communicating movement to scientists of other disciplines (Xavier & Dodge, 2014). Although dynamic characteristics of movement restrict the usability of static representations, surprisingly there is not much work on movement animation and evaluation of animated visual displays to study movement.

This paper aims to highlight the significant role of geographic visualization as an important component of computational movement analytics. It briefly discusses possible opportunities and ongoing challenges facing the geographic visualization community to advance visualization techniques and support computational approaches.

## Opportunities and the Role of Visualization

Availability of large volumes of data sets and advances in computing systems facilitate development of new dynamic and interactive visualization techniques. Using fast and high-performance computers and powerful graphic cards, it is now possible to render and stream high resolution and high-quality dynamic visualizations in higher frequencies. Parallel advances in computational movement analytics continue to enable visualization of complex movement patterns by providing techniques to effectively summarize existing patterns in data sets (e.g. trajectory segmentation techniques, similarity analysis, trajectory clustering, etc.). Visualization is identified as the backbone of movement analytics and simulation models because it can communicate the outcomes of computational models in meaningful and effective ways and can facilitate cognitive processing of complex movement patterns and their relationships to context (Dodge 2016; Holloway and Miller, 2018). Figure 2 illustrates how visualization can help computational movement analytics. It provides an excellent tool to not only evaluate raw data sets, but also evaluate and validate the outcome of computational techniques, and movement models. Visualization should be embraced as a powerful means for interpreting extracted patterns through computational methods and inferring behavior. It can also support the validation process of computational methods by facilitating the evaluation of work process of the algorithms, interpretation of their results, and providing real-time feedback for examining movement simulations.

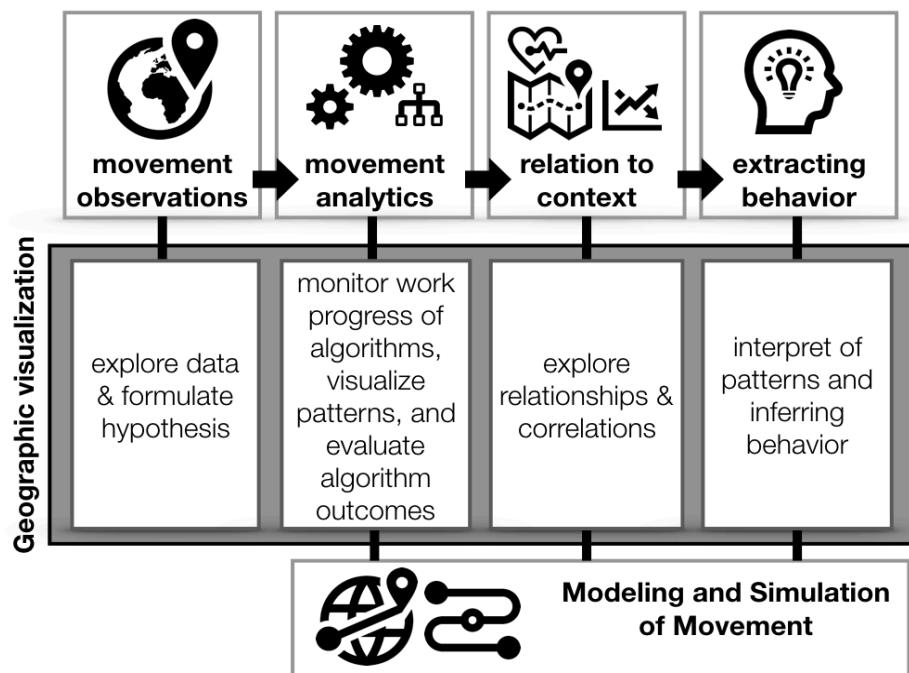


Figure 2. Geographic visualization contributes to other computational movement analytics by providing a means for data exploration, formulating new hypothesis, monitoring and evaluating the work progress of algorithms, exploring relationships between movement and context, and interpretation of patterns and inferring behavior

## Challenges in Visualization of Movement

Movement represents a change in the position of an individual over time. It is driven by an individual's behavior, by its interaction, and its surrounding environment at multiple spatial and temporal scales (Dodge, 2016). In order to represent movement in a plausible way and facilitate perception of complex patterns, it is important to consider all the interconnected driving components of movement: an individual's location, time, context, and scale. Although a significant body of literature documents visualization techniques to capture *change* over time and spatiotemporal patterns (e.g. time plots, snapshot series, animation, self-organizing maps) (Guo et al., 2006), there is little empirical work to evaluate how effective these tools are for representation of movement (Lautenschütz, 2011).

Advanced tracking technologies and new multi-modal sensors provide an unprecedented opportunity to study the associations between movement and its embedding context and exploring interaction among moving individuals with their environment. This also presents a challenge to development of new visualization approaches to effectively integrate all these dimensions in cognitively plausible visual forms to facilitate discovery of patterns, understanding interaction, and complex relationships between movement and environment. Griffin et al. (2017) also highlights the existing research gap in cartography for visualizing space-time data and complex spatiotemporal phenomena that are often captured in large data sets.

Other remaining challenges in this area include:

- Visualization of complex movement patterns dynamically, and proper highlighting and brushing techniques when visualizing large tracking data sets
- Alleviating privacy concerns in visualization of human mobility
- Difficulty for other disciplines to use existing complex GIS packages to map movement.
- Lack of simple and open source visualization tools to serve as a preliminary exploratory data analysis platform
- Multi-scale visualization of movement patterns

## Conclusion and Work Ahead

This paper highlighted the importance of visualization, and briefly discussed its ongoing challenges in the area of computational movement analytics. It is time to operationalize visualization in multidisciplinary work to study movement, particularly since a generic ontology and a common language for defining movement patterns is still lacking as suggested in Dodge et al. (2008). GIScience community should seek to develop more effective ways of illustrating patterns in movement data sets using simple, dynamic, and interactive visualization approaches. In the full version of this paper, using animation and visual analytics of movement, I will demonstrate how visualization can be used to enlighten multidisciplinary research and facilitate analytical approaches for exploring behavior of individuals and their interaction in ecological systems.

## References

- Andrienko, G., Andrienko, N., Bak, P., Keim, D. and Wrobel, S., (2013). *Visual analytics of movement*. Springer Science & Business Media.
- Demšar, U., & Virrantaus, K. (2010). Space–time density of trajectories: exploring spatio-temporal patterns in movement data. *International Journal of Geographical Information Science*, 24(10), 1527-1542.
- Dodge, S., (2016). From Observation to Prediction: The Trajectory of Movement Research in GIScience. In Onsrud, H. and Kuhn, W., (Eds.), *Advancing Geographic Information Science: The Past and Next Twenty Years*. Ch. 9. pp. 123 – 136. GSDI Association Press.
- Dodge, S., Weibel, R., Ahearn, SC., Buchin, M., Miller, J. (2016). Analysis of movement data, *International Journal of Geographical Information Science*, Volume 30, Issue 5, pages 825–834
- Dodge, S., Weibel, R., & Lautenschütz, A. K. (2008). Towards a taxonomy of movement patterns. *Information visualization*, 7(3-4), 240-252.
- Griffin, A. L., Robinson, A. C., & Roth, R. E. (2017). Envisioning the future of cartographic research. *International Journal of Cartography*, Vol 3, pp 1-8.
- Guo, D., Chen, J., MacEachren, A. M., & Liao, K. (2006). A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE transactions on visualization and computer graphics*, 12(6), 1461-1474.
- Holloway, P., & Miller, J. A. (2018). Analysis and Modeling of Movement. In Huang, B., (Eds.), *Comprehensive Geographic Information Science*, pp 162 –180, Volume 1, Chapter 13, Elsevier.
- Kraak, M. J. (2003). The space-time cube revisited from a geovisualization perspective. In *Proc. 21st International Cartographic Conference* (pp. 1988-1996).
- Lautenschütz, A. K. (2011). *Assessing the relevance of context for visualizations of movement trajectories* (Doctoral dissertation).
- Slingsby, A., Dykes, J., & Wood, J. (2008). Using treemaps for variable selection in spatio-temporal visualisation. *Information Visualization*, 7(3-4), 210-224.
- Song, Y. and Miller, H.J., 2012. Exploring traffic flow databases using space-time plots and data cubes. *Transportation*, 39(2), pp.215-234.
- Xavier, G., & Dodge, S. (2014). An exploratory visualization tool for mapping the relationships between animal movement and the environment. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Interacting with Maps* (pp. 36-42). ACM.

**Somayeh Dodge**, Assistant Professor, Department of Geography, Environment, and Society, University of Minnesota, Twin Cities, NM 55455

# Automatic Alignment of Geographic Features in Contemporary Vector Data and Georeferenced Historical Maps Using Reinforcement Learning

Weiwei Duan and Yao-Yi Chiang

**ABSTRACT:** With large amounts of digital map archives becoming available, the capability to automatically extracting information from scanned maps is important for many domains that require long-term geographic data, such as understanding the development of the landscape and human activities. Convolutional Neural Networks (CNN) have shown impressive performance in image recognition. In our previous work, we built a Fully Convolutional Neural Network to recognize geographic features in georeferenced maps automatically and achieved good performance. However, Neural Networks need the large number of representative training data, which makes manually generating training data unpractical to process tens of thousands of scanned maps. Our solution is to use contemporary vector data to automatically label examples of the geographic feature of interest in scanned maps as training samples for the model. However, positional misalignment exists between geographic features in maps and vector data due to the fact that the two datasets are published in different years and generated at different scales, which causes the automatically generated training data less representative. In this poster, we introduce our alignment algorithm to solve the problem using reinforcement learning. We formally model the alignment problem using the reinforcement learning framework and efficiently train a model. Our algorithm can align various types of geographic features in vector format from a contemporary source to scanned maps while other existing work focus on road features. The experiment shows that our alignment algorithm achieved promising results and can improve the overall performance for automatic feature extraction from maps.

**KEYWORDS:** Vector-and-raster alignment; Reinforcement Learning; USGS topographic maps

**Weiwei Duan**, PhD student, Department of Computer Science, University of Southern California, Los Angeles, CA 90007

**Yao-Yi Chiang**, Professor, Spatial Sciences Institute, University of Southern California, Los Angeles, CA 90007

# Automatic Generation of Precisely Delineated Geographic Features from Georeferenced Historical Maps Using Deep Learning

Weiwei Duan, Yao-Yi Chiang, Craig A. Knoblock, Johannes H. Uhl and Stefan Leyk

**ABSTRACT:** Historical map scans are now accessible through a number of online archives, but how to efficiently use the information in these map images in an analytic environment remains a challenge. Deep Convolutional Neural Networks (DCNNs) have demonstrated impressive performance in image recognition, including extracting geographic features from historical maps and converting the features to a vector format. In a typical image recognition process using DCNNs, the input image goes through multiple down-sampling processes and hence can result in the loss of spatial accuracy and poorly delineated object boundaries. In this paper, we present an analysis of two state-of-the-art DCNNs for digital map processing. We show that with carefully designed network architectures, DCNNs can extract precisely delineated boundaries of geographic features from scanned historical maps.

**KEYWORDS:** digital map process, deep learning, image recognition, image segmentation, historical maps

## Introduction

Historical maps store valuable information documenting human activities and natural features on Earth over long time periods. Such information is very useful for analysis that requires detailed historical geographic data. With hundreds of thousands of historical maps scanned and stored in digital archives, existing map processing methods (e.g., discussed in Chiang et al. 2014) that require manual user intervention remain inefficient. Deep Convolutional Neural Networks (DCNNs) (e.g., He et al. 2016) have shown impressive performance in automatic image recognition when sufficient training data is available. However, DCNNs abstract the input images through multiple down-sampling processes (e.g., convolutional and max-pooling layers), which can result in the loss of local image information and poorly delineated object boundaries. For many computer vision applications, well-recognized object boundaries are often not a strict requirement (e.g., recognizing individual persons in an image); however, for the extracted map features to be useful in scientific studies, the delineated boundaries need to be precise given that a one-pixel offset in the feature geometry may correspond to a distance of several meters on the earth surface.

In our previous work, we used the VGG network (Simonyan et al. 2014) for feature extraction from scanned maps (Duan et al. 2017), but due to the architecture limitations, the recognition results often have inaccurate and shifted boundaries. To overcome such limitations, new methods (e.g., Long et al. 2015, Yu et al. 2015, Bertasius et al. 2015) have been proposed and achieved significant improvements on non-document images (e.g., public datasets such as PASCAL-Context). However, maps are a specific type of document image. They contain cartographic symbols with boundary representations significantly different from the (non-document) image objects that most of the existing segmentation models have been tested on. First, compared with images in the public datasets, image pixels representing a geographic feature of interest in a map document occupy only a small proportion of the entire image (Figure 1). **Therefore, even**

**slightly reducing the spatial resolution during the training of a Deep Learning model can result in a significant drop of the spatial accuracy in the extraction results.** Second, the graphical representations of cartographic symbols belonging to different map layers can be very similar (Figure 2), which can result in high proportions of **false positives** if there are geographic features that resemble the feature of interest.

In this paper, we present an analysis of two state-of-the-art Deep Learning segmentation models (a fully convolutional network (FCN) (Long et al. 2015) and the context module (CM) (Yu et al. 2015)) to extract precisely delineated boundaries of geographic features from scanned maps. The goal of this paper is to provide a basic understanding of the performance of Deep Learning models in digital map processing and identify the limitations and future directions of using these models.

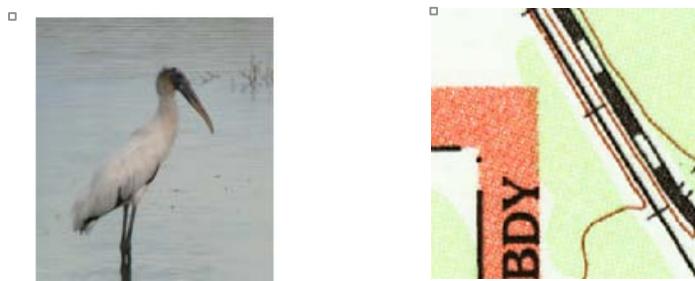


Figure 1: The object-of-interest (a bird) in typical datasets for segmentation (left) occupies a large region in the image compared to the railroads in maps (right, about 2.5% of the entire image).

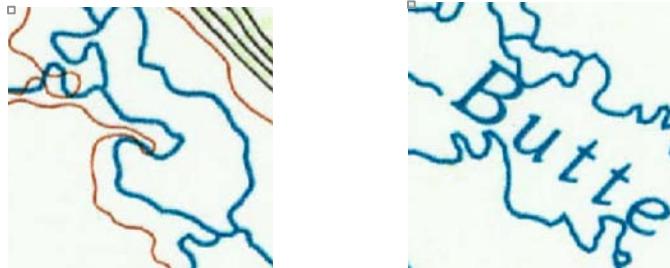


Figure 2: Examples of waterlines (left) and characters (right) in a USGS topographic map having the same graphical image representation (blue lines)

## Segmentation Models

This section briefly introduces the two types of segmentation models we built based on FCNs (Long et al. 2015) and the CM (Yu et al. 2015) architectures. As explained below, FCNs change the spatial resolution of the input images while CM does not.

**Fully Convolutional Network:** Long et al. 2015 proposed the Skip Architecture, which combines the final and intermediate results in DCNNs to recover the lost image resolution. Figure 3 shows our FCN architecture. Note that the resolution of the input image is reduced after each of the pooling layers (i.e., larger grid size). Our architecture combines the results from the second max-pooling and final layer. The idea behind this combination is that the intermediate

layers contain more local information that could be useful to preserve boundary details in the final results.

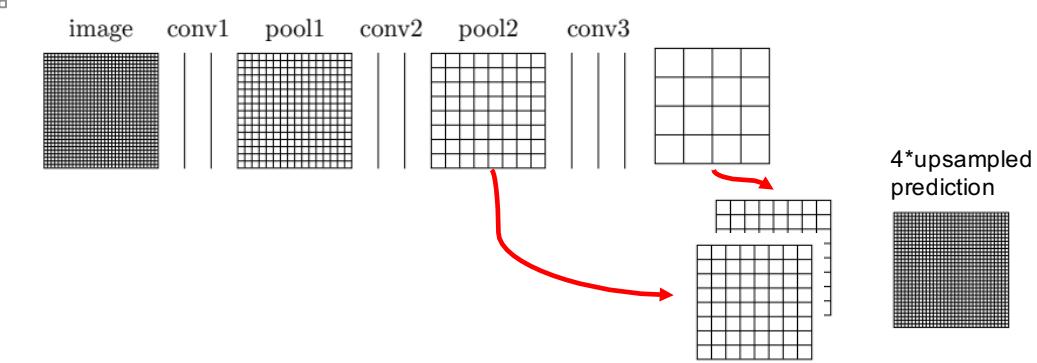


Figure 3: Our FCN architecture

**Context Module:** Yu et al. 2015 proposed the CM, which does not use pooling layers and **hence maintains the spatial resolution of the input images** in DCNNs but uses dilated convolution layers to enlarge the receptive fields. The dilated convolution process enlarges the filter sizes and can capture more local information than conventional convolution filters. Table 1 shows the architecture of our CM, which consists of eight layers.

Table 1 The architecture of our CM. The last layer is the classification layer, and the first seven layers have different dilation rates.

Layer	1	2	3	4	5	6	7	8
Convolution kernel size	3*3*32	3*3*64	3*3*128	3*3*256	3*3*256	3*3*512	3*3*512	1*1*2
Dilation	1	1	2	4	8	16	1	1
Receptive field	3*3	5*5	9*9	17*17	33*33	65*65	67*67	67*67

## Preliminary Results & Outlook

We tested the segmentation of railroads and waterlines on two historical USGS topographic map sheets: Bray, California, 2001 and Louisville, Colorado, 1965 (Historical maps here refer to the maps generated before the practice of map generation in electronic form. For USGS topographic maps, it is before 2009). We trained both the FCN and CM using automatically generated training data (Duan et al. 2017), and we used the trained models to classify each pixel in the maps. We manually labeled locations of railroads and waterlines in the test maps to generate the ground truth and used correctness and completeness (Heipke et al. 1997) as the evaluation metrics (Table 2). Overall, both models extracted railroad and waterline features reliably (i.e., high completeness) with very few gaps that cause the discontinuity in linear features. One notable result is that the CM generated considerably fewer false positives than the FCN for railroads extracted from the map of Louisville but more false positives for waterlines extracted from the map of Bray. This could be because the CM uses the larger field of view, which could

successfully distinguish dashed and continuous black lines from graphic representations of railroads (Figure 4) but misclassified some blue-tone characters as waterlines (Figure 5).

These results give us confidence in using the intermediate layers as well as large filters to preserve object boundaries when using DCNNs for segmenting geographic features in scanned map documents. In addition, an important finding is that without sacrificing the spatial resolution (and hence achieving higher levels of recognition accuracy), we are able to focus on the design of convolution filters to improve the segmentation quality. Currently, we are in the process of building the techniques to automatically learn boundary conditions to further improve the segmentation results (e.g., Bertasius et al. 2015). Finally, we will also incorporate the topological and geometric characteristics of the features (e.g., railroads should be straight within some distance) to formalize rules that improve the overall results.

Table 2: The experiment results of two models (1-pixel buffer)

	Bray				Louisville	
	Railroads		Waterlines		Railroads	
	Correctness	Completeness	Correctness	Completeness	Correctness	Completeness
FCN	84.74%	97.46%	92.59%	98.01%	68.60%	96.64%
Context Module	82.90%	97.86%	88.81%	98.26%	78.83%	96.73%

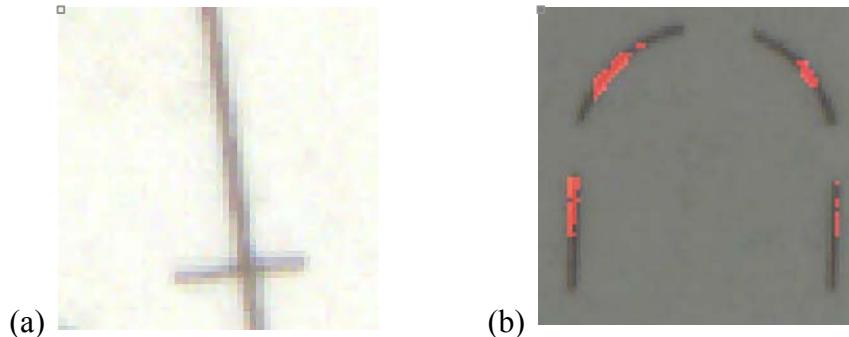


Figure 4: An example of false recognition results of the FCN. (a) A railroad segment in the map of Louisville; (b) Red lines are falsely recognized as railroads by the FCN in the map of Louisville.

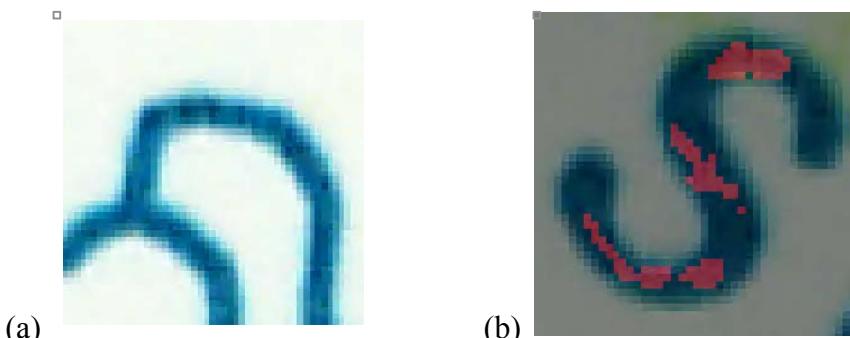


Figure 5: An example of false recognition results of the CM. (a) An example of waterlines in the map of Bray; (b) Red lines are falsely recognized as waterlines by CM in the map of Bray

## References

- Bertasius, G., Shi, J., & Torresani, L. (2015). High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision. In *IEEE ICCV* (pp. 504-512).
- Chiang, Y.-Y., Leyk, S., and Knoblock, C. A. (2014). A Survey of Digital Map Processing Techniques. *ACM Computing Surveys*, 47(1):1–44. doi: 10.1145/2557423
- Duan, W.; Chiang, Y.; Knoblock, C. A.; Jain, V.; Feldman, D.; Uhl, J. H.; and Leyk, S. (2017). Automatic Alignment of Geographic Features in Contemporary Vector Data and Historical Maps. In ACM SIGSPATIAL GeoAI Workshop.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE CVPR* (pp. 770-778).
- Heipke, C., Mayer, H., Wiedemann, C., & Jamet, O. (1997). Evaluation of automatic road extraction. *IAPRS*, 32(3 SECT 4W2), 151-160.
- Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *IEEE CVPR* (pp. 3431-3440).
- Qiao, K., Chen, J., Wang, L., Zeng, L., & Yan, B. (2017). A top-down manner-based DCNN architecture for semantic image segmentation. *PloS one*, 12(3), e0174508.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *IEEE CVPR* (pp. 1-9).
- Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

**Weiwei Duan**, Ph.D. student, Department of Computer Science, University of Southern California, Los Angeles, CA

**Craig A. Knoblock**, Professor, Department of Computer Science, University of Southern California, Los Angeles, CA

**Yao-Yi Chiang**, Professor, Spatial Sciences Institute, University of Southern California, Los Angeles, CA

**Johannes H. Uhl**, Ph.D. student, Department of Geography, University of Colorado, Boulder, CO

**Stefan Leyk**, Professor, Department of Geography, University of Colorado, Boulder, CO

# Optimizing Activity Locations in GIS using a Multi-Objective Trajectory Approach

**Xin Feng, Shaohua Wang, Alan T. Murray, Yuanpei Cao, Song Gao**

**ABSTRACT:** Human movement and interaction in space over time provides opportunity to structure economic and social activities. Having access to rich information provided by current location-based technologies, people are able to make good decisions to satisfy social activity participation needs and travel behavior. Identifying an optimal trajectory connecting desired activity locations via the road network when there are multiple attendees with space-time constraints is a challenging task. This spatial organization problem is formulated mathematically as a sequential, multi-objective optimization model. A framework consisting of context knowledge, geographic information system, and spatial optimization is structured to solve it. The framework allows for the integration of geographical context and other geospatial inputs to support social networking. The proposed approach offers a way to balance tradeoffs attributable to activity location, enabling travel cost, personal preference, quality rating, etc. to simultaneously be considered in planning and decision making. A case study is detailed involving the organization of a series of activities reflected in points of interest provided by Yelp for a group of individuals. The application results highlight the utility and insight afforded through the use of the proposed spatial analytical framework.

**KEYWORDS:** Multi-Objective trajectory approach; network analysis; sequential spatial optimization

**Xin Feng**, Ph.D. candidate, Department of Geography, University of California, Santa Barbara, CA 93106

**Shaohua Wang**, Ph.D., Institute of Geographic Science and Natural Resources Research, Chinese Academy of Science, Beijing, China

**Alan Murray**, Professor, Department of Geography, University of California, Santa Barbara, CA 93106

**Yuanpei Cao**, Ph.D., Airbnb, San Francisco, CA 94103

**Song Gao**, Assistant Professor, Department of Geography, University of Wisconsin, Madison, WI 53706

# Unmanned Aircraft Systems and the Atmospheric Boundary Layer: A New Frontier for Geospatial Data Science?

Amy E. Frazier and Benjamin L. Hemingway

**ABSTRACT:** The lowest portion of the Earth's atmosphere, known as the atmospheric boundary layer (ABL), plays an important role in the formation of weather events. Collecting simple meteorological measurements from within the approximately 1-km thick ABL, such as temperature, humidity, and wind velocity, is critical for understanding the processes operating therein, but the ABL is one of the most challenging regions of the atmosphere to measure. Geospatial technologies such as satellites, radar, and weather balloons that can sample the atmosphere have contributed tremendously to our understanding of the physical processes occurring within the ABL, but each of these approaches has spatial and temporal shortcomings. Small, unmanned aircraft systems (suAS), also known as drones, have emerged as versatile and dynamic platforms for atmospheric sensing that can fill the spatio-temporal sampling gaps left by conventional, geospatial surveillance methods, but their utility for conducting geospatial investigations in the ABL has been understudied. Many questions surrounding the appropriate sampling scales for capturing these measurements need to be addressed before they can be integrated into routine weather monitoring.

In this research, we use a commonly used geostatistical technique—variogram modeling—to capture the spatial structure of key thermodynamic variables including temperature and relative humidity. Variogram modeling can be translated from geographic phenomena to atmospheric phenomena since the physical processes affecting the spatial variation of both are complex, making their behavior appear random. Atmospheric phenomena also adhere to similar physical laws since their movements are governed by the non-linear effects of turbulence. Thus, there remains an inherent structure to atmospheric data, and their values have a statistical relationship relative to their location in space. By modeling this structure using variograms, we can determine the spatial continuity of the data at different distances between sampling points to identify the optimal spatial sampling scales at which autocorrelation is no longer affecting measurements.

We flew vertical profiles with a 3DR Iris during summer 2016 and captured atmospheric measurements of temperature and humidity using an iMet sensor. Variogram analysis indicated that vertical sampling scales of approximately 3m for temperature and 2m for relative humidity were sufficient to capture the spatial structure of these phenomena under the conditions tested. Future work will focus on testing these findings across a variety of geographic and climatic conditions as well as extending our understanding of atmospheric structures to two- and three-dimensions using variograms.

**KEYWORDS:** unmanned aerial vehicles, drones, meteorology, atmospheric physics, geostatistics

**Amy E. Frazier**, Assistant Professor, Department of Geography, Oklahoma State University, Stillwater, OK 74078 and the School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85281

**Benjamin L. Hemingway**, Ph.D. Student, Department of Geography, Oklahoma State University, Stillwater, OK 74078

# Gerrymandering and Geospatial Analysis of Redistricting Plans

## Lee Hachadoorian

**ABSTRACT:** Gerrymandering is a growing national concern. Prior to the 1960s, Congressional districts often respected county and municipal boundaries, but wildly differing populations among districts led to the establishment of the one person, one vote standard. In an effort to conform to the requirements of one person, one vote, Congressional districts became increasingly less compact, and increasingly more gerrymandered. With cheaper computing power and the increasing adoption of GIS, the redistricting implemented by several states following the 2000 and 2010 censuses fare poorly on measures of compactness and have arguably been among the most egregious gerrymanders the United States has ever seen. In different states, gerrymanders benefit both Republicans (Pennsylvania, Wisconsin, North Carolina) and Democrats (Maryland, Illinois). In the face of increasingly obvious gerrymanders, several court cases have successfully challenged Congressional districts in lower courts, and some of them have been taken up by the Supreme Court for review.

In this talk, I discuss the use of open data and open source software (primarily R and QGIS) to analyze redistricting and model election outcomes of hypothetical districts. Additionally, I discuss my work with Concerned Citizens for Democracy, a nonpartisan Pennsylvania-based advocacy group. Pennsylvania, like several other state constitutions, requires that counties and municipalities be split only if “absolutely necessary”. We demonstrate that it is possible to achieve equal district populations with many fewer county and municipal boundary splits than in the current, recently invalidated PA district plan. I will also discuss redistricting guidelines that, if adopted by the courts, will severely limit the ability of political parties to create gerrymandered districts.

**KEYWORDS:** redistricting; gerrymandering; FOSS4G; political geography; compactness

**Lee Hachadoorian**, Assistant Professor of Instructions, Department of Geography and Urban Studies, Temple University, Philadelphia, PA 19122

# Forecasting County-Level Maize Yield with Deep Learning and Satellite Remote Sensing

**Yanghui Kang and Mutlu Özdogan**

**ABSTRACT:** To maintain food security for an exploding world population, it is more imperative than ever before to provide timely and reliable estimates of crop yield. The timeliness and robustness of such prediction can significantly impact decision making about field management, grain market, and crop insurance. In this project, we exploited the merits of satellite remote sensing data and a deep learning algorithm to improve maize yield prediction at county-level for 10 US States. We compiled yield data (2001 to 2016) from the National Agricultural Statistic Service (NASS), and constructed feature vectors using various satellite and weather data extracted from Google Earth Engine. The feature space is comprised of time varying vegetation indices, soil moisture, land surface temperature, and weather observations, as well as non-sequential features including geographic coordinates and year of harvest. A Long-Short Term Memory (LSTM) neural network was trained based on 10-year observations and tested using data from the following year. The end-of-season estimation had Root Mean Squared Error (RMSE) of 14.7 bushels/acre (8.5%) on average for years 2011 to 2016. Yield forecast in early August (two and a half months before harvest) had a coefficient of variance of RMSE ranging between 9% and 14% while early September forecast was within 1% of the end-of-season estimation for most years. Satellite derived Enhanced Vegetation Index (EVI) and soil moisture were found to have the highest correlation with final yield. This work contributes to the operationalization of large-scale yield forecasting and improves our understanding of the role atmospheric/environmental variables play in estimating of crop yields over large areas.

**KEYWORDS:** Yield, MODIS, Deep Learning, Google Earth Engine

**Yanghui Kang**, PhD student, Department of Geography, University of Wisconsin - Madison, Madison, WI 53706

**Mutlu Özdogan**, Associate Professor, Forest Ecology and Environmental Studies, University of Wisconsin - Madison, Madison, WI 53706

**Geovisual Text Analytics for Exploring Public Discourse on Twitter:  
A Case Study of Immigration Tweets Before and After the  
January 27, 2017 Travel Ban**

**Caglar Koylu, Bryce J. Dietrich and Ryan L. Larson**

**ABSTRACT:** Social media provides a unique opportunity to study geographic variation and evolution of content and sentiment of publicly shared opinions. However, it is challenging to identify the general patterns and trends in public discourse due to the complexity of the tone, topic, geographic and temporal variation in the way people express their opinions. This paper examines how public discourse varies across political geographies and over time by introducing a geovisual analytics environment that integrates topic modeling and sentiment analysis with spatiotemporal visualization. Our proposed study makes two major contributions. First, we contribute to the analysis of public discourse on immigration by introducing a natural experiment by collecting and analyzing tweets related to immigration a month before and a month after the 2017 Travel Ban. Second, we introduce a geovisual analytics framework that allows simultaneous and linked views of topical themes and sentiment patterns of public discourse on immigration as well as the variation in patterns across states and the eight-week time period before and after the ban. Preliminary results of this study revealed that the intensity of tweets about immigration substantially increased right after the immigration ban, and the geographic intensity of online discourse was correlated with the protest events held at a number of airports.

**KEYWORDS:** Geovisual analytics, topic modeling, sentiment analysis, Twitter, immigration, public discourse

## **Introduction**

Diffusion of policies are often influenced by citizens who express their opinion through public discourse on Twitter (Vasi et al., 2015). Public discourse often varies by states, political and administrative areas and evolves in reaction to policies and real world events. Social media provides a unique opportunity to study geographic variation and evolution of content and sentiment of publicly shared opinions and identify political geographies of public discourse. For example, on January 27, 2017, President Donald Trump suspended the entry of people into the U.S. from a number of predominantly Muslim countries. In response, thousands of people flooded airports across the country to protest what the travel ban meant for democracy in the United States and elsewhere. Similar protests were observed on Twitter. This paper explores the regional, national and temporal progression of tenor of these online protests, both in terms of their tone and topic.

Our proposed study makes two major contributions. First, we contribute to the analysis of public discourse on immigration by introducing a natural experiment by collecting and analyzing tweets related to immigration a month *before* and a month *after* President Trump signed Executive Order 13769. The vast majority of previous studies analyzed Twitter protests, including those that focused on the Arab Spring, looked at tweets after

the even occurred. For example, Bruns, Highfield and Burgess (2013) studied the Arab Spring using tweets from January to November 2011 which is a full month *after* the Arab Spring began on December 17, 2010. Our data collection and analysis enable the natural experiment in which we can determine how and in what context the discussion of immigration changed *after* the travel ban was announced. Second, we introduce a geovisual text analytics framework that allows simultaneous and linked views of topical themes and sentiment patterns of public discourse on immigration as well as the variation in patterns across states and the eight-week time period before and after the ban.

## Data and Methods

### Data

Using the Twitter Streaming API, we collected tweets that contain the keywords related to immigration, and specifically Muslim refugees and immigrants (i.e., “immigration”, “immigrant”, “Muslim”, “Islam”, “refugee”) in a number of languages including English, Arabic, French, Spanish, Turkish, and Persian a month before and a month after the Immigration Ban put in place on the January 27<sup>th</sup>, 2017. There were 84,159,489 immigration tweets world-wide from December 27, 2016 to February 27, 2017. The locations of 99 % of these tweets can be identified at the state or country level, and 21% (~17 million tweets) of the geo-located immigration tweets were generated in the Continental U.S by 759,171 users. Figure 1 illustrates the temporal distribution of immigration tweets, which highlights a large increase following the day the first travel ban was announced. This provides an initial evidence of the extensive online protest we reference in the introduction and demonstrates the “natural experiment” we aim to leverage in our study.

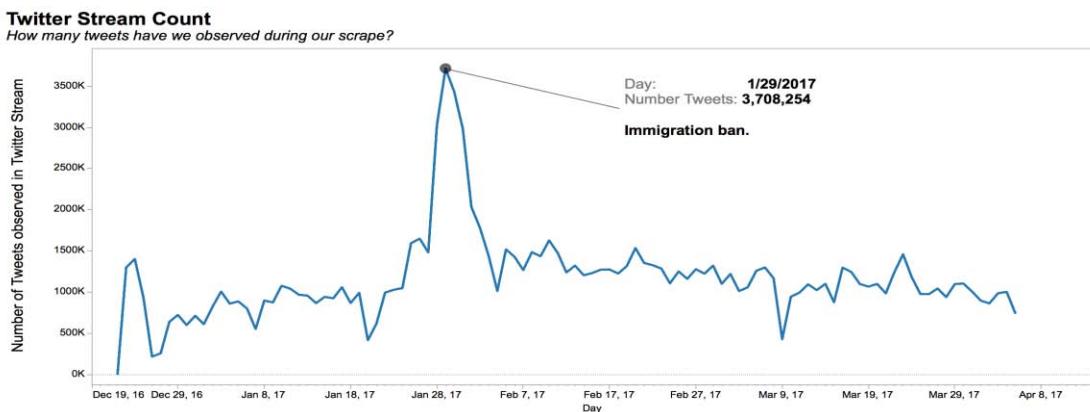


Figure 1: Immigration tweets per 1,000 people

Out of the 17 million tweets generated within the U.S., 49% of the tweets were re-tweets. Retweets were excluded in topic modeling and sentiment analysis, while they were used in identifying opinion leaders. Figure 2 shows the multiplicity the diversity of Twitter participants including celebrities, journalists, academics, politicians and activists whose

tweets were retweeted extensively. This finding reveals that the online protest was not focused on one specific group, but ultimately reached a diverse audience.

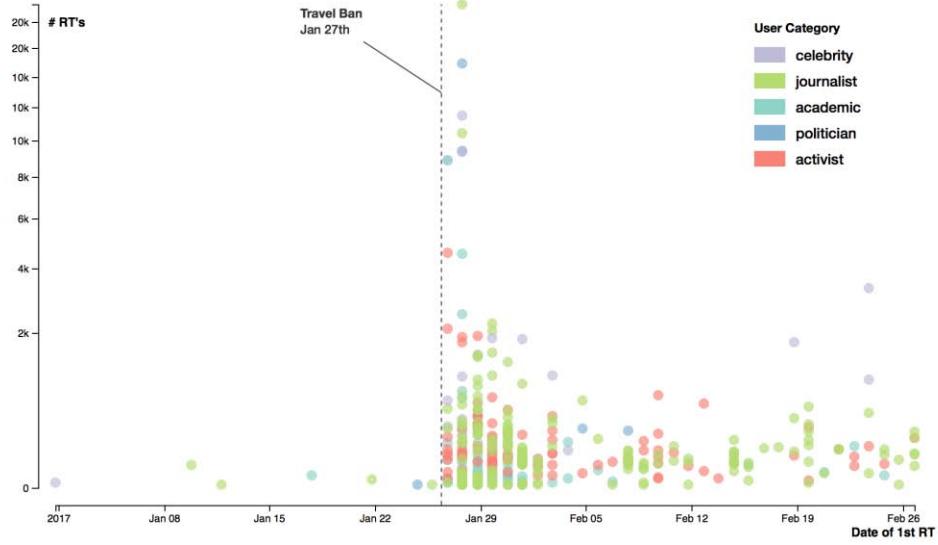


Figure 2: Opinion leaders: users with the most number of retweets before and after the ban

### ***Extracting topics and sentiment from tweets***

First, pre-processing phase is used to clean and partition the data into eight weekly time periods a month before and after the travel ban. At this phase, tweets within each state are combined into documents into the eight week-period before and after the ban to obtain temporally organized document collections. Aggregation by states address the short-text problem in topic modeling (Yan et al., 2013) caused by the 140 character limit, and inability of words to occur within a single tweet. We employ a separate Latent Dirichlet Allocation (LDA) on the document collection for each of the eight-week time period before and after the ban. LDA is a generative process that relies on term frequency-inverse document frequency (tf-idf), which reflects how important each word is to a document in a collection of documents or corpus, and its value increases proportionally to the number of times a word appears in a document (Goldstone & Underwood, 2012). LDA has been used in a variety of studies to identify event locations, geographical variation of linguistics, and topical themes, and provide recommendations based on location from Twitter data (Chae et al., 2012; Koaylu, 2018a, 2018b; Lai, Cheng, & Lansley, 2017; Lansley & Longley, 2016; Liu, Ester, Hu, & Cheung, 2015; Pozdnoukhov & Kaiser, 2011). We employ LDA to identify representative topics, ultimately ranging from specific issues (i.e., xenophobia or border security) to different types of phrasing (i.e., formal versus informal language).

In addition to extracting topical themes from immigration tweets, we employ the Linguistic Inquiry Word Count (LIWC) in order to identify positive and negative sentiment, and the patterns across states over the eight-week period. Previous studies (Alpers et al., 2005; Bantum & Owen, 2009; Kahn et al., 2007) found that the “negative emotion” category outperformed all other LIWC categories, leading these authors to

conclude that this category could be used to track changes in expression of negative emotions in on-line groups. Kahn et al. (2007) argued that the LIWC provides a meaningful indicator of emotion that may be used as an alternative or complement to self-reports of emotion. Similar to these previous studies, we utilize the count of the number of “positive” and “negative” words in identifying overall sentiment of public discourse grouped by states and the eight-week time period.

## Preliminary Results

Figure 3 illustrates the number of immigration tweets within the eight-week period normalized by state population. Figure 3 suggests that the highest density of tweets centered on protest events in important airports like those found in New York, Chicago, and Boston. Since airports were the focal point of many protest events, we can use the proximity to airports to explore whether those demonstrations actually influenced the way people discussed immigration. If they did have an effect, then you would expect the greatest change in both the topic and tone of Twitter discussions to occur in close proximity to airports. Figure 3 demonstrates the initial layout, and the potential of the geovisual analytics framework for spatiotemporal comparisons of public discourse.

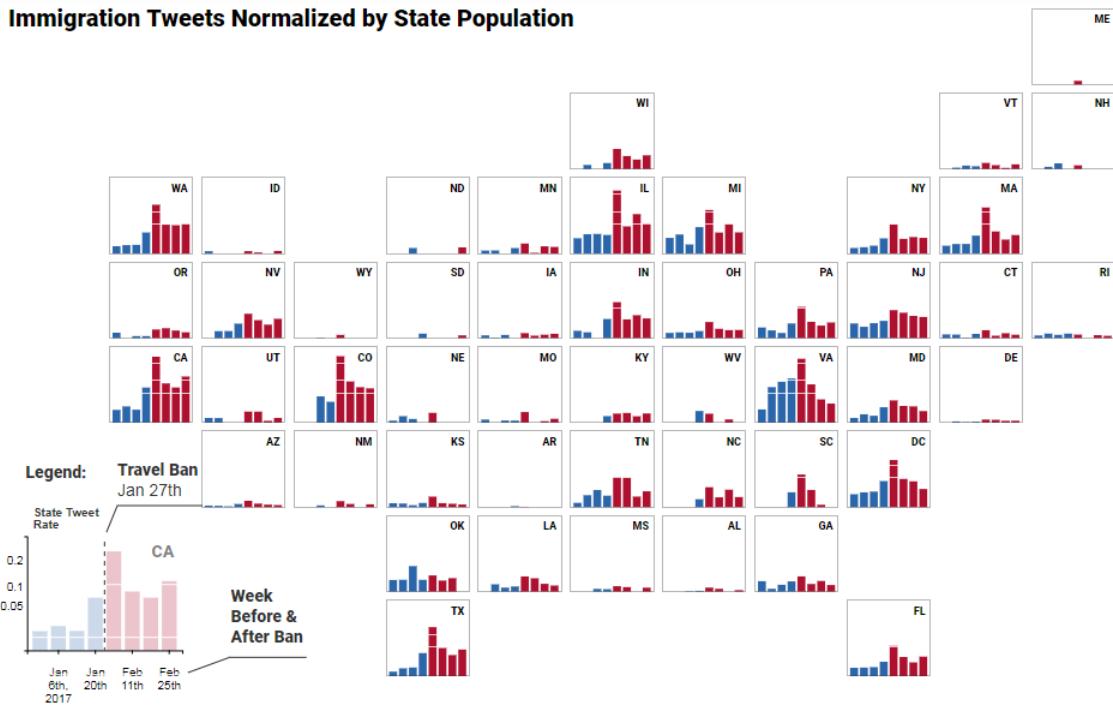


Figure 3: Spatiotemporal cartogram that illustrates the normalized frequency of immigration tweets per state.

## Conclusion and Future Work

With this paper, we not only provide a methodology for gaining a greater understanding of a very important recent moment in American political history, but we also contribute

to the broader discussion of how online protests emerge and ultimately influence broader political discussions. The next step is to integrate the results of topic modeling and sentiment analysis in a geovisual analytics environment to explore both regional and national variation, and the changing patterns of content and sentiment of public discourse on immigration over the course of the eight-week period including the month before and after the first travel ban. The layout presented in Figure 3 will be transformed into coordinated views of (1) the temporal progression of the public discourse both in terms of its content and sentiment; (2) topic word clouds illustrating the extracted themes from topic modeling, and (3) the sentiment visualizations that highlight the changing patterns of the tone in public discourse on immigration.

## References

- Alpers, G. W., Winzelberg, A. J., Classen, C., Roberts, H., Dev, P., Koopman, C., & Taylor, C. B. (2005). Evaluation of computerized text analysis in an Internet breast cancer support group. *Computers in Human Behavior*, 21(2), 361-376.
- Bantum, E. O. C., & Owen, J. E. (2009). Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological assessment*, 21(1), 79.
- Bruns, A., Highfield, T., & Burgess, J. (2013). The Arab Spring and social media audiences: English and Arabic Twitter users and their networks. *American Behavioral Scientist*, 57(7), 871-898.
- Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D. S., & Ertl, T. (2012). *Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition*. Paper presented at the Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on.
- Goldstone, A., & Underwood, T. (2012). What can topic models of PMLA teach us about the history of literary scholarship. *Journal of Digital Humanities*(2 (1)), 39-48.
- Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American journal of psychology*, 263-286.
- Koylu, C. (2018a). Modeling and visualizing semantic and spatio-temporal evolution of topics in interpersonal communication on Twitter. *International Journal of Geographical Information Science*, 1-28. doi:10.1080/13658816.2018.1458987
- Koylu, C. (2018b). Uncovering Geo-Social Semantics from the Twitter Mention Network: An Integrated Approach Using Spatial Network Smoothing and Topic Modeling. In S.-L. Shaw & D. Sui (Eds.), *Human Dynamics Research in Smart and Connected Communities* (pp. 163-179). Cham: Springer International Publishing.

Lai, J., Cheng, T., & Lansley, G. (2017). Improved targeted outdoor advertising based on geotagged social media data. *Annals of GIS*, 23(4), 237-250.

Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, 58, 85-96.

Liu, Y., Ester, M., Hu, B., & Cheung, D. W. (2015). *Spatio-Temporal Topic Models for Check-in Data*. Paper presented at the Data Mining (ICDM), 2015 IEEE International Conference on.

Pozdnoukhov, A., & Kaiser, C. (2011). *Space-time dynamics of topics in streaming text*. Paper presented at the Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks.

Vasi, I. B., Walker, E. T., Johnson, J. S., & Tan, H. F. (2015). "No fracking way!" Documentary film, discursive opportunity, and local opposition against hydraulic fracturing in the United States, 2010 to 2013. *American Sociological Review*, 80(5), 934-959.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). *A biterm topic model for short texts*. Paper presented at the Proceedings of the 22nd international conference on World Wide Web.

**Caglar Koylu**, Assistant Professor, Department of Geographical and Sustainability Sciences, University of Iowa, Iowa City, IA, 52242

**Bryce Dietrich**, Assistant Professor, Department of Political Science, University of Iowa, Iowa City, IA, 52242

**Ryan L. Larson**, Computer Science, University of Iowa, Iowa City, IA, 52242

# Area-Preserving Simplification of Polygon Features

**Barry J. Kronenfeld, Lawrence V. Stanislawski, Tyler Brockmeyer  
and Barbara P. Buttenfield**

**ABSTRACT:** Developing simplified representations of a two-dimensional polyline is an important problem in cartographic data analytics where datasets must be integrated across spatial resolutions. This problem is generally referred to as line simplification, and is increasingly driven by preservation of specific analytic properties such as positional accuracy and high-frequency detail. However, the distinction between linear features and polygon boundaries is rarely considered. Polygonal features differ fundamentally from linear features in that they represent areal regions covering a portion of the earth's surface. As such, they lend themselves naturally to the objective of preserving area across scale.

Only a few studies have investigated the preservation of area in polyline simplification. Bose et al. (2006) explore the problem under the restriction that Steiner points (points that are introduced when solving a geometric optimization to improve upon solutions based only on the original point set) are not allowed. They prove that it is NP-hard, meaning that the computing time required to find a solution increases so quickly with additional inputs (here, added vertices) that the solution becomes intractable. Using Steiner points, however, the problem becomes tractable. This is proved by Meulemans et al. (2010), who present an area-preserving schematization of polygon tessellations that uses Steiner points. Their schematization is rectilinear and provides effective schematic simplification of artificial features such as buildings, but is less suitable for lakes and other natural features.

We present a novel area-preserving polyline simplification algorithm using Steiner points. Unlike Meulemans' schematization, our algorithm produces a realistic facsimile similar in geometry to the original feature. The algorithm works by iteratively collapsing segments to points, and is referred to as Area-Preserving Segment Collapse (APSC). The concept of segment collapse is also used in Raposo's (2013) resolution-driven hexagonal spatial means algorithm. APSC however is driven by analytical error-minimization objectives. Specifically, selection and collapse are controlled by the simultaneous goals of minimizing areal displacement and balancing area preservation across both sides of the simplified boundary. Procedurally, APSC is similar to the Visvalingam and Whyatt (1992) effective area (VEA) algorithm, but instead of removing a single vertex at each step, two vertices are removed and one new vertex is added.

After discussing the rationale for preserving area and for using Steiner points, we describe the algorithm design and illustrate the mechanics of segment collapse, with a specific strategy for optimized vertex placement to preserve area. The method is illustrated using a sample of 10 lakes from across the contiguous United States formed from alpine, Karst, glacial, and arid desert processes as well as artificial dams, with areas ranging from 30 to 60,000 square kilometers. For comparison, features are simplified to a controlled number of vertices using APSC, VEA, the Ramer-Douglas-Peucker algorithm (Ramer 1972, Douglas and Peucker 1973) and Raposo's (2013) spatial means algorithm. The results are evaluated for areal displacement, area preservation and introduction of self-intersections.

Although self-intersections can occur with APSC due to protrusions created by Steiner points, preliminary results confirm that APSC preserves polygon area as intended. In addition, it

produces simplified features with the lowest areal displacement among the algorithms tested and is computationally efficient (worst case  $O(n \log n)$ )..

**KEYWORDS:** cartographic generalization, line simplification, polygons, area

## References

Bose, P., Cabello, S. Cheong, O., Gudmundsson, J., van Kreveld, M. and Speckmann, B. (2006) Area-preserving approximations of polygonal paths. *Journal of Discrete Algorithms*, 4, 4, pp. 554-566.

Douglas, D. H. and Peucker, T.K. (1973) Algorithms for the reduction of the number of points required to represent a digitised line or its caricature. *The Canadian Cartographer*, 10, 2, pp. 112-122.

Meulemans, W., van Renssen, A. and Speckmann, B. (2010) Area-preserving subdivision schematization. In Fabrikant, S.I., Reichenbacher, T., van Kreveld, M. and Schlieder, C., *Proceedings of the 6th International Conference on Geographic Information Science (GIScience 2010)*, Zurich, Switzerland, Sep. 14-17.

Ramer, U. (1972) An iterative procedure for the polygonal approximation of plane curves. *Computer Graphics and Image Processing*, 1, 3, pp. 244-256.

Raposo, P. (2013) Scale-specific automated line simplification by vertex clustering on a hexagonal tessellation. *Cartography and Geographic Information Science* 40, 5, pp. 427-443.

Visvalingam, M. and J.D. Whyatt. 1992. *Line generalisation by repeated elimination of the smallest area*. Discussion Paper 10, Cartographic Information Systems Research Group (CISRG), The University of Hull.

**Barry J. Kronenfeld**, Associate Professor, Department of Geology and Geography, Eastern Illinois University, Charleston, IL 61920

**Lawrence V. Stanislawski**, Research Scientist, U.S. Geological Survey, Center of Excellence for Geospatial Information Sciences, Rolla, MO 65401

**Tyler Brockmeyer**, Computer Science Developer, Missouri State Technical University, Rolla, MO 65401

**Barbara P. Buttenfield**, Professor, Department of Geography, University of Colorado, Boulder, CO 80309-0260

# GeoVisual Analytics for the Exploration of Complex Movement Patterns on Arterial Roads

**Irma Kveladze and Niels Agerholm**

**ABSTRACT:** Visualization of complex spatio-temporal traffic movements on the road network is a challenging task since it requires simultaneous representation of vehicle measurement characteristics and traffic network regulation rules. Previously proposed visual representations addressed issues related to traffic congestion to explore movements along the road networks. However, those studies did not focus on the traffic safety aspects for the vulnerable road users in densely populated areas. This research deals with traffic safety issues for pedestrians on arterial roads located within the urban areas. Arterial roads are important for the mobility and connectivity of modern society, but they also have traffic regulations that are not always followed by the vulnerable road users. In order to understand complex movement behaviors between vehicle drivers and pedestrians on the arterial roads, a GeoVisual Analytics approach was developed in dialog with traffic experts. The exploratory interactive tools have assisted experts to extract unknown information about movement patterns from large traffic data at different levels of details. The results of the analysis revealed detailed patterns of speed variations and demonstrated the effectiveness of the proposed GeoVisual Analytics solutions.

**KEYWORDS:** GeoVisual Analytics, traffic movements, arterial roads, visual exploration

## Introduction

Road networks can be characterized as a hierarchy of different road types depending on capacity and function. At the top of the hierarchy are highways and motorways that provide uninterrupted movement for high-speed traffic flow. They are followed by the arterial roads with a primary task to carry large volumes of traffic and connect various destinations for urban and rural areas. Accordingly, their constant analysis is essential for optimized functioning and improved traffic safety.

The traffic domain distinguishes between free and controlled traffic flow on road network. However, despite freeways in some countries, highways and motorways are traffic controlled with speed limits, while urban arterial roads also are controlled by traffic lights, humps, pedestrian crossings, etc. Although arterial roads located in densely populated areas allow high-speed movements due to their curvature and design, traffic regulations are restricting vehicle drivers to low speeds between 30 - 60 km/h. Most research has been conducted on speed calming measures and speed control of vehicles from a safety perspective (Agerholm et al. 2016), but, not much has been researched on the Vulnerable Road Users' (VRU) behavior and its effect on the traffic flow and safety.

Even with strict regulations, specifically in densely populated areas, VRU ignore traffic rules on arterial roads that might lead to accidents. According to studies from the traffic domain, speeding remains an important worldwide problem that is considered as a large contributor to road trauma (Eksler et al. 2009; Elvik 2009). Besides, The European Commission (2010) also reports that in Europe alone during 2009 more than 1.500.000 people were injured in traffic and more than 35.000 died. Even though, The International Traffic Safety Data and Analysis Group (IRTAD) (2014) reported slightly improved traffic safety results over the past years only in high income countries, a report published by Statistics Denmark, show only an insignificant improvement. To the best of our memory, despite of the high importance, traffic safety of VRU has never been in the center of research from a cartographic perspective. The existing studies mainly are dealing with finding solutions for visual exploration of large complex traffic dataset to reveal traffic congestion patterns, but studies related to traffic safety are missing. Thus, to fill this missing gap and address traffic safety issues of VRU in urban arterial roads a number of research questions in close cooperation of traffic experts were defined. Accordingly, this research focuses on the investigation of where, when and how often do VRU cross the streets by neglecting traffic rules on arterial roads, and weather the vehicle drivers obey speed limit rules on the selected roads. To address these questions, Floating Car Datasets (FCD) collected over one year within Denmark were used. Based on the low speed patterns located relatively far from the traffic controlling elements, the study will be able to identify illegal pedestrian crossings. Besides, using a temporal aggregation approach, we also will be able to identify the median speed of the vehicles over the weekdays for each street.

The remainder of this research is organized as follows: in the next section we will discuss existing visual representations used for the analysis of large complex traffic network data. They are followed by the methodology section describing use cases studies, data collection methods and the deployed exploratory environment. After the methodology section, we present examples of constructed visual representations and derived results. Finally, we will conclude our findings in the conclusion section and provide recommendations for future work.

## Literature review

### ***Visualization solutions and analysis for traffic movements***

Visual analysis of large traffic dataset is an effective way for detecting movement patterns, and current visualization techniques do offer solutions to uncover where and when particular events take place (Andrienko et al., 2011). Understanding traffic phenomena and their proper functioning has been the study objective for Andrienko and Andrienko (2008). They proposed representation of flow maps for visual analysis of vehicle movements on a traffic network in Milan. Based on the spatio-temporal aggregation methods for massive traffic-oriented movements, they designed task oriented interactive visual representations. First, a mosaic diagram was developed to explore median speed variation patterns over the weekdays. Then, they proposed a directional bar

diagram to reveal movement directions of vehicles on roads. Later, in another piece of research Andrienko et al. (2011) introduced a generic visual analytics procedure to explore place-oriented patterns of events from movement data. Based on the occurrences of the spatial events relevant to the research interest, an event-based view was introduced that is suitable for the exploration of movement characteristics such as rhythmic processes and high-low speed. The suggested generic analytics procedure has been used for the analysis of dynamic processes of congestions on road and air traffic. Slingsby et al. (2008) also introduced new visualization methods of treemaps and road maps to address issues related to the representation of large multivariate movement data along the road network. The proposed treemaps offer a spatio-temporal ordering of a dataset resulting in summarized visual representations. Based on the overview - detail approach (Shneiderman 1996), representations simultaneously can display space, time and variable-constrained subsets of large complex traffic dataset for visual analysis.

Spatio-temporal analysis of movement trajectories along the road network was also the study interest for Tominski et al. (2012). The authors introduced a novel visualization method of a 2D / 3D trajectory wall implemented in a Space-Time Cube to support exploratory analysis of complex movements from a space, time and attribute perspective. The developed interactive trajectory wall reveals traffic network characteristics of congestions through color-coded trajectory bends, and facilitates the analysis of relevant tasks. Alike Tominski et al. (2012), Cheng et al. (2013) also developed visualization methods to reveal traffic congestions from road networks located in densely populated areas. In particular, they presented three 3D exploratory visualization techniques of isosurface, constrained isosurface and wall map implemented in a Space-Time Cube as well. Proposed representations have different advantages to explore data, thus, the authors recommended their integrated use. A visual analysis system was presented by Zuchao Wang et al. (2013) to investigate urban traffic congestion issues on the major roads of a traffic network. The developed visual analysis system consisted of multiple representation views of map, pixel table, route filtering, route flow and animation. The suggested visual analysis system allows multivariate exploration of traffic patterns and traffic jam conditions.

Cartographic literature offers various method for the visual exploration and analysis of traffic network data (Boyandin et al. 2011; Andrienko et al. 2011). The literature presented above are only a few of them that have been considered relevant to this research. They address traffic congestion issues and, therefore, propose different solutions suitable for the particular task analysis and visual exploration. Those studies specify the urge of use of combination of multiple interactive visual representations to derive meaningful knowledge and understand the relationships within multivariate spatio-temporal aspects of complex movement patterns in traffic network. However, the proposed solutions are dealing with visual exploration and analysis of traffic congestions and focus on the analysis of vehicle movements, while essential aspects related to the traffic safety are omitted. Therefore, differing from other studies, in this research we focus on the traffic safety of VRU on urban arterial roads that sometimes neglect traffic rules and evoke dangerous situations to a certain degree. On the other hand, we also analyse the speeding behavior of motorists regarding to the speed limitation for urban arterial roads.

## Method

### Use case study

In closer cooperation with domain experts five different street segments of Gugvej, Hadsundvej, Vesterbro, Kastetvej and Sankt Peders Gade located in different parts of Aalborg, Denmark were selected (Figure 1).



Figure 1: Location of the selected five streets in Aalborg, Denmark.

The arterial roads have been chosen based on two criteria; first, they have a high volume of functions that means many potential crossings and VRU, and second, ensure passability in different parts of the city. For instance, Kastetvej and Vesterbro are located in the city center and compared to other three streets are busy over the course of a day. Gugvej and Hadsundsvej are located in the eastern and southern part of the city, although, as arterial roads they do connect peripheral parts of the city but are less busy compare to the previous two roads. Sankt Peders Gade is located in the northern part of Aalborg and represents an important arterial road since it connects city to the airport. Except Gugvej, all streets segments do contain traffic controlling elements such as traffic signals, zebra crossings, bumps, etc. to ensure smooth movement regulation on the streets. In addition, all selected streets have speed limits of 30 - 60km/h (see Table 1, below).

Table 1: Streets used in this article and their parameters relevant to the study.

<i>Street name</i>	<i>Speed limit (km/h)</i>	<i>Length of the street segment (m)</i>	<i>Number of traffic controlling elements</i>
Sankt Peders Gade	30	647,11	7
Kastetvej	50	569,77	2
Vesterbo	50	680,11	7
Hadsundvej	50	480,15	1
Gugvej	60	440,11	-

### ***Data collection and characteristics***

The primary goal of our research was to understand and report the complex traffic functioning of arterial roads in densely populated urban areas. For this purpose, Floating Car Datasets (FCD) collected from 425 vehicles during 2014 within Denmark was used (Tøfting et al. 2014). The data was gathered through On-Board Unit (OBU) tracking devices transmitting information on location, time, actual speed, movement direction, and various related technical parameters of the vehicles into a database system (Rashidi et al., 2012). Generated FCD was scattered around the road network and required further processing for matching to the real-world traffic network. To do so, map matching techniques widely used in traffic network studies for examining vehicle movements were used. Map matching is a technique that combines an automatically collected GPS, FCD, etc. dataset with a digital electronic map to correct for spatial errors originating from the GPS units. It is one of the essential steps in traffic network data processing and its implementation parameters for pattern recognition defines accurate vehicle positioning in traffic (Greenfeld 2002; Quddus et al. 2007; Jagadeesh et al. 2004; Tradisauskas et al. 2009). Therefore, based on the tracking ID, the time sequence of the trajectory and movement direction of coordinates, FCD was map-matched to the digital road network derived from the OpenStreetMap (see Figure 2 next page).

In this particular case, Mapillary's freely available mapmatching (MapMatching) algorithm based on PostgreSQL, Postgis and pgRouting was adapted. The Map-matched FCD required further data cleaning from irrelevant or erroneous information. First, the data outside the spatial range of selected road segments were disregarded from database. Then, the recordings not including temporal information, speed or movement direction of the vehicle were filtered and removed. Furthermore, based on the movement direction of the vehicles, the dataset was divided in to two directional movements that correspond to the real time movement direction on the roads. And lastly, for better understanding the overall median speed variation on the roads, a one-year dataset was aggregated first based on the weekdays and then based on the months. At the same time the total number of

vehicle trips in per weekday over the months were summed up. The processed dataset was further used to investigate speed variation patterns along traffic network.

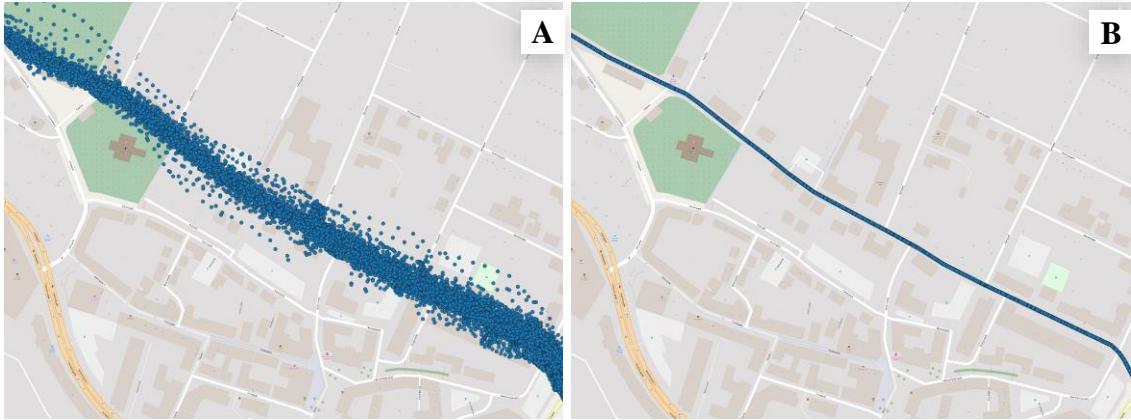


Figure 2: Result of the map-matching of FCD to road network. FCD before (**A**) and after (**B**) map-matching.

### *Applied visualizations*

To reveal complex speed-flow relationships, an interactive GVA environment allowing multilevel exploration of FCD was constructed (Shneiderman, 1996; Keim et al., 2008). The dataset contains the variables of significant importance for the traffic safety study, and to analyze them the visual representations of time cartograms and speed profile graphs were integrated in an exploratory interface with a two-dimensional map view. The rectangle time cartograms were constructed for the analysis of temporal distribution of median speed on roads over the weekdays in relation to the appointed speed limit regulations (see Table 1, above). Besides, cartograms were also used to reveal the total amount of the generated daily trips in order to judge the results of the analysis and make conclusions.

The other representation developed were speed profile graphs for two directional movement analyses of vehicles on arterial roads. Alike Keim et al. (2002), we also visualized complex traffic network data without further aggregation. Based on the pixel bar chart idea to present each data value through a single pixel, we have constructed profile graphs using the distance of the streets and displayed each speed record over time using color code. This approach helped not only displaying detailed speed characteristics but also keep the location of a traffic elements on the streets. The types of traffic controlling elements also were encoded into the color. Since each point derived from FCD was carrying valuable information on speed and other technical parameters. The other approach we used for street profile graph was a mosaic diagram (Andrienko et al. 2008). Differing from the mosaic diagram, in our graph the horizontal axes represents the space with location of traffic controlling elements and FCD, while vertical axes show time information as generated initially, which allows to make a selection of different temporal granularities. The developed representations were integrated in the GVA environment.

The exploratory GVA environment supports recognition of speed change within the traffic flow and establishes correlation with traffic controlling elements. The accurate examination of speed-flow relationship for each segment of selected arterial roads, will lead to revealing the critical locations where pedestrians use to cross the roads and neglect traffic regulation rules. For instance, based on the interrupted speed flows detected on speed profile graphs, traffic experts can determine the location of unexpected pedestrian crossings on the map view.

## Results

### ***Temporal distribution of traffic movements on arterial roads***

The exploration of traffic data showed some differences between lifestyle of the streets. For instance, the time cartogram for *Sankt Peders Gade* shows no difference in the traffic volume for the movements toward north-west and toward south-east (Figure 3 next page), however, there is an obvious difference in median speed distribution between two directions, the vehicles do exceed speed limitation of 30 km/h mostly while are traveling towards north. This trend could be related to the traveling from work place to home and also traveling to the airport, since *Sankt Peders Gade* connects the city to the airport. The cartogram also shows differences in trip distributions over the weekdays. In particular, weekdays appear to be busier than weekends, but, weekends can be distinguished with more speed violation than any other weekday. Differing from *Sankt Peders Gade*, no difference between moving south-east and north-west were found on *Kastetvej*. The only obvious difference was found in the distribution of the activities over the weekdays. Sundays here appeared to be as busy as other weekdays, however, Saturdays show less movement activities. In addition, no speed violation was detected - here speed limit is 50km/h - while results of the analysis detected only a median speed of 39,3km/h. On *Gugevej* the traffic volume showed similar patterns as the above mentioned two streets. Alike on *Sankt Peders Gade*, Sundays and Saturdays were less active, besides no difference in traffic volume was found between north-west and south-east. The speed limit on this street is 60km/h and the time cartogram revealed that the median speed of the vehicles did not exceed 46 km/h.

The vehicle movement volume on *Hadsundsvej* also looks similar to the other streets. No difference between movements towards northwest and southeast were found, however, there is a difference in the median speed distribution of the cars. In particular they have the opportunity to move towards southeast with slightly higher speed. The reason could be fewer traffic jams during rush hours towards southwest, while towards northwest, i.e. towards city traffic jams are more and vehicles move slow. This resulted in a median speed that does not exceed 34 km/h when the real speed limitation is only 50 km/h. This trend of the median speed steadily reappears for almost every weekday. *Vesterbro* is located in the center of the city, and the analysis of the datasets revealed some differences in traffic volume compare to the above discussed arterial roads. Besides, no differences were detected in movement direction towards north and towards south on *Vesterbro*. The distribution of the traffic volume over the weekdays appeared to be similar. Like the arterial roads discussed above, the median speed limitation also did not exceed 39 km/h

when the speed limit is 50 km/h. The reason could be the large traffic volume on *Vesterbro* that does not allow high-speed movement for vehicles.

Sankt Peders Gade - towards North - West



Sankt Peders Gade - towards South - East

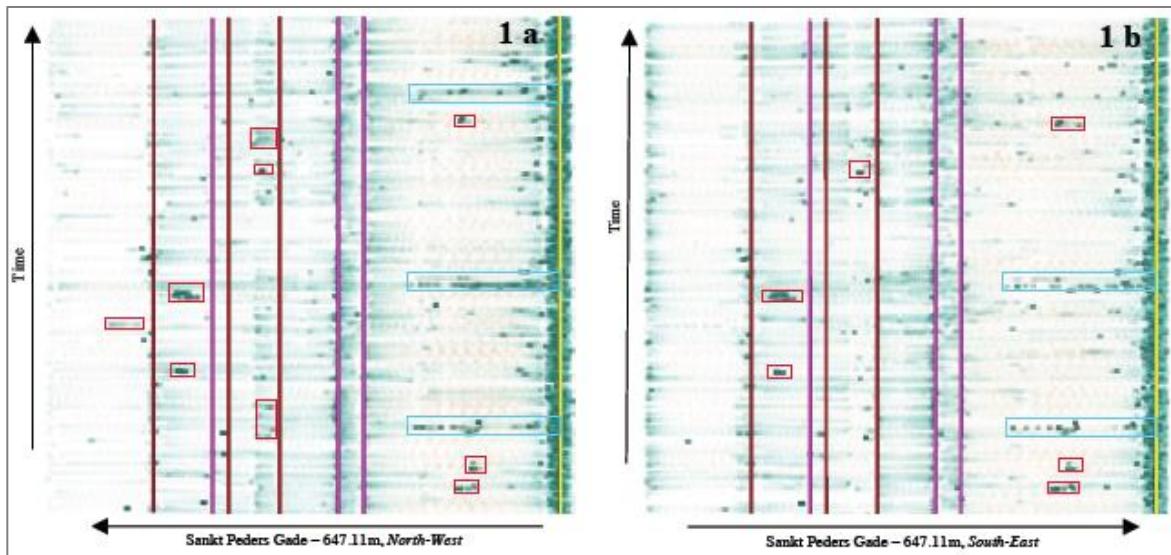


Figure 3: A time cartogram representing the temporal distribution of vehicle movements and their median speed on Sankt Peders Gade. The big rectangles represent weekdays and small rectangles inside 12 months. The size of the rectangle indicates the amount of the trips: min 7 – 14 and max 45 – 65 per weekday over the months. The color saturation of the rectangles represents median speed. The darker the saturation the higher the median speed and vice versa.

This section shows that arterial roads do play important role in the connectivity and associability of the different parts of the city. The incoming traffic flow during rush hours from the local roads, influences their passability which results in a low median speed. The morning and evening rush hours have increased traffic volume that does not enable high speeds. In a result, the allowed speed limit exceeds the real median speed of the traffic movement on arterial roads.

## The speed interruption patterns

Temporal distribution of the speed patterns along the streets was one of the key aspects for revealing places of illegal intersections. The street profile graphs shown on Figure 4, below reveal vivid differences between low - high distribution on the streets. The speed distribution of *Sankt Peders Gade* shows very few places of illegal traffic crossing activities performed by pedestrians. In the majority of the cases low speed locations coincide with the location of bumps, and pedestrian crossings. It is to be noted, that alike discussed in the section above, there is a difference between speeding patterns moving towards north-west and south-east. The low speed patterns towards north-west appear more often and also show more places of long traffic jams than the south-east part. Differing from *Sankt Peders Gade*, the speed interruption patterns on *Kastetvej* are more intense. They mainly appear between pedestrian crossing and the traffic light zone. Similar to *Sankt Peders Gade*, here is also some difference between both sides of roads. In particular, the south-east direction slightly exceeds with speed interruption patterns. The speed interruption patterns on *Vesterbro* were more than on the previous two streets, but the reason here could be the slow movement of the traffic and five traffic lights within a stretch of 680,11m. The representation also revealed places of traffic jams. The slow speed of the traffic did not allow to detect vivid patterns of the pedestrian illegal crossings. The low speed patterns also appeared to be the same for north and south direction movements. Alike *Vesterbro*, *Hadsundsvej* also shows intensive low speed variations between traffic lights and pedestrian crossing. This representation revealed obvious patterns of low speeds at the traffic lights. Moreover, the north-west direction appeared to be with more speed interruption patterns than south-east. This obviously indicates illegal speed interruptions by pedestrians who cross the street not very far from the zebra pedestrian crossing. Differing from other streets *Gugvej* showed some additional aspects influencing speed interruptions.



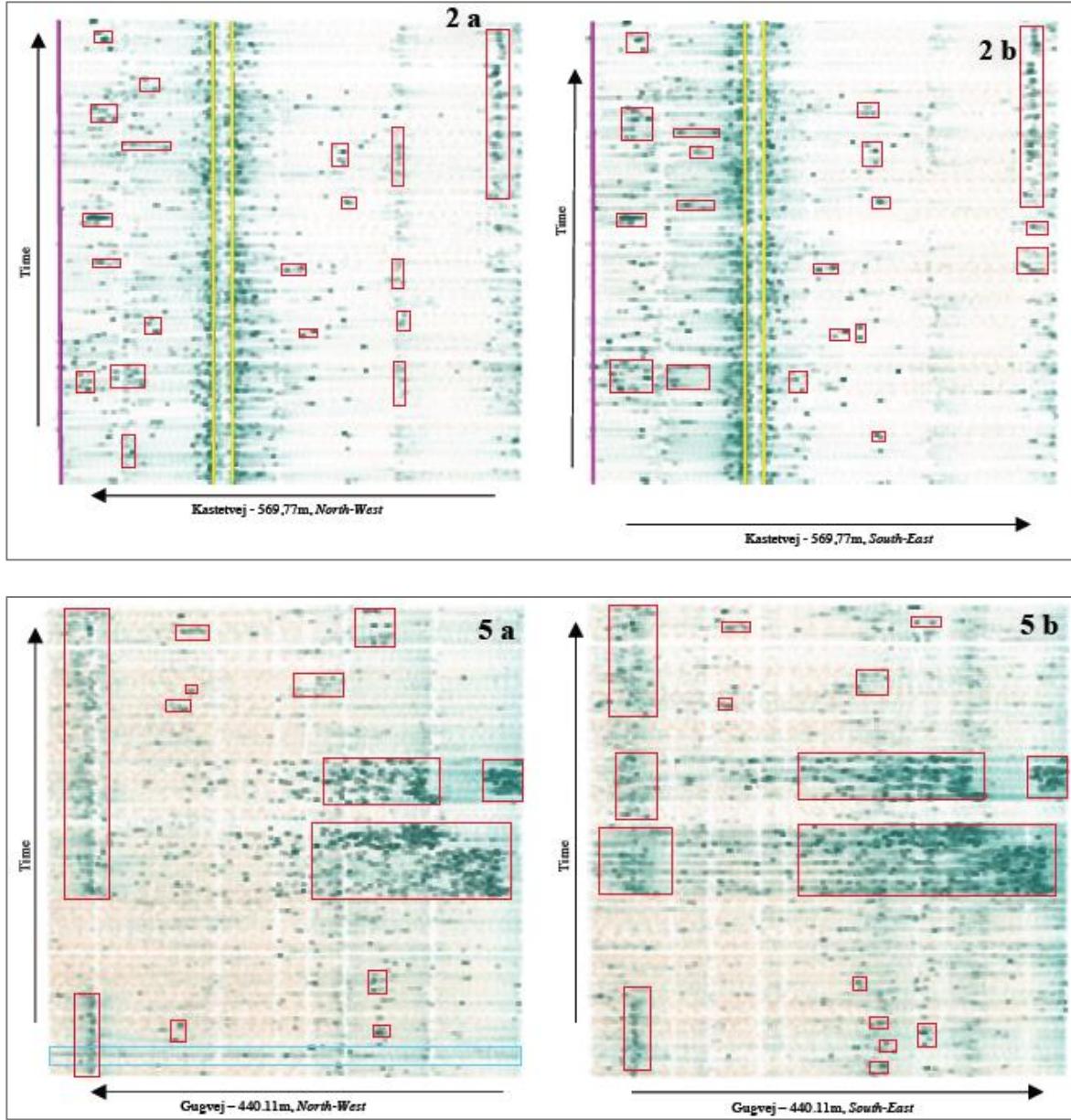


Figure 4: Examples of street profile graphs representing the temporal distribution of speed along the streets with speed limits **1.** 30 km/h, **2.** 50km/h and **5.** 60 km/h over the year. On the representations is applied transparency to detect places of the speed interruptions in red and traffic jams in cyan rectangles. **a.** movement directions towards *North -West*. **b.** movement directions towards *south-east*. The vertical axis represents time and the horizontal axis street distance. The colours on the graph show speed variation from high to low, the dark green indicates low speed, and brown indicates high speed. The stationary lines show the location of the speed controlling elements: brown – bumps, yellow – traffic signal and pedestrian crossings and purple – zebra crossings.

The representation on Figure 4 shows two long zones of the massive traffic congestion. The investigation showed that the reason was local traffic work performed by the city municipality who replaced some pipes located underground. This roadwork lasted for quite some months and over this time period it is hard to detect nor to the behavior of

pedestrians regarding to illegal crossings. Despite of it, we still could find some places with vivid speed interruptions that appeared over time at the same place.

The location of the arterial roads was one of the main influential factors of the intensity of the movement activities. *Vesterbro* that is located in the center of the city appeared to be more active than other streets located further away from the city center. The representations also show some rhythmic processes in traffic jams and illegal pedestrian crossings at particular locations. The examination of the street segments also showed the decrease in the speed patterns at the segments when drivers approached and departed traffic controlling elements. A sudden drop in speed between traffic controlling elements shown on street profile graphs, primarily indicate interrupted driving activities by VRU. According to studies in the traffic domain, VRU do cross streets unexpectedly during rush hours by ignoring traffic rules. Thus, the exploratory environment proposed in this research, assisted experts to perform behavioral characterization, comparison, and search tasks to examine traffic movement patterns. By establishing the link between visual representations, the users could interact with date characteristics and extract information at the different level of details.

## Conclusions

Modern technological achievements in positioning and communication systems allow more accurate tracking and recording of information for various transport modes. This information offers extensive knowledge on whereabouts and various technical parameters of tracked vehicles for comprehensive exploration. To make sense of such information suitable visual representations and tools for the analysis are needed. To do so, in this research we presented an exploratory analytical environment to analyze FCD on arterial roads located within densely populated urban areas with many functions with related crossing activities for VRU. Using FCD to investigate congestions and detect queue, became the prevailing way in traffic and transportation domain. Based on the research questions, FCD revealed changes in the distribution of speed over space and time on arterial roads. The overall temporal distribution of movement show difference in activities for weekdays over the months (Figure 3, above). While the detailed examination of daily rhythms, showed the influence of the day light on the driving behavior of the vehicles. The time cartogram also revealed some similarities and differences over weekdays in the distribution of traffic volume and their median speed. The knowledge derived from data analysis may help scientists in the transportation domain to gain an extensive understanding on various movements patterns in complex traffic networks. Besides, it also reveals speeding related issues in traffic that supports domain experts in better planning and design of road network.

Highly interactive visual representations were sufficient to investigate traffic related aspects from location, attribute and time perspective and facilitate exploration process. The proposed solution demonstrated the advantage of GeoVisual Analytics tools for knowledge extraction from complex traffic movement data. Therefore, by processing all relevant information in one interactive analytical environment traffic experts get support

for a decision making for traffic safety regulations with respect to future development and recommendations.

## Acknowledgements

Many thanks to Bernhard Snizek, SWECO Denmark A/S for his valuable help.

## References

- Agerholm N, Knudsen D, Variyeswaran K (2016) Speed-calming measures and their effect on driving speed – Test of a new technique measuring speeds based on GNSS data. *Transportation Research Part F: Traffic Psychology and Behaviour*.
- Andrienko G, Andrienko N (2008) Spatio-temporal aggregation for visual analysis of movements. In: *2008 IEEE Symposium on Visual Analytics Science and Technology*. IEEE, pp 51–58
- Andrienko G, Andrienko N, Hurter C, et al (2011) From movement tracks through events to places: Extracting and characterizing significant places from mobility data. In: *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, pp 161–170
- Andrienko N, Andrienko G, Pelekis N, Spaccapietra S (2008) Basic Concepts of Movement Data. In: Giannotti F, Pedreschi D (eds) *Mobility, Data Mining and Privacy*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 15–38
- Boyandin I, Bertini E, Bak P, Lalanne D (2011) Flowstrates: An Approach for Visual Exploration of Temporal Origin-Destination Data. *Computer Graphics Forum* 30:971–980.
- Cheng T, Tanaksaranond G, Brunsdon C, Haworth J (2013) Exploratory visualisation of congestion evolutions on urban transport networks. *Transportation Research Part C: Emerging Technologies* 36:296–306.
- Eksler V, Popolizio M, Allsop R (2009) *How far from Zero?* European Transport Safety Council , Brussels
- Elvik R (2009) *The Handbook of Road Safety Measures*. Emerald
- European Commission. (2010). *Towards a European road safety area: policy orientations on road safety 2011-2020*. Brussels. Retrieved from COM(2010) 389 final
- Greenfeld JS (2002) Matching GPS Observations to Locations on a Digital Map.

*Transportation Research Board 13*

Map Matching, [https://github.com/mapillary/map\\_matching](https://github.com/mapillary/map_matching)

International Traffic Safety Data and Analysis Group (IRTAD) (2014) Road Safety Annual Report 2014. *OECD Publishing*, Paris, France

Jagadeesh GR, Srikanthan T, Zhang XD (2004) A Map Matching Method for GPS Based Real-Time Vehicle Location. *Journal of Navigation* 57:429–440.

Keim DA, Hao MC, Dayal U, Hsu M (2002) Pixel Bar Charts: A Visualization Technique for Very Large Multi-Attribute Data Sets. *Information Visualization* 1:20–34.

Quddus M a., Ochieng WY, Zhao L, Noland RB (2003) A general map matching algorithm for transport telematics applications. *GPS Solutions* 7:157–167.

Quddus M a (2006) High Integrity Map Matching Algorithms for Advanced Transport Telematics Applications. *Centre for Transport Studies Department of Civil and Environmental Engineering*, Imperial College London, United Kingdom 270

Quddus MA, Ochieng WY, Noland RB (2007) Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies* 15:312–328.

Shneiderman B (1996) The eyes have it: a task by data type taxonomy for information visualizations. In: *Proceedings 1996 IEEE Symposium on Visual Languages*. IEEE Comput. Soc. Press, Boulder, Colorado, pp 336–343

Slingsby A, Dykes J, Wood J (2008) Using treemaps for variable selection in spatio-temporal visualisation. *Information Visualization* 7:210–224.

Tøfting S, Lahrmann H, Agerholm N, et al (2014) ITS Platform: Development of intelligent traffic. In: *10th ITS European Congress*. ERTICO - ITS Europe, Helsinki, Finland, p 9

Tominski C, Schumann H, Andrienko G, Andrienko N (2012) Stacking-Based Visualization of Trajectory Attribute Data. *IEEE Transactions on Visualization and Computer Graphics* 18:2565–2574.

Tradi. auskas N, Juhl J, Lahrmann H, Jensen CS (2009) Map matching for intelligent speed adaptation. *IET Intelligent Transport Systems* 3:57.

Zuchao Wang, Min Lu, Xiaoru Yuan, et al (2013) Visual Traffic Jam Analysis Based on Trajectory Data. *IEEE Transactions on Visualization and Computer Graphics* 19:2159–2168.

**Irma Kveladze**, Postdoc researcher, Department of Civil Engineering, Aalborg University,  
Thomas Manns Vej 23, 9220 Aalborg, Denmark

**Nils Ageholm**, Associate Professor, Department of Civil Engineering, Aalborg University,  
Thomas Manns Vej 23, 9220 Aalborg, Denmark

# Incorporating Changes in Multi-scale Databases

**Dan Lee, Nobbir Ahmed and Iffat Chowdhury**

**ABSTRACT:** Many GIS organizations and national mapping agencies (NMAs) build and maintain multi-scale databases. It is important for them to keep the data accurate and up to date for the intended geographical analysis and mapping purposes. Typically changes need to be incorporated into the master database, that is, the largest scale database; and the affected areas are to be incrementally updated across smaller scale databases. However, these organizations and agencies may face a few challenges: (1) features in their master databases don't always line up with those they acquire from other sources; and (2) the lack of linkages for corresponding features within the multi-scale database makes the propagation of changes difficult. Too often we see that road data owned by a city government or an NMA may differ in shape, position, and level of detail from what they obtained through their state government or another source; rivers or other features along the borders of neighboring areas are inconsistent or disconnected. The increasing needs for change detection, data harmonization, and linkages in multi-scale databases for incremental updating remain.

Solving the above problems requires a unique process that identifies or matches corresponding features in various data sources and reconciles the differences for the best accuracy, completeness, and consistency. This process is known as conflation. In the past few years a set of conflation tools has been developed based on common use cases and requirements and released in ArcGIS Desktop. Given the wide range of data varieties and complexities, the automatic tools may not produce 100% correct results. Our efforts have also been devoted to building and testing workflows that can facilitate post evaluations and corrections as efficiently as possible.

This paper highlights a few typical conflation scenarios and presents the processing workflows and results. Some of the workflows involve spatial adjustment to bring spatially inconsistent data together, for example to align outdated parcel data to newly available and more accurate parcel data. Other workflows transfer attributes between correspondent features for the purposes of enriching information and establishing linkages for multi-scale databases, for example to transfer road names from legacy road data to new GPS data or to transfer unique identifiers from large scale buildings to generalized buildings at smaller scales as their linkages. Conflation plays indispensable roles in data reconciliation, change detection, and incremental database updating. Changes in newly updated master databases can be propagated to smaller scales through linkages and generalization of the affected areas. Our future focuses aim at enhancing the existing tools and formalizing workflows to meet the growing demands for conflation.

**KEYWORDS:** Change detection, conflation, feature matching, multi-scale, database updating

**Dan Lee**, Senior Product Engineer, Analysis and Geoprocessing, Esri Inc., Redlands, CA 92373

**Nobbir Ahmed**, Product Engineer, Analysis and Geoprocessing, Esri Inc., Redlands, CA 92373

**Iffat Chowdhury**, Software Developer, Analysis and Geoprocessing, Esri Inc., Redlands, CA 92373

# Viewshed Analysis for UAS Flight Planning

**Samuel Levin and May Yuan**

**ABSTRACT:** In this project, we developed a mapping algorithm to identify observer locations in planning for Unmanned Aircraft System (UAS) surveys. This algorithm can be used to ensure safety of all participants and bystanders during active surveying and to maintain compliance with Federal Aviation Administration regulations for safe drone operation. These regulations require that ground crews maintain visual line-of-sight with the drone during flight. This safety precaution is especially relevant when flying in densely populated areas with visual obstacles, like buildings and tall vegetation. We used the University of Texas at Dallas (UTD) as a study site for the development of this algorithm as part of our smart campus project. The UTD campus covers 200 hectares with clusters of buildings and facility structures assembling an urban landscape. Using lidar data, we developed a digital surface model of UTD that details the structure of buildings, tall trees, and other obstructing features. This model was partitioned into survey areas to accommodate UAS flight time constraints. Based on the digital surface model, we applied viewshed analysis in relation to the planned flight paths to examine potential observer locations and cumulative viewshed across each survey area. For each area, the number and locations of observation stations necessary to maintain constant visual contact with the drone throughout its flight were identified. This was accomplished using a Python processing algorithm, which selects ideal observers from valid ground locations visible along the flight path until total coverage is achieved. The resulting algorithm determines the minimum number of stations required to maintain the line-of-sight and the locations of these stations. Additionally, the portion of the drone's flight path visible to each ground observer is mapped, delineating zones of observer responsibility. These findings assure an effective distribution of ground crews to maintain UAS safety and FAA compliance. Comparable lidar survey data are now available in many urban areas where maintaining visual line-of-sight presents a critical issue for UAS surveys. This algorithm may facilitate the planning and assessment of observation stations in future UAS surveys using similar datasets. This assessment may be used as a supplement to FAA waiver and airspace authorization applications and reinforces the safety precautions that UAS operators should undertake when requesting such exemptions.

**KEYWORDS:** UAS; Viewshed; Smart Campus; Python; Lidar

**Samuel Levin**, M.S. Student, Department of Geospatial Information Sciences, The University of Texas at Dallas, Richardson, TX 75080

**May Yuan**, Ashbel Smith Professor, Department of Geospatial Information Sciences, The University of Texas at Dallas, Richardson, TX 75080

# Using Deep Learning and Google Street View Images to Quantify the Shade Provision of Street Trees in Boston, Massachusetts

Xiaojiang Li and Carlo Ratti

**ABSTRACT:** Urban areas are the places of mass interactions between human and nature, and homes to a large proportion of the global population. In 2008, more than half of the global population lives in urban areas, and the number is keep increasing rapidly. Making our cities more livable and sustainable becomes more and more important. Thermal comfort plays an important role in determining the quality of life in cities. Study on how to increase thermal comfort levels in urban spaces, particularly under the contexts of global warming and rapid urbanization, has become a prominent topic in urban studies and of key interest to decision makers. Shade provision by street trees during the hot summer months is a primary factor for the thermal comfort of people in urban areas. As such, quantitative information on the ecosystem services provided by street trees, particularly in terms of potential temperature reductions, would provide an important reference for urban greening projects in order to maximize those services.

Traditionally, remotely sensed data has been widely used to measure the performance of urban heat island mitigation. However, such overhead view information cannot necessarily reflect how much solar radiation is blocked by street trees, as it lacks the vertical dimension needed to estimate or model the path of light from the sun; and hence shading. In this study, we proposed to use Google Street View (GSV) panorama to estimate the shade provision of street greenery. Hemispherical images, which provide the bottom-up view of street canyons, were generated from GSV panorama based geometric transform and used to quantify the shade provision of street trees. The bottom-up view hemispherical images help us to study how the solar radiation reaching the ground, which would provide very important tools for us to understand the energy balance in street canyons. We further used the state-of-the-art deep learning algorithms to segment street level images into vegetation, sky, and building pixels. Based on the quantitative information derived based on mathematical analyses of GSV panoramas, we further estimated the spatio-temporal distribution of sky view factor (SVF) and the sunlight duration at the street canyon level. We further mapped and analyzed the influence of street enclosure on solar radiation reaching the street canyons by estimating the sunlight duration in urban street canyons. The results show that street trees help to decrease the SVF by 24.61% in Boston, Massachusetts.

This study showed that that GSV is a very promising data source for urban studies considering its public accessibility and global availability. The developed workflow for quantifying the shade provision of street trees in this study is totally automatic and without any human intervention. Therefore, it is possible to simply and rapidly estimate the SVF and sunlight duration at street level for any city with GSV service available. The GSV would also be a surrogate for those study areas with high-resolution digital city models not available. Other researchers may find the method illustrated in this study is directly deployable for different studies related to urban form analysis. The results of this study would shed new light on future urban studies using the publicly accessible and globally available GSV data.

**KEYWORDS:** Deep learning, Google Street View, thermal comfort, shade provision, street trees.

**Xiaojiang Li**, Postdoc Fellow, MIT Senseable City Lab, Cambridge, MA 02139

**Carlo Ratti**, Professor, MIT Senseable City Lab, Cambridge, MA 02139

# Geospatial Machine Learning: Predicting Accident-Prone Road Segments Using GIS and Data Mining

Xiao Li, Daniel W. Goldberg, Tracy Hammond and Xingchen Chen

**ABSTRACT:** Traffic crashes have become the seventh leading cause of preventable death in the United States. Existing geostatistical methods work well for identifying and visualizing existing high-risk traffic zones (“black” zones) from historical crash data. However, geostatistical modeling alone cannot be used to predict driving risk for newly-constructed roads. This study reports on the use of data mining techniques for estimating a priori crash risk from road-related features. The performance of this approach is compared to existing geostatistical methods for identifying and predicting accident-prone road segments. A case study was conducted in Polk County, Iowa, with historical crash data collected between 2011 and 2016. The results demonstrate that while geostatistical methods can be highly effective for determining crash intensity on road segments, data mining techniques enable researchers and practitioners to accurately predict crash risk based on carefully chosen road related features.

**KEYWORDS:** Crash Analysis, Accident-Prone Road Segments, Data Mining, Geostatistical Methods.

## Introduction

Traffic injuries are one of the most severe public health problems, which not only cause a lot of deaths, injuries, and property damages but also produce a significant economic loss (World Health Organization, 2015). Studies have demonstrated that traffic accidents are not randomly scattered. There are some circumstantial relationships between road-related features and accident occurrence (Al Haji, 2005; Effati et al., 2012). Based on this hypothesis, various studies have been conducted for facilitating traffic safety (Evans, 2004; Goodchild, 2015; Olusina and Ajanaku 2017; Yao et al., 2015).

Using GIS-based methods for crash analysis dates to the 1970s (Moellering, 1976). Hotspot analysis and Kernel Density Estimation (KDE) are two of the most commonly used geostatistical methods for identifying traffic “black” zones. The KDE is a powerful solution to estimate the probabilistic density of crashes for road segments and to study the spatial patterns of car accidents (Anderson, 2009; Mohaymany et al., 2013; Thakali et al., 2015). Moran’s I and Getis-Ord Gi\* function can be used to investigate the spatial autocorrelation of car crashes and map “hot spots” (Prasannakumar et al., 2011; Songchitruksa and Zeng, 2010). However, these methods are not efficient to forecast the driving risk for newly-constructed roads. Studies have demonstrated that data mining techniques can investigate the linkage between the accident-related factors (i.e. type of car, weather condition, driver’s age et al.) and car accident severity (e.g. “slight injury”, “severe injury”, “fatal” et al.), which has the potential to help with crash risk prediction

without using historical crash data. (Beshah and Hill, 2010; Effati et al., 2015; Kumar and Toshniwal, 2015; Shah et al., 2017).

This study takes advantage of data mining techniques and focuses on investigating the correlation between driving risk and road-related features. Compared to previous studies, this paper has made new contributions for better understanding and identifying reasons for the occurrence of car crashes: 1) different geostatistical methods and data mining techniques are compared and evaluated for accident-prone road segment identification; 2) the relationship between driving risk and road-related features are innovatively investigated; 3) a detailed temporal variation analysis of crashes is included.

## Method

### *Geostatistical Methods*

Hotspot analysis (Getis-Ord Gi\*) can identify the significant spatial clusters of high values (hot spots) and low values (cold spots) by calculating the Getis-Ord Gi\* statistic. In this study, Emerging Hot Spot Analysis is implemented to investigate spatial-temporal trends of the crash variation (Harris et al., 2017). Crashes are restructured as space-time bins based on their time stamp and location. The Getis-Ord Gi\* statistic is calculated for labeling each bin as a “hot spot” or “cold spot.” The Mann-Kendall trend test is performed to highlight the changing of “hot spots” over time. The KDE is used to estimate the density of features in a neighborhood around those features. A continuous density surface will be generated and placed over each car crash. The density of each grid is calculated by overlapping and summing the crashes’ density surfaces, which can be used to extract the crash intensity road segments (Anderson, 2009; Thakali et al., 2015).

### *Crash Rate Assessment*

Crash intensity road segments are mapped well using geostatistical methods, however, “high-risk” road segments are not identified. Crash risk is defined as “the number of crashes compared to the level of exposure.” In this study, crash rate assessment is adapted from (Federal Highway Administration, 2011), which is calculated as:

$$R = \frac{C \times 100,000,000}{365 \times N \times V \times L}$$

where  $R$  = crash rate of a road segment defined as “crashes per 100 million vehicle-miles of driving”;  $C$  = number of crashes occurred along a road segment;  $N$  = number of years;  $V$  = Average Annual Daily Traffic (AADT) volumes;  $L$  = road segment length in miles.

### *Data Mining Methods*

Data mining methods have been widely used to discover patterns in a massive volume of data, which can promote the understanding of relationships between crash rate levels (dependent variable) and road-related features (independent variables) (Feeelders et al., 2000). In this study, various road-related features are generated for each road segment as

independent variables. Meanwhile, crash risk levels are labeled based on the calculated crash rate. Data preprocessing is performed to handle missing data, remove irrelevant attributes and transform data format. Correlation-Based Feature Selection is used to select the most relevant features. Three data mining methods: Decision Tree, Naïve Bayes, and Support Vector Machine (SVM) are implemented to predict the crash risk for road segments.

## Results

In this paper, 45,898 car accidents of Polk County, Iowa between 2011 and 2016 were snapped to their nearest road segments. Emerging Hot Spot analysis was implemented to test the spatial autocorrelation and temporal variation trend of car accidents at different spatiotemporal scales. Crashes were restructured into space-time hexagon bins. Getis-Ord  $Gi^*$  statistic was calculated for each bin (e.g., 1-mile hexagons with 12-week) for identifying hot spots. The temporal trend of hot spots was assessed by performing the Mann-Kendall trend test on the time-series hexagon layers, as shown in Figure 1.

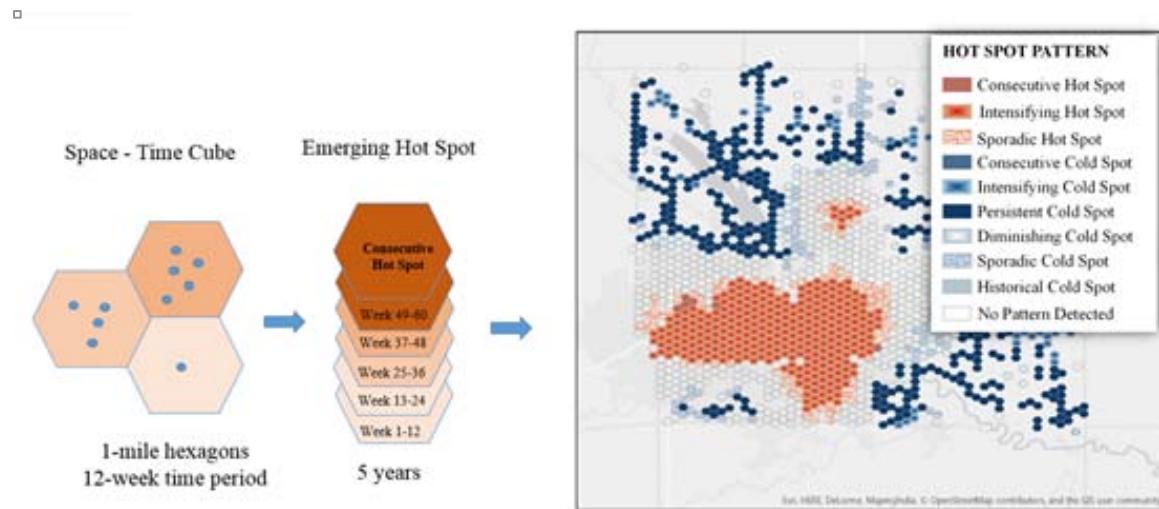


Figure 1: An example of emerging hot spot analysis result (Spatial scale: 1-mile hexagon; Temporal scale: 12-week).

Kernel Density Estimation was performed to calculate the crash density for each road segment. Since there is no fixed threshold to define crash intensity road segments, road segments were evenly divided into ten different levels (Level 1 – Level 10) based on their estimated values, as shown in Figure 2. The higher the level of the road segment labeled (i.e., Level 10), the more dangerous it was.

Twenty-six road-related features were selected and aggregated as the attributes of each road segment. Crash rate was calculated to present the driving-risk for each road

segment. Regarding crash rate, road segments were categorized as “Safe,” “Low-Risk,” and “High-Risk.” 6,324 road segments with 26 attributes, including 25 road-related features (independent variables) and labels (the dependent variable), were fed to a data-mining engine for predicting the driving-risk levels of each road segment. Five-fold cross-validation was performed to assess the performance of models.

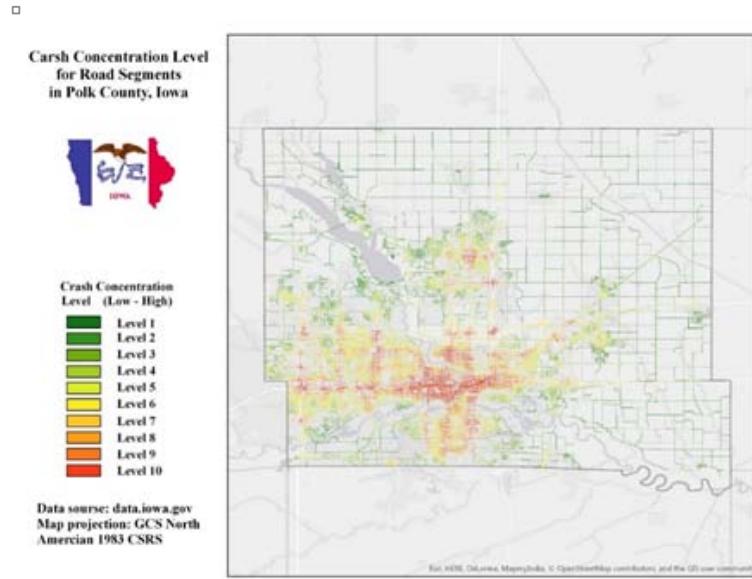


Figure 2: Kernel density estimation for identifying crash intensity road segments.

Overall, results demonstrate that data mining methods are effective for labeling the road segments as “Safe,” “Low-Risk,” and “High-Risk” based on the selected features as shown in Table 1. Of the three methods tested, Decision Tree has the highest accuracy of 82.82%. Additionally, the “IF-THEN” rules generated from the Decision Tree can help identify the road-related leading factors of crash rate (e.g., if  $\text{NUMLANES} \leq 3$  and  $\text{TYPEPARK} = \text{Parallel Parking}$  and  $\text{IRI} < 26$  then Safe).

Table 1: Performance of three classifiers.

	<i>Decision Tree</i>	<i>Naïve Bayes</i>	<i>SVM</i>
Accuracy	82.823%	79.553%	80.033%
Kappa	0.658	0.666	0.640

As shown in Table 2, a large portion of crashes occurred around road intersections. To

eliminate the impact of road intersections, crashes within a 5-meter-buffer were removed, additional road-related features were generated. By doing so, the accuracy of Decision Tree can be improved to 85.03%.

Table 2: Crashes occurred in the intersection buffers.

<i>Buffer</i>	<i>3 Meters</i>	<i>5 Meters</i>	<i>10 Meters</i>
# crashes	18630	20870	22718
% crashes	40.768%	45.669%	49.713%

The temporal trend of car crashes was also conducted. The result shows that car crashes significantly increased over the past five years in Polk County. Winter months (Dec, Jan, and Feb) have more crashes, which may be caused by poor weather conditions and fewer daylight hours. Also, weekdays have more crashes than weekends. 7 am, 4 pm and 5 pm are the top 3 high-risk hours for driving due to high traffic volume, as shown in Figure 3.

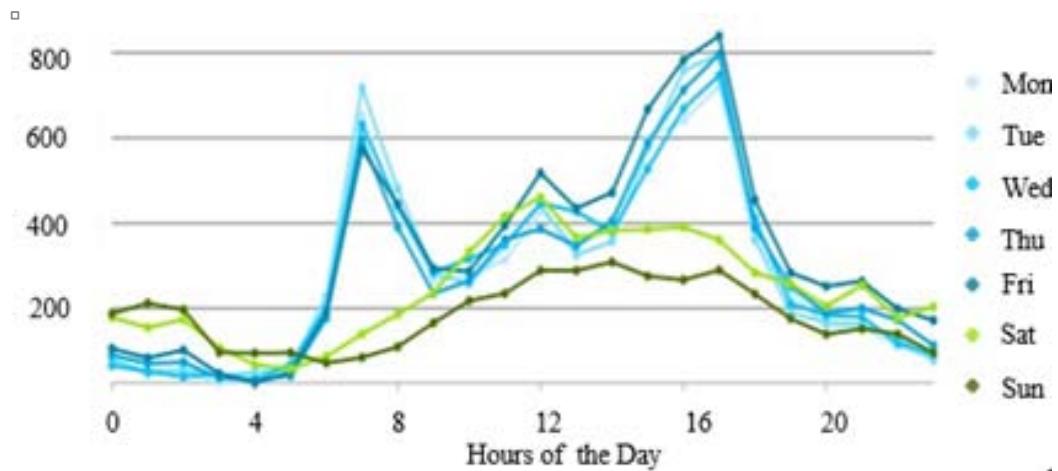


Figure 3: Daily and hourly variation of car crashes.

## Conclusions

Geostatistical methods work well for mapping existing traffic “black” zones. However, they cannot be used to predict the crash rate for newly-constructed roads. Data mining methods, which are able to link crash risk with road-related features, can overcome this limitation. Despite the promising results in this study, real-life applications of data mining methods can be challenging due to large volumes of road segment data required to generate sufficient road features.

## References

- Anderson, T.K. (2009) Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, 41, 3, pp. 359-364.
- Beshah, T. and Hill, S. (2010, March) Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia. In *AAAI Spring Symposium: Artificial Intelligence for Development*.
- Effati, M., Rajabi, M. A., Samadzadegan, F., and Blais, J. A. (2012) Developing a novel method for road hazardous segment identification based on fuzzy reasoning and GIS. *Journal of Transportation Technologies*, 2, 1, pp. 32-40.
- Effati, M., Thill, J.C. and Shabani, S. (2015) Geospatial and machine learning techniques for wicked social science problems: analysis of crash severity on a regional highway corridor. *Journal of Geographical Systems*, 17, 2, pp. 107-135.
- Evans, L. (2004) Traffic safety. Science Serving Society, Bloomfield Hills, MI.
- Federal Highway Administration. (2011, June & July) Crash Rate Calculations. [https://safety.fhwa.dot.gov/local\\_rural/training/fhwasa1109/app\\_c.cfm](https://safety.fhwa.dot.gov/local_rural/training/fhwasa1109/app_c.cfm) Last visited 11/30/2017.
- Feelders, A., Daniels, H., and Holsheimer, M. (2000) Methodological and practical aspects of data mining. *Information & Management*, 37, 5, pp. 271-281.
- Goodchild, M.F. (2015) Space, place and health. *Annals of GIS*, 21, 2, pp. 97-100.
- Harris, N. L., Goldman, E., Gabris, C., Nordling, J., Minnemeyer, S., Ansari, S., ... and Potapov, P. (2017) Using spatial statistics to identify emerging hot spots of forest loss. *Environmental Research Letters*, 12, 2, pp. 012-024.
- Kumar, S., and Toshniwal, D. (2015) A data mining framework to analyze road accident data. *Journal of Big Data*, 2, 1, pp. 26
- Moellering, H. (1976) The potential uses of a computer animated film in the analysis of geographical patterns of traffic crashes. *Accident Analysis & Prevention*, 8, 4, pp. 215-227.
- Mohaymany, A. S., Shahri, M., and Mirbagheri, B. (2013) GIS-based method for detecting high-crash-risk road segments using network kernel density estimation. *Geospatial Information Science*, 16, 2, pp. 113-119.
- Olusina, J., and Ajanaku, W. A. (2017) Spatial analysis of accident spots using weighted severity index (WSI) and density-based clustering algorithm. *Journal of applied sciences and environmental management*, 21, 2, pp. 397-403.

Prasannakumar, V., Vijith, H., Charutha, R., and Geetha, N. (2011) Spatio-temporal clustering of road accidents: GIS based analysis and assessment. *Procedia-Social and Behavioral Sciences*, 21, pp. 317-325.

Shah, S., Brijs, T., Ahmad, N., Pirdavani, A., Shen, Y., and Basheer, M. (2017) Road Safety Risk Evaluation Using GIS-Based Data Envelopment Analysis—Artificial Neural Networks Approach. *Applied Sciences*, 7, 9, pp. 886.

Songchitruksa, P. and Zeng, X. (2010) Getis-Ord spatial statistics to identify hot spots by using incident management data. Transportation Research Record: *Journal of the Transportation Research Board*, 2165, pp. 42-51.

Thakali, L., Kwon, T. J., and Fu, L. (2015) Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. *Journal of Modern Transportation*, 23, 2, pp. 93-106.

World Health Organization. (2015). Global status report on road safety 2015. World Health Organization. [http://www.who.int/violence\\_injury\\_prevention/road\\_safety\\_status/2015/en/](http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/) Last visited 11/30/2017.

Yao, S., Loo, B. P., and Zi, Y. B. (2015) Traffic collisions in space: four decades of advancement in applied GIS. *Annals of GIS*, 22, 1, pp. 1-14.

**Xiao Li**, Ph.D. Student, Department of Geography, Texas A&M University, College Station, TX 77840

**Daniel W. Goldberg**, Assistant Professor, Department of Geography, Texas A&M University, College Station, TX 77840

**Tracy Hammond**, Director, Sketch Recognition Lab, Professor, Department of Computer Science & Engineering, Texas A&M University, College Station, TX 77840

**Xingchen Chen**, Master Student, Department of Geography, Texas A&M University, College Station, TX 77840

**Deep Learning for Geospatial Big Data Analytics:  
Terrestrial Ecological Systems Recognition and Classification Assessment**  
**Qingmin Meng**

**ABSTRACT:** Terrestrial landscape management and ecological conservation/restoration are typically based on the data of land cover categories, which represent the interface between natural environment and human activities across space and over time. Land cover as the composition and characteristics of land surface elements is the key environmental information, and the status and dynamics are the important indications for natural resource management and policy purposes. Integrated remote sensing and GIS has been an efficient and quick way to observe and monitor biophysical characteristics of terrestrial landscape dynamics. Among these geospatial technology, geographic feature classification and mapping are the effective approaches to characterizing natural resources and landscape ecology. The applications of remote sensing data in landscape ecology, modeling, and planning have been as broad as landscape science itself. Satellite imagery based geospatial modeling has been an efficient way to observe and monitor biophysical characteristics of large area land cover, land cover changes, and landscape dynamics over time. Large area terrestrial ecosystems monitoring using remote sensing provides a basic prerequisite for many scientific applications including but not limited to land use land cover management, landscape conservation, environmental study, and ecological analysis.

The significant advancements of data science in the last few years provide deep learning and big data for geospatial big data analytics and its potential applications to terrestrial ecological conservation and mapping. However, there is very limited research focusing on geospatial big data driven ecological feature recognition using deep learning. The objectives of this study are to:(1) Design a typical deep learning approach for geospatial big data driven terrestrial ecosystems recognition, and total more than 100 types of ecological systems are classified and mapped in the Eastern GCPO region.(2) Understand how deep learning improve effective ecosystem feature extractor and boost recognition accuracy for geospatial big data classification. And, (3) assess the differences in computation cost, recognition efficiency, and classification accuracy between deep learning and other common machine learning approaches including the support vector machine classifier (SVMc), neural network classifier (NNc), and random forests classifier (RFC),

**KEYWORDS:** Geospatial big data; deep learning classifier; support vector machine classifier; neural network classifier; random forests classifier

**Qingmin Meng**, Assistant Professor, Department of Geosciences, Mississippi State University, MS 39762

# Development of a GIS-Based Model to Examine Alternative Scenarios for Last-Mile Freight Delivery

Amy M. Moore

**ABSTRACT:** For years, research has been conducted to look at better ways to optimize freight routing to result in fewer miles traveled. New technologies have allowed for the development of hybrid and electric trucks, with ongoing research promising to integrate autonomous electric trucks into the road network in the very near future. With this said, there is still a continuing need to further investigate various modalities with respect to local, or intra-city freight movement to compliment the ongoing research focused on long-haul freight efficiency, and new technologies using alternative energy sources.

The purpose of this study is to examine traditional truck movements within the city of Columbus, Ohio, and upon further examination of these varying routes, develop multi-modal scenarios where innovative technologies can be applied to improve efficiency and reduce energy usage specifically for intra-city freight movement. This study is part of the Multi-Modal Pillar of the Department of Energy (DOE)'s Systems and Modeling for Accelerated Research in Transportation (SMART) Mobility project. Columbus, Ohio is the SMART City Challenge winner and has been the focus of ongoing efforts to improve the infrastructure, job prospects for residents, and the live-ability of Columbus. This study is a partnership with the Oak Ridge National Laboratory, National Renewable Energy Laboratory, and Idaho National Laboratory, and with data provisions from the Mid-Ohio Regional Planning Commission (MORPC) and the United Parcel Service (UPS). The deliverables from this study aim at providing other cities with a useful tool to further examine intra-city freight movement and alternative modes to improve efficiency and reduce energy.

The GIS-based model is currently being developed using ArcGIS and TransCAD software. UPS provided Global Positioning System (GPS) data from their Columbus Depot fleet of standard trucks. This GPS data was processed and imported into ArcGIS for analysis. MORPC provided Traffic Analysis Zone (TAZ)-level socioeconomic, business, and land use data. Locational variables were developed in the GIS. Using all of these data sources, a model was developed to estimate freight deliveries in all 1,086 TAZs within Franklin County, Ohio. This model is currently being improved but will be useful for other cities to estimate freight delivery demand in areas lacking actual freight delivery data. Using actual and estimated locations of freight deliveries, TransCAD software was used to develop optimized routes to simulate truck tours. Additionally, optimized routes were developed using various scenarios incorporating the use of Electric Vehicle (EV) trucks, EV delivery vans, parcel lockers, drones, and Uber-style delivery systems. Estimates of kilowatt-hour per mile energy usage were calculated to compare the different scenarios.

In this study, GIS has proven to be an invaluable tool for model development, both for freight delivery demand estimation and tour model development for trucks and alternative modes. GIS improved the data analysis process and data visualization of the UPS data and the route modeling. As this project expands to incorporate additional scenarios and new technologies, and additional larger GPS datasets, covering larger cities or regions, GIS will continue to be the appropriate tool for further analysis and model development.

**KEYWORDS:** Tour-Based Freight Model, Freight Delivery, Last-Mile, SMART Mobility, TransCAD

**Amy M. Moore**, Postdoctoral Research Associate in Transportation Modeling and Energy Analysis, National Transportation Research Center, Oak Ridge National Laboratory, Knoxville, TN 37932

# From Point Clouds to Tactile Maps: How Lidar and Photogrammetry Can Improve Maps for People with Visual Impairments

Thomas J. Pingel, Matthew W. Mendez and Earle W. Isibue

**ABSTRACT:** Although many advances in turn-by-turn, GPS-enabled guidance for people with visual impairments have been made in recent years, research suggests that map-based learning significantly improves long-term spatial memory and wayfinding performance. As part of the long tradition of touch-based maps, we have developed methods for creating 3D printed maps for people with visual impairments based on laser scans and photogrammetrically reconstructed 3D models of the environment. These methods provide a powerful mechanism to capture local areas at ultra-high resolution, allowing highly detailed indoor/outdoor models of the built environment to be created. We expect that such 3D printed maps will provide an improved medium through which people with visual impairments can more efficiently and accurately build cognitive maps of their local environment.

**KEYWORDS:** Lidar, Unmanned Aerial Vehicles, Photogrammetry, Tactile Maps, Visual Impairment

## Introduction

Many aids to navigation for people with low vision or blindness have been developed in recent years, including GPS-enabled guidance systems (Loomis, Golledge, and Klatzky, 1998; Marston et al., 2006; Katz et al., 2012) and near-field beacon systems for place identification and indoor navigation (Cheraghi et al., 2017). While tactile maps have proven effective for both place-learning and navigation (Tatham, 2003; Simonnet et al., 2011; Brock et al., 2015) cost of construction, among other factors, has prevented their widespread adoption.

Even with direct wayfinding support from personal guidance systems, there remains a role for maps to support spatial awareness. Sighted persons that use GPS-enabled personal guidance systems have been shown to develop cognitive maps of the environment more slowly, and with greater error than those using non-digital maps (Ishikawa et al., 2008; Ishikawa and Takahashi, 2014), and there is good reason to think that while personal guidance systems for the blind provide a net benefit, they may also have similar costs. Maps, unlike turn-by-turn directions, provide the opportunity to query and explore, providing a measure of the kind of interactivity shown to improve cognitive map development (Brock et al., 2015).

The rise in availability of inexpensive 3D printers now creates the possibility for wide dissemination of tactile maps. Similarly, the wide availability of airborne laser scan or lidar data over much of the country provides a ready data source for the creation of sub-meter accuracy maps and models. Inexpensive terrestrial laser scanners and photometrically-derived point clouds sourced from unmanned aerial vehicles (UAVs or drones) provide even higher resolution output at low cost. The confluence of these

technologies affords great opportunity to create detailed renderings of both indoor and outdoor spaces via 3D printed tactile maps.

This approach has several advantages, including reduced cost of construction and ease of distribution of maps / models through existing free 3D model distribution networks such as Thingiverse. Such models of neighborhoods, buildings, and interiors have the capability to represent complex shapes and textures, potentially easing burdensome symbolization requirements (Tatham, 2003). Such maps may also feature interaction capabilities already developed for tactile maps, including haptic and auditory feedback (Rice et al., 2005; Simonnet et al., 2011).

We present the results of our development of such maps focused on two research questions. First, what laser-based and photogrammetric techniques are most effective at capturing the local environment? And second, what extensions of the cartographic techniques of simplification, generalization, exaggeration, and labeling can best be employed to produce physical representations of buildings and neighborhood-scale environments that are the most helpful in aiding people who are visually impaired to develop accurate cognitive maps of their environment.

## References

- Brock, A. M., Truillet, P., Oriola, B., Picard, D. and Jouffrais, C. (2015) Interactivity Improves Usability of Geographic Maps for Visually Impaired People. *Human–Computer Interaction*, 30, 2, pp. 156-194.
- Cheraghi, S. A., Namboodiri, V. and Walker, L. (2017) GuideBeacon: Beacon-Based Indoor Wayfinding for the Blind, Visually Impaired, and Disoriented. IEEE International Conference on Pervasive Computing and Communications, pp. 121-130.
- Ishikawa, T., Fujiwara, H., Imai, O. and Okabe, A. (2008) Wayfinding with a GPS-based Mobile Navigation System: A Comparison with Maps and Direct Experience. *Journal of Environmental Psychology*, 28, 1, pp. 74-82.
- Ishikawa, T., & Takahashi, K. (2014) Relationships between Methods for Presenting Information on Navigation Tools and Users' Wayfinding Behavior. *Cartographic Perspectives*, 75, pp. 17-28.
- Katz, B. F., Kammoun, S., Parseihian, G., Gutierrez, O., Brilhault, A., Auvray, M., Truillet, P., Denis, M., Thorpe, S. and Jouffrais, C. (2012) NAVIG: Augmented Reality Guidance System for the Visually Impaired. *Virtual Reality*, 16, 4, pp. 253-269.
- Loomis, J. M., Golledge, R. G. and Klatzky, R. L. (1998) Navigation System for the Blind: Auditory Display Modes and Guidance. *Presence: Teleoperators and Virtual Environments*. 7, 2, pp. 193-203.

Marston, J. R., Loomis, J. M., Klatzky, R. L., Golledge, R. G. and Smith, E. L. (2006) Evaluation of Spatial Displays for Navigation Without Sight. *ACM Transactions on Applied Perception (TAP)*. 3, 2, pp. 110-124.

Rice, M., Jacobson, R. D., Golledge, R. G. and Jones, D. (2005) Design Considerations for Haptic and Auditory Map Interfaces. *Cartography and Geographic Information Science*, 32, 4, pp. 381-391.

Simonnet, M., Vieilledent, S., Jacobson, R. D. and Tisseau, J. (2011) Comparing Tactile Maps and Haptic Digital Representations of a Maritime Environment. *Journal of Visual Impairment & Blindness*, 105, 4, pp. 222-234.

Tatham, A. F. (2003) Tactile Mapping: Yesterday, Today and Tomorrow. *The Cartographic Journal*, 40, 3, pp. 255-258.

**Thomas J. Pingel**, Assistant Professor, Department of Geographic and Atmospheric Sciences, Northern Illinois University, DeKalb, IL 60115

**Matthew W. Mendez**, Undergraduate Student Researcher, Department of Geographic and Atmospheric Sciences, Northern Illinois University, DeKalb, IL 60115

**Earle W. Isibue**, Graduate Student Researcher, Department of Geographic and Atmospheric Sciences, Northern Illinois University, DeKalb, IL 60115

# Position Paper: Understanding the Analytical Affordances of Absence in Spatial Data Science

**Anthony C. Robinson**

**ABSTRACT:** The arrival of big spatial data prompts us to consider the ways in which analysts can recognize and account for missing information in mapping and spatial analysis. This position paper argues for new GIScience research to address problems associated with visual representation and analytical reasoning with missing data and data that have attributes of absence. It argues that there are important analytical affordances associated with absence in spatial data science, and that they must be approached from visual as well as sensemaking perspectives.

**KEYWORDS:** visualization, absence, analytical reasoning, spatial data science, cartography

## Introduction

The spatial big data era promises to provide detailed observations of people and our planet with great density, coverage, and update frequency. Here we argue that there is an emerging challenge for cartographers to develop and evaluate techniques for visually representing and drawing attention to the absence of spatial data observations, rather than just their presence. For example, where people are tweeting during a disaster may rightfully be the center of a great deal of analytical attention, but the locations in which people have suddenly stopped tweeting may in fact represent the worst impacted place. Sensor network coverage may provide detailed observations on changes in soil moisture on a farm field, but being able to recognize when one or more sensors has gone offline will become essential in order to properly deliver on the promises of precision agriculture. These scenarios prompt a key question for GIScientists to solve: how do we leverage the analytical affordances of absence in spatial data science?

Understanding where there are gaps in coverage (missing data) represents a different analytical scenario than understanding the quality dimensions of data that are already present (attributes of absence in data). For example, as discussed in our precision agriculture scenario above, the analytical affordances of knowing that several sensors that normally capture data have stopped doing so are qualitatively different than what can be gleaned from a situation in which the sensors are working but are indicating attributes that signify absence (e.g. lack of any measurable soil moisture).

The importance of absence in Geography is not a new idea, having some of its earliest recognition in work by Hägerstrand (1984). This problem space is also closely related to the significant body of work in GIScience that has focused on methods for measuring and communicating uncertainty (Coulclelis, 2003; Kinkeldey et al., 2014; MacEachren et al., 2005). We argue however that new efforts need to consider what to do with representing missing information as well as how to highlight attributes of absence found in existing

information. For example, parcels in a cadaster that are categorized as vacant may not be represented at all in a typical basemap. One need only to look at the Lower 9th Ward in New Orleans in any popular web map to see the effect that this has on what users see – someone who understands the place and its context will recognize the long shadow of Hurricane Katrina, but nothing on the map itself helps to draw attention explicitly to the attributes of absence found in New Orleans parcel data. It is not the main purpose of a basic reference map to draw such attention, but we argue that this example highlights the potential power associated with exploring new ways to handle and represent attributes of absence.

## Representation and Reasoning with Absence

We hypothesize that representation methods for revealing the presence of absence may be more effective if designed using cues that humans routinely perceive to signify absence. For example, although it is possible to simply highlight missing data with a distinct hue or not draw it at all, we propose that representing those data with shadows, transparency, blur, desaturation, or texture may aid map reading tasks that require users to identify and characterize absence on maps. Simply put, we need to evaluate the extent to which specific visual methods for representing absence may influence map reading and spatial analytical reasoning.

The basis for our hypothesis that representing absence may require special cues is the body of previous work in perceptual science on the topic of visual search asymmetry, which suggests that performance varies depending on whether or not users are searching for the presence or absence of visual attributes (Treisman & Souther, 1985). This literature suggests that identifying features that include the presence of a cue is faster and easier compared to identifying features that are missing a cue (Wolfe, 2001). While the precise mechanism for understanding why visual asymmetry occurs remains a source of ongoing research (Moran et al., 2016), this effect has a potentially important connection to GIScience efforts to leverage the analytical affordances of absence in spatial data through visual representations. Map readers may have a difficult time with visual search tasks to find missing values unless we provide distinct cues to draw attention, and we do not have evidence yet from cartographic research to suggest which cues we should be using.

In addition to better understanding how we might represent absence in big spatial data, we also need to know more about how the process of analytical reasoning (Pirolli & Card, 2005) is influenced by missing information. What are the touchpoints in the reasoning process in which we can prompt analysts to explore these aspects, and how can our geospatial tools help them accomplish that kind of task?

Finally, we note the need for future analysts to be able to leverage the analytical affordances of missing/absent spatial information. To do so, we must adjust educational approaches in GIScience to encourage projects that incorporate structured analytical reasoning techniques that require students to explain supporting evidence as well as contradictory evidence, and to explicitly define what information is present versus which is missing or contains attributes of absence. In other words, we need to help students

recognize and account for their blind spots, and to transform that knowledge into an analytical affordance in its own right.

## References

- Couclelis, H. (2003). The Certainty of Uncertainty: GIS and the Limits of Geographic Knowledge. *Transactions in GIS*, 7(2), 165-175.
- Hägerstrand, T. (1984). Presence and absence: A Look at Conceptual Choices and Bodily Necessities. *Regional Studies*, 18(5), 373-379.
- Kinkeldey, C., MacEachren, A. M., & Schiewe, J. (2014). How to Assess Visual Communication of Uncertainty? A Systematic Review of Geospatial Uncertainty Visualisation User Studies. *The Cartographic Journal*, 51(4), 372-386.
- MacEachren, A. M., Robinson, A. C., Hopper, S., Gardner, S., Murray, R., & Gahegan, M. (2005). Visualizing Geospatial Uncertainty: What We Know and What We Need to Know. *Cartography and Geographic Information Science*, 32(3), 139-160.
- Moran, R., Zehetleitner, M., Liesefeld, H. R., Müller, H. J., & Usher, M. (2016). Serial vs. Parallel Models of Attention in Visual Search: Accounting for Benchmark RT-Distributions. *Psychonomic Bulletin & Review*, 23(5), 1300-1315.
- Pirolli, P., & Card, S. (2005, May 2-6). The Sensemaking Process and Leverage Points for Analyst Technology as Identified through Cognitive Task Analysis. Paper presented at the *International Conference on Intelligence Analysis*, McLean, VA.
- Treisman, A., & Souther, J. (1985). Search Asymmetry: a Diagnostic for Preattentive Processing of Separable Features. *Journal of Experimental Psychology*, 114(3), 285-310.
- Wolfe, J. M. (2001). Asymmetries in Visual Search: An Introduction. *Perception & Psychophysics*, 63(3), 381-389.

**Anthony C. Robinson**, Assistant Professor, GeoVISTA Center, Department of Geography, The Pennsylvania State University, University Park, PA 16802 <[a-robinson@psu.edu](mailto:a-robinson@psu.edu)>

# MapStudy:

## An Open Source Survey Tool for Studying Interactive Web Maps

### Robert E. Roth, Carl Sack, Meghan Kelly, Nick Lally and Kristen Vincent

**ABSTRACT:** New geospatial web technologies have radically changed how maps are created and used. Yet, many of our most basic cartographic design principles were developed in a pre-digital era, with empirical studies rarely replicated in an interactive, digital medium. Further, user studies increasingly are administered not for the goals of basic science, but for usability engineering to gather feedback on a single interactive or web map design. Studying interactive map design is difficult, though, requiring new technical skills to develop the experimental materials, new procedures that capture free exploration, and new measures to determine design success.

Here, we describe MapStudy, a survey tool facilitating rapid design of empirical studies on interactive web maps. MapStudy is based on open web standards and can display maps loaded directly from a URL or created dynamically through the modularized design library built atop Leaflet.js. A new survey can be configured through a point-and-click set-up form or directly using JSON notation. MapStudy supports a range of survey questions and measures (e.g., multiple choice, rating scales, short answer), including response time and interaction logging. Participant responses can be forwarded via email or logged into an SQL database (requiring additional technical set-up). Documentation and source code for MapStudy are available for extension and reuse at <http://github.com/uwcart/mapstudy/>.

**KEYWORDS:** interactive maps, web maps, methodology, user studies, open source

**Robert E Roth**, Faculty Director, University of Wisconsin Cartography Lab, Associate Professor, University of Wisconsin-Madison Department of Geography, Madison, WI 53706

**Carl Sack**, Geography/GIS Faculty and Geospatial Technologies Program Coordinator, Fond du Lac Tribal & Community College, Cloquet, MN 55720

**Meghan Kelly**, PhD Candidate, University of Wisconsin-Madison Department of Geography, Madison, WI 53706

**Nick Lally**, Assistant Professor, University of Kentucky, Lexington, KY 40508

**Kristen Vincent**, MSc, University of Wisconsin-Madison Department of Geography, Madison, WI 53706

# Aerial Imaging and Lidar Point Cloud Fusion for Low-Order Stream Identification

Ethan Shavers and Lawrence Stanislawski

**ABSTRACT:** Accurate headwater mapping and classification are essential for hydrologic modeling and watershed-related research. Remote sensing techniques have proven useful for waterbody mapping, yet identification of low-order streams in low topographic relief and heavily vegetated environments has remained a challenge. Here we present a model to automatically identify first-order streams in low-relief agricultural areas, and differentiate them from ephemeral channels. The model employs decision tree feature extraction from two datasets with current or forecasted national coverage and availability: airborne light detection and ranging (lidar) point cloud data, and four-band National Agricultural Imagery Program (NAIP) orthophotos.

Ground point density, return intensity, surface elevation, vegetation height, and vegetation complexity are measures derived from lidar point cloud data and used in the stream identification model. Ground point density is an indicator of signal dropout caused by the presence of water in a channel. Lidar return intensity is low at high incidence angles and high at low angles relative to most land cover. Elevation information is used to derive local low and profile curvature thresholds for channel identification. Vegetation height and complexity correlate with riparian zone development in response to stream permanence, and with the presence of vegetation buffer strips that may be maintained along ephemeral channels. NAIP images are used to identify open water using a modified band ratio,  $\sigma(NIR) * blue/NIR$ . The NAIP images also complement the lidar-derived vegetation height and complexity measures for identifying the presence and development of riparian zones and vegetation buffer strips.

A 170 square kilometers watershed in central Iowa is used to develop the model in this study. The area has low topographic relief with average, minimum, and maximum elevation of 306.8, 266.7, and 330.2 meters. Slope average, minimum, and maximum values are 3.2, 0.0, and 77.8 percent rise, respectively. An established weighted flow accumulation model is applied on 1/3 arc-second elevation data to extract surface-water drainage lines in the study area that are likely perennial or intermittent streams. The decision tree feature extraction model is used to identify linear trends in low ground point density and in vegetation complexity, and thereby corroborate or disprove the elevation-derived lines. The model is tested against two adjacent 300 square kilometer watersheds with similar low topographic relief. The research estimates whether or not the presence of riparian zone vegetation heterogeneity is sufficient to differentiate stream channels in low relief areas where vegetation obscures accurate channel or water body mapping. Results of this work are expected to enhance low-order stream mapping and classification to improve the National Hydrography Dataset, and elucidate riparian zone dynamics. Challenges encountered thus far are related to data heterogeneity, in particular NAIP collection solar angle. This can potentially be addressed using multi-temporal data averages.

**KEYWORDS:** lidar point cloud, object-based feature extraction, flow accumulation, elevation-derived, riparian zone

**Ethan Shavers**, Postdoctoral Research Scientist, U.S. Geological Survey, Center of Excellence for Geospatial Information Sciences, Rolla, MO 65401

**Lawrence V. Stanislawski**, Research Scientist, U.S. Geological Survey, Center of Excellence for Geospatial Information Sciences, Rolla, MO 65401

# Generalizing Linear Stream Features to Preserve Sinuosity for Analysis and Display: A Pilot Study in Multi-Scale Data Science

Lawrence V. Stanislawski, Barry J. Kronenfeld, Barbara P. Buttenfield  
and Tyler Brockmeyer

**ABSTRACT:** Cartographic generalization can impact geometric properties of geospatial data and subsequent analyses. This study evaluates simplification methods with the goal of preserving geometric details, such as sinuosity. We evaluate two recently developed line simplification algorithms that introduce Steiner points: Raposo's Spatial Means, and Kronenfeld's new area-preserving segment collapse algorithm, and compare them with several well-known algorithms. Results indicate the area-preserving segment collapse algorithm optimally simplifies linear stream features with minimal horizontal displacement and the best retention of sinuosity.

**KEYWORDS:** Cartographic data modeling, multi-scale data science, hydrography, sinuosity, geometric pattern recognition

## Introduction

An ongoing challenge in cartographic data modeling through generalization is preservation of specific geometric and semantic properties of geospatial features as levels of detail (LoDs) are reduced for representation at smaller scales. Increasing data volumes in the past half century forced a transition from manual to automated generalization strategies. Concurrently, researchers have sought methods to preserve visual semantics (i.e., overall appearance and recognizability), a difficult problem because visual assessments are not feasible as data volumes transition from megabytes to gigabytes and larger.

While visual semantics has received much attention in the literature, conservation of the analytic utility in generalized data incurs additional and often complex challenges. For example, a generalized stream channel network should preserve evidence of erosion and deposition. Straight lines and right angles should be maintained in irrigation channels. Rounded meanders and non-orthogonal confluences should be maintained in naturally occurring streams.

Previous work has demonstrated generalization strategies preserving length (Buttenfield et al., 2011), angularity (Gökgöz et al., 2015; Jenks, 1983), proportion of high frequency detail (e.g., fractal dimension) (Bernhardt, 1992), network feature density (e.g., thinned roads, stream channel networks) (Stanislawski et al., 2012), vector displacement (Opheim, 1982; Reumann and Witkam, 1974), areal displacement (Ramer, 1972; Douglas and Peucker, 1973; Visvalingam and Whyatt, 1992; Bose et al., 2006; Tong et al., 2015, Shen et al., 2018), topology (e.g., self-crossings and network connectivity) (Saalfeld, 1999; Marquez and Wu, 2003), and self-adjusting tolerance values across a feature or

network to reflect variations in the amount of detail (Steiniger, 2007). Challenges in automating the modification of detail are primarily issues of pattern recognition (Duda et al., 2001; Bishop, 2013) and geometric learning (Shoujue and Jiangliang, 2005), because it is not possible to automatically preserve a property that cannot be detected and characterized automatically.

This paper reports on data modeling that generates hydrographic features at reduced scales while preserving sinuosity. Higher-order properties, such as sinuosity, have proven more difficult to preserve than simple geometric properties such as length, in spite of their importance for hydrologic modeling, hydrographic analysis, and cartographic display.

With most simplification algorithms, feature length and sinuosity can only decrease. However, algorithms that introduce Steiner points, added during geometric optimization to create a better solution than would be possible from the original points alone (Hwang et al., 1992), can increase or decrease sinuosity at each step. We assess sinuosity preservation using two recently developed algorithms that introduce Steiner points: Raposo's algorithm (2013), and Kronenfeld's new area-preserving segment collapse algorithm (2018, not yet published). Results are compared with several well-known algorithms (Visvalingham and Whyatt, 1992; Ramer, 1972; Douglas and Peucker, 1973; and Wang and Muller, 1998) that preserve subsets of original points. This work quantifies sinuosity and shows how to constrain sinuosity reduction in generalized data. A set of 50 stream features collected from different landscapes across the conterminous United States are tested. For statistical comparison, features are simplified to a controlled number of vertices. Results are evaluated on retained sinuosity, and vector and areal displacement.

## Methods

### *Sample Data*

Fifty sections of linear hydrographic stream segments, including headwater-to-confluence and confluence-to-confluence sections, were selected from the National Hydrography Dataset (NHD) in various climate and terrain conditions (Figure 1). The stream sections were initially selected from 1:10,000,000-scale National Atlas hydrographic data, with the condition that all features are also represented at 1:24,000 (24K). At 24K, the 50 stream sections are comprised from more than 15,000 NHD flowline features, because of divisions at network confluences and feature type changes. NHD flowline feature types include Stream/River, Artificial Path, Canal/Ditch, Pipeline, and Connector (USGS, 2000). Consecutive 24K flowline features were combined to form a set of 50 continuous stream sections. Beyond this, the data did not undergo prior processing.

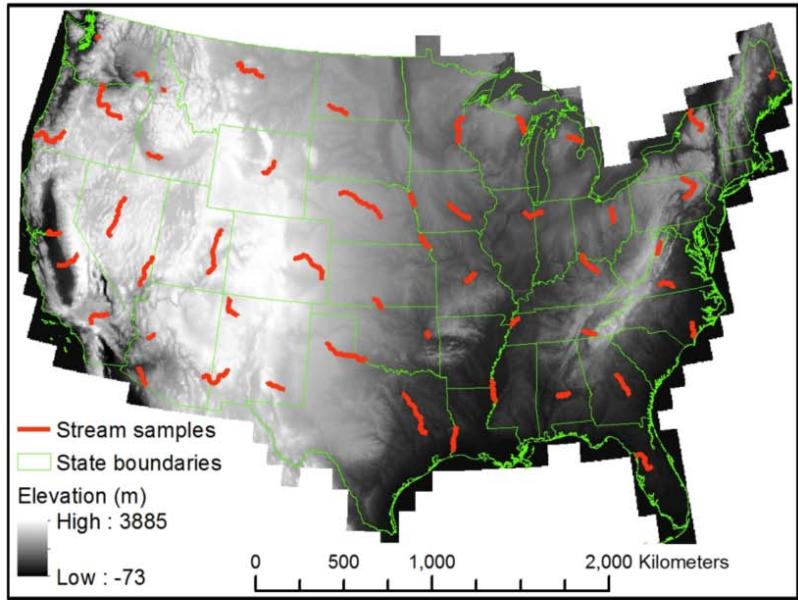


Figure 1: Distribution of 50 sample stream segments over the conterminous United States. Features are selected from the flowline feature class of the 1:24,000-scale National Hydrography Dataset. Five-kilometer resolution data show the approximate range of elevation conditions within the study.

### ***Scale-based Legibility Constraint***

A scale-based legibility constraint based on the pixel resolution of a typical computer display screen is used for this study. A 19-inch computer monitor with a display resolution of 1280x1024 has a pixel size of 0.29 millimeters (mm) and pixel diagonal of about 0.4 (mm). Therefore, a 0.4 mm legibility constraint is applied at 12 scales (1:5,000; 1:10,000; 1:24,000; 1:50,000; 1:100,000; 1:250,000; 1:500,000; 1:1,000,000; 1:2,000,000; 1:5,000,000; 1:10,000,000; and 1:20,000,000) to form the associated minimally visible spacing in meters (2.0; 4.0; 9.6; 20.0; 40.0; 100.0; 200.0; 400.0; 800.0; 2000.0; 4000.0; and 8,000.0) for displayed cartographic features. Some values were graphically verified by measuring the diameter of several small bends in the hydrographic lines to approximate the smallest scale at which the bend collapses. The legibility constraint is more conservative than a positional accuracy constraint, such as the National Map Accuracy Standard of 1/50<sup>th</sup> inch (0.51 mm) at 1:20,000-scale and smaller (U.S. Bureau of Budget, 1947), thus enabling retention of additional shape details at smaller scales, respecting visual and analytical data character.

### ***Comparison of Simplification Algorithms***

Five simplification algorithms are compared. Three algorithms do not introduce Steiner points. Wang and Muller's Bend-Simplify (BS) algorithm eliminates vertices that represent bends smaller than the tolerance. The Ramer-Douglas-Peucker (RDP) algorithm reduces vertices by retaining all vertices that are further than the tolerance distance from a hierarchy of baselines. The Visvalingam Effective-Area (VIS) algorithm eliminates

vertices that minimize areal displacement. Because these algorithms only eliminate vertices, the length and sinuosity of the resulting line is always reduced.

The Raposo Spatial Means (RSM) algorithm uses a hexagonal tessellation to retain a minimum number of shape points for each feature, replacing each contiguous sequence of vertices within a hexagon with a single Steiner point. Raposo's Steiner points are computed as the spatial mean of vertices within a hexagon. As such, the results can only compress bends (thus reducing sinuosity), and we hypothesize that sinuosity will be reduced similar to the BS, RDP, and VIS algorithms.

Kronenfeld's new simplification algorithm, referred to as Area-Preserving Segment Collapse (APSC) replaces consecutive pairs of vertices with a single new vertex (Steiner point) whose location is chosen to minimize areal displacement on either side of the input line. APSC proceeds similarly to VIS, but because Steiner points may be placed both inside and outside of bends, bends can be exaggerated as well as eliminated. For this reason, it is hypothesized that the APSC algorithm will preserve sinuosity and minimize areal displacement.

Parameter selection to guide simplification of hydrographic features is challenging because there are various features types, collection methods, and environmental conditions that must be considered. To make fair comparisons among the five algorithms, our methods force the same number points to be retained for each simplified feature at each scale. This is accomplished by first simplifying the features with BS using the 12 legibility constraints (2.0 through 8,000.0 meters) and determining the number of points retained for each simplified feature. Subsequently, the original features are simplified with each of the four other methods, constraining each simplification to retain the same number of points determined from its associated BS version. Original and simplified features are compared using metrics of length, sinuosity, modified Hausdorff distance (MHD), fractal dimension, and areal displacement. For computational efficiency, MHD is the larger of two distances: either the furthest distance between the vertices of the simplified line and its original version, or the furthest distance between the vertices of the original line and its simplified version. Because these two distances are not always vertex-to-vertex, they will be unequal in almost every case, as long as some vertex reduction occurs.

## Results and Discussion

This analysis compares five simplification algorithms in terms of retaining sinuosity for all types of streams. Results are summarized across the contiguous United States, to highlight algorithm treatments, rather than to highlight regional differences, which should be expected under differing landscape conditions, bedrock types, and terrain slopes. The expectation of regional differences draws upon earlier development of ecosystem maps by Bailey (1998), by Comer et al (2003), and Sayre et al (2009), showing that ecosystem differences depend upon variations in climate, vegetative cover, landforms, soils, surface water and groundwater. Stream sinuosity will vary and reflect these landscape factors, and our sample of streams is distributed across a variety of ecological divisions.

Average MHDs indicate a pattern of displacement showing gradually larger increases in MHD with smaller scales and fewer retained points (Figure 2). The BS algorithm shows the highest horizontal displacement, except at the least aggressive simplifications, where the RSM method generates greatest average displacement. Given the high frequency of vertices along some parts of these stream sections, RSM does not appear capable of reducing just a few points to retain small positional displacement, especially for sinuous lines. The RDP method tends to minimize vertex displacement at all LoDs. APSC shows the second smallest MHD averages, lower than VIS for intermediate and aggressive simplifications.

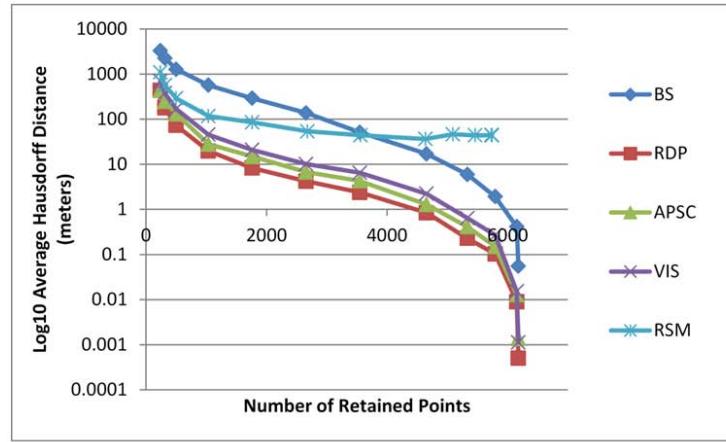


Figure 2: Average Modified Hausdorff distances determined between original and simplified features for 50 hydrographic stream sections simplified through five algorithms [Bend Simplify (BS), Ramer-Douglas-Peucker (RDP), Area Preserving Segment Collapse (APSC), Visvalingam Effective-Area (VIS), and Raposo Spatial Means (RSM)]. Averages are plotted on a base-ten log scale to enhance visual separations between methods for 12 levels of detail based on the number of points retained.

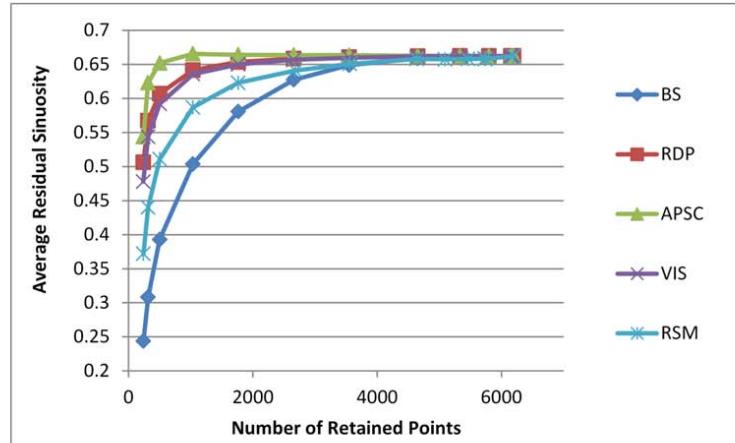


Figure 3: Average residual sinuosity values determined for 50 original and simplified hydrographic stream sections simplified through five algorithms [Bend Simplify (BS), Ramer-Douglas-Peucker (RDP), Area Preserving Segment Collapse (APSC), Visvalingam Effective-Area (VIS), and Raposo Spatial Means (RSM)]. Averages are shown for 12 levels of detail based on the number of points retained.

As expected, APSC preserves average sinuosity better than all other methods at all LoDs, with RDP and VIS in second and third place, respectively (Figure 3). Among all the tested algorithms, BS is least effective at preserving sinuosity as fewer points are retained. Figure 4 shows an example of a simplified hydrographic feature by these methods compared to the original line.

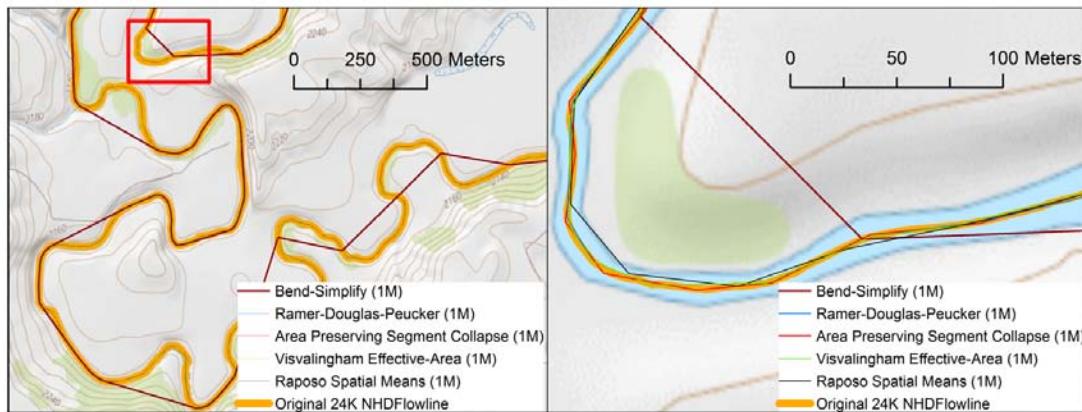


Figure 4: Example of linear feature simplified by five algorithms (Bend-Simplify, Ramer-Douglas-Peucker, Area Preserving Segment Collapse, Visvalingam Effective-Area, and Raposo Spatial Means) to 1:1,000,000 detail. Original feature is shown with broad orange line. Background graphic is a 1:24,000-scale U.S. Topographic map. Area of detail view in right panel is outlined with red box in left panel.

This paper has addressed a preliminary question about the relationship between simplification and sinuosity. Presented results indicate APSC optimally simplifies linear stream features with minimal horizontal displacement and the best retention of sinuosity. Related questions assessing the impact of simplification algorithms on regional differences in stream sinuosity forms a topic for further statistical analysis and metrics. Examination of contextual issues such as the possibility of sinuosity variations between and within geographic regions is another area for future research.

## References

- Bailey, R.G. (1998). *Ecoregions Map of North America: Explanatory Note*. Washington DC: U.S. Department of Agriculture Forest Service, Miscellaneous Publication No. 1548, 10 pp.
- Bernhardt, M.C. (1992). Quantitative characterization of cartographic lines for generalization. Report 425, Department of Geodetic Science and Surveying, Ohio State University, Columbus, Ohio. 142 pp.
- Bishop, C.M. (2013). *Pattern Recognition and Machine Learning*. Berlin: Springer Series on Information Science and Statistics.

Bose, P., Cabello, S., Cheong, O., Gudmundsen, J., van Kreveld, M., and Speckman, B. (2006). Area-preserving approximations of polygonal paths. *Journal of Discrete Algorithms* 4, 4, pp. 554-566. DOI 10.1016/j.jda.2005.06.008

Buttenfield, B.P., Stanislawska, L.V. and Brewer, C.A. (2011). Adapting generalization tools to physiographic diversity for the USGS National Hydrography Dataset. *Cartography and GIScience* 38, 3, pp. 289-301.

Comer, P., Faber-Langendoen, D., Evans, R., Gawler, S., Josse, C., Kittel, G., Menard, S., Pyne, M., Reid, M., Schulz, K., Snow, K., and Teague, J. (2003). Ecological systems of the United States, A working classification of U.S. terrestrial systems: Arlington, Va., NatureServe, 75 p.

Douglas, D. H. and Peucker, T.K. (1973). Algorithms for the reduction of the number of points required to represent a digitised line or its caricature, *The Canadian Cartographer*, 10, 2, pp. 112-122.

Duda, R.O., Hart, P.E, Stork, D.G. (2001). *Pattern Classification*. 2nd Ed. NY: Wiley.

Gökgöz, T., Sen, A., Memduhoglu, A. and Hacar, M. (2015). A new algorithm for cartographic simplification of streams and lakes using deviation angles and error bands. *ISPRS International Journal of Geo-Information* 4, pp. 2185-2204. DOI 10.3390/ijgi4042185.

Hwang, F.K., Richards, D.S., and Winter, P. (1992). The Steiner Tree Problem. *Annals of Discrete Mathematics*, 53, Elsevier, 339 pp. <https://www.sciencedirect.com/science/bookseries/01675060/53>.

Jenks, G.F. (1983). Line simplification algorithm based on segment length and angularity. Never published formally but reported in numerous outlets (e.g., in McMaster, R.B. 1983. *Mathematical Measures for the Evaluation of Simplified Lines on Maps*, PhD dissertation University of Kansas; Clayton, V.H. 1985. *Cartographic Generalization: A Review of Feature Simplification and Systematic Point Elimination Algorithms*. Rockville, MD: NOAA Technical Report NOS 112, [https://www.ngs.noaa.gov/PUBS\\_LIB/TRNOS112CGS5.pdf](https://www.ngs.noaa.gov/PUBS_LIB/TRNOS112CGS5.pdf)

Marquez, M.R.G., and Wu, S.T. (2003). A non-self-intersection Douglas-Peucker algorithm. *Proceedings 16th Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPH 2003)*. IEEE Society: 60-66.

Opheim, H 1982. Fast reduction of a digitized curve. *Geoprocessing* 2, pp. 33-40.

Ramer, U. (1972). An iterative procedure for the polygonal approximation of plane curves. *Computer Graphics and Image Processing*, 1, 3, pp. 244-256.

Raposo, P. (2013). Scale-specific automated line simplification by vertex clustering on a hexagonal tessellation. *Cartography and Geographic Information Science* 40, 5, pp. 427-443. <http://dx.doi.org/10.1080/15230406.2013.803707>

Reumann, K., and Witkam, A.P.M. (1974). Optimizing curve segmentation in computer graphics. *International Computing Symposium*. Amsterdam: North Holland Publishing Company: 467-472.

Sayre, R., Comer, P., Warner, H., and Cress, J. (2009). *A New Map of Standardized Terrestrial Ecosystems of the Conterminous United States*. Reston VA: U.S. Geological Survey Professional Paper 1768, 24pp.

Saalfeld, A. (1999). Topologically consistent line simplification with the Douglas-Peucker algorithm. *Cartography and Geographic Information Science*, 26, 1, pp. 7-18. <https://doi.org/10.1559/152304099782424901>

Shoujue, W. and Jiangliang, L. (2005). Geometrical learning, descriptive geometry, and boionimetic pattern recognition. *Neurocomputing* 67: 9-28.

Shen, Y., Ai, T., and He, Y. (2018). A new approach to line simplification based on image processing: a case study of water area boundaries. *International Journal of Geo-Information* 7, 41, 25 pp. DOI:10.3390/ijgi7020041

Steiniger, S. (2007). *Enabling pattern-aware automated map generalization*. PhD. Dissertation, Department of Geography, University of Zurich, Switzerland.

Stanislawski, L.V., Briat, M., Punt, E., Howard, M., Brewer, C.A., and Buttenfield, B.P. (2012). Density-stratified thinning of road networks to support automated generalization for The National Map. *Proceedings 15th ICA Workshop on Generalization*, September 13-14, 2012, Istanbul, Turkey, 10 pp.

Tong, X., Jin, Y., Li, L., and Ai, T. (2015). Area-preservation simplification of polygonal boundaries by the use of the structured total least squares method with constraints. *Transactions in GIS* 19, 5, pp. 780-799. DOI 10.1111/tgis.12130.

Toussaint, G.T. (1982). Computational geometric problems in pattern recognition. Chapter 10 in: Kittler J., Fu K.S., Pau LF. (Eds) *Pattern Recognition Theory and Applications*. Dordrecht: Springer, NATO Advanced Study Institutes Series (Series C - Mathematical and Physical Sciences), vol. 81, pp. 73-91. [https://doi.org/10.1007/978-94-009-7772-3\\_7](https://doi.org/10.1007/978-94-009-7772-3_7).

US Bureau of the Budget (1947). *United States National Map Accuracy Standards*. Washington DC, June 17.

U.S. Geological Survey (2000). *The National Hydrography Dataset: concepts and contents (February 2000)*, United States Geological Survey. [http://nhd.usgs.gov/chapter1/chp1\\_data\\_users\\_guide.pdf](http://nhd.usgs.gov/chapter1/chp1_data_users_guide.pdf). (pdf accessed 27 Feb 2017).

Visvalingam, M. and Whyatt, J.D. (1992). Line generalisation by repeated elimination of the smallest area. *Discussion Paper 10*, Cartographic Information Systems Research Group (CISRG), The University of Hull.

Wang, Z. and Muller, J.C. (1998). Line generalization based on analysis of shape characteristics. *Cartography and Geographic Information Science* 25, 3-15.

**Lawrence V. Stanislawski**, Research Scientist, U.S. Geological Survey, Center of Excellence for Geospatial Information Sciences, Rolla, MO 65401

**Barry J. Kronenfeld**, Associate Professor, Department of Geology and Geography, Eastern Illinois University, Charleston, IL

**Barbara P. Buttenfield**, Professor, Department of Geography, University of Colorado, Boulder, CO 80309-0260

**Tyler Brockmeyer**, Computer Science Developer, Missouri State Technical University, Rolla, MO 65401

# Visualizing Sea Level Rise Induced Migration Using Hexagonal Grids

Hoda Tahami, Bo Zhao, David J.Wrathall and Majidreza Hosseinieh Farahani

**ABSTRACT:** Increased attention to future climate change in recent years has resulted in a wide variety of interactive and online maps displaying the anticipated impacts and consequences of climate change, including sea level rise (SLR). The climate change stressors such as SLR are likely to have dramatic effects on the human migration patterns. By accounting for population growth trends, migration system analysis and the advancements in SLR-adaptation technology, recent studies have estimated that 1.8 meters of sea level rise could displace an estimated 13.1 million people by 2100, living in US coastline counties. Understanding this phenomenon, its causes, processes, and impacts often start from analyzing and visualizing its spatiotemporal patterns. Although the current maps and visualization products demonstrate the areas vulnerable to SLR, they lack to illustrate the destination of the potential displaced migrants and the scale of potential migrations. Projecting that possibility for the United States, the paper aims to create a migration map to show how the United States coastline population might shift due to projected sea level rise and show where the potential migrants will relocate, and how this could alter the U.S. population landscape. In this paper, we proposed and implemented a geo-visualization approach to map the SLR-induced migration in the United States, the specific locations at risk of SLR in future as well as the areas in the US where are likely to experience the most significant population migration due to SLR. The developed application builds an online platform for users to visualize human migration through space and time interactively. A hexagon binning approach as a data aggregation technique is proposed to simplify the shapes of geographic features representing US counties to improve the aesthetics and functionality of the visualization. Developing the proposed web-based platform also required determining the appropriate hexagon size, matching and transferring the migration records with pre-established administrative counties to hexagons; symbolizing the migration flow by arcs of varying weight based on the migration flow rate; connecting the centroids of the origin and destination hexagons; and developing a responsive web thematic map by applying several java scripts libraries including Bootstrap, CartoDB, Leaflet, D3, SpatialSankey and the web-based map tile layers from CartoDB and Leaflet. Combining our geo-visualization technique with the pre-established method for the SLR- derived population projection reveals the impact of sea level inundation on US migrant preferences and suggests areas of high vulnerability of absorbing high rate of migration in future due to SLR. Such comprehensive analytical platform forms a basis for studying the consequences of such migrations, addressing the issues related to the future population distribution, resource allocation, and infrastructure planning and assessing economic and environmental impacts for both inland and coastal communities.

**KEYWORDS:** Geo-Visualization, Hexagonal Binning, SLR-induced Migration

## References

- Hauer, M. (2017) Migration induced by sea-level rise could reshape the US population landscape, *Nature Climate Change*, 7, pp. 321–325

**Hoda Tahami**, PhD Student, College of Engineering, Department of Civil Engineering, Oregon State University, Oregon, OR 97331

**Bo Zhao**, Assistant Professor, College of Earth Ocean and Atmospheric Science, Oregon State University, Oregon, OR 97331

**David J.Wrathall**, Assistant Professor, College of Earth Ocean and Atmospheric Science, Oregon State University, Oregon, OR 97331

**Majidreza Hosseinieh Farahani**, Graduate Student, College of Engineering, Department of Civil Engineering, Oregon State University, Oregon, OR 97331

# **Flyover Country: Mobile Visualization of Geoscience Data**

**Ross Thorn and Shane Loeffler**

**ABSTRACT:** Flyover Country is a mobile app for geoscience data discovery and exploration from anywhere, allowing for offline exploration of data in a region of interest pertaining to a flight path, road trip, or a researcher's field area. Flyover Country uses open data in concert with open source mapping and mobile development technologies to combine geoscience data from many sources into its interactive map. Many types of geoscience data contain variables that vary through time or depth, creating a major visualization challenge. Visualizing and interacting with complex multivariate/spatiotemporal datasets on a mobile device is made increasingly challenging due to smaller screens, reduced processing power, and more limited data connectivity than traditional laptop or desktop computers. After analyzing visualizations in paleoecological papers from the Neotoma database and mobile applications comparable to Flyover Country, we created cartographic design solutions for visualizing these data on a mobile device, testing the effectiveness of each in a user study.

**KEYWORDS:** geoscience, data visualization, mobile maps, UI/UX

**Ross Thorn**, MS Student, Department of Geography, University of Wisconsin – Madison, Madison, WI 53706

**Shane Loeffler**, MS Student, Department of Earth Sciences, University of Minnesota, Minneapolis, MN 55455

# Exploring the Potential of Deep Learning for Settlement Symbol Extraction from Historical Map Documents

Johannes H. Uhl, S. Leyk, Y.-Y. Chiang, W. Duan and C. A. Knoblock

**ABSTRACT:** Historical map documents are unique witnesses of landscapes in the past and contain valuable information about the spatiotemporal evolution of geographic phenomena such as forest coverage, transportation networks, and human settlement patterns. However, this information needs to be extracted (from map documents) and converted into machine-readable data to be used for quantitative analysis. The extraction of information from historical maps is a persistent challenge due to the low graphical quality of the scanned documents and the massive data volume of digital map archives, which can hold hundreds of thousands of scanned map sheets. Recently, several digital map archives have been made available to the public including the United States Geological Survey (USGS) historical topographic maps (Fishburn et al., 2017) and the Sanborn Fire insurance map collection (Library of Congress 2018). For example, the USGS has systematically scanned and georeferenced approximately 200,000 historical topographic map sheets at scales of up to 1:24,000, produced between 1884 and 2006. This map archive is publicly available and represents a unique data source for various applications and studies requiring retrospective geographical data.

However, information extraction from such large amounts of data covering considerable spatial and temporal extents requires systematic, robust, and automated approaches lending from the fields of computer vision, image processing, and machine learning. Convolutional neural networks (CNNs) and other deep learning methods have recently shown promising performance in image recognition tasks in general and applied to geospatial data, such as remotely sensed earth observation data (Ball et al., 2017). In this case study, we present preliminary results of a deep-learning based framework for the extraction of human settlement symbols from historical map documents. Contemporary contextual geospatial data (i.e., building footprints and housing data) are employed to guide an automated collection of training data. The training data are then used for learning and feature extraction using CNNs and image segmentation techniques. The use of contextual geographic information during the training data collection steps overcomes the need for user interventions, which has typically been required for traditional map processing tasks and impeded higher levels of automation. In addition, spatial offsets and temporal inconsistencies between the contextual data and map data require the application of spatial alignment methods and the use of image-processing based filtering methods to ensure clean and representative training samples.

First experimental steps using CNNs (Uhl et al., 2017, Uhl et al., 2018) have shown promising results but need further refinement. Optimally, a pixel-based semantic segmentation of the map documents into the target classes (i.e., individual buildings, dense urban settlement, and no settlement) is desired. However, the sparsely available contextual information (e.g., existing data of building footprints) and the above mentioned discrepancies between contextual and map data do not allow the creation of reliable training labels at pixel level but rather at image level for map subsets cropped around the locations given by the contextual data. In order to overcome this problem of “weakly annotated training data”, unsupervised image segmentation and shape-based analysis are tested to further refine the CNN-based feature extraction results. We are currently exploring suitable ways to incorporate segmentation techniques in the extraction process of settlement symbols.

**KEYWORDS:** Map processing, Deep learning, Convolutional neural networks, Human settlement modelling, Image segmentation

## References

- Ball, J. E., Anderson, D. T., and Chan, C. S. (2017) Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community, *Journal of Applied Remote Sensing*, 2017, 11(4), 042609
- Fishburn, K.A., Davis, L.R., and Allord, G.J. (2017) Scanning and georeferencing historical USGS quadrangles: *U.S. Geological Survey Fact Sheet 2017-3048*, 2 p., <https://doi.org/10.3133/fs20173048>.
- Library of Congress, Geography and Map Division (2018)  
<https://www.loc.gov/collections/sanborn-maps/>, Last visited 1/08/2018.
- Uhl, J. H., Leyk, S., Chiang, Y.-Y., Duan, W. and Knoblock C.A. (2017) Extracting Human Settlement Footprint from Historical Topographic Map Series Using Context-Based Machine Learning, *Conference Proceedings of 8th International Conference on Pattern Recognition Systems (ICPRS), Madrid, Spain*.
- Uhl, J. H., Leyk, S., Chiang, Y.-Y., Duan, W. and Knoblock C.A. (2018) Spatializing uncertainty in image segmentation using weakly supervised convolutional neural networks: A case study from historical map processing, *submitted*.

**Johannes H. Uhl**, PhD Student, Department of Geography, University of Colorado Boulder, Boulder, CO 80309-0260

**Stefan Leyk**, Professor, Department of Geography, University of Colorado Boulder, Boulder, CO 80309-0260

**Yao-Yi Chiang**, Associate Professor (Research), Spatial Sciences Institute, University of Southern California, Los Angeles, CA 90089-0374

**Weiwei Duan**, PhD Student, Department of Computer Science, University of Southern California, Los Angeles, CA 90089-0374

**Craig A. Knoblock**, Professor, Information Sciences Institute, University of Southern California, Los Angeles, CA 90089-0374

# The Evolution of Cartography in the Digital Age: From Digitizing Vertices to Intelligent Maps

**E. Lynn Usery and Dalia Varanka**

**ABSTRACT:** Theoretical and technical developments in cartography, and advancements in computers, networks, and the Internet have resulted in the development of intelligent maps that will be pervasive and ubiquitous using Web technologies. The current state of the art and science of cartography began with digitizing vertices of lines on paper maps into x and y coordinates to represent their basic geometry in computer format. After initial x, y position locations were collected, the next advance was to include identifiers for points, lines, and areas with the associated coordinates. Attributes of these shapes were then added to the identifiers. Data formatting and processing techniques were developed and advanced to consider the topology of these basic geometries, and relational databases were adopted to handle the attributes and topological relationships among the spatial objects. The point, line, and area objects became organized as geographic features of roads, streams, cities, mountains, and other natural and manmade entities. Features have associated characteristics and relations to other features that have enabled the development of a semantic model of the underlying geospatial data. This semantic model has led to the concept of the “Map as a Knowledgebase” generating intelligent and self-reacting features, which can respond to logical inference creating new data through connections, automatically respond to symbolization requirements, and connect with other features in response to user queries or machine requests. The ever-increasing capabilities of computers, networks, and the Worldwide, Semantic, and Geospatial Webs have created a rich backdrop that has greatly enabled the advancement of cartographic representation. The interplay of technological innovations such as sensor networks providing continuous information about connected feature locations and advancements in geographic feature representation have led to the semantically enabled smart map. The concept of the “Map as a Knowledgebase” serves as the foundation for this new development, which can be implemented and used directly in the pervasive and ubiquitous Web environment. Hence, the age of the intelligent map is upon us.

**KEYWORDS:** Intelligent maps, Web, knowledgebase

**E. Lynn Usery**, U.S. Geological Survey, Rolla, MO 65401

**Dalia Varanka**, U.S. Geological Survey, Rolla, MO 65401

# Maps as Graphs: An Implementation for Cartographic Retrieval of Geospatial and Geographical Linked Data

Dalia E. Varanka, Logan J. Powell and William L. Baumer

**ABSTRACT:** Linked Data (LD) for geospatial data systems benefit from map interfaces for visualization. Although map interface capabilities such as plotting geographic coordinates retrieved with SPARQL Protocol and RDF Query Language (SPARQL) queries are widely used, these forfeit the advantages LD offer for browsing a graph. The described project explores the challenges and initial solutions to browse data triples through a map interface.

**KEYWORDS:** Linked Data, Map Interface, Data-Driven Documents (D3), Leaflet, Web Application

## Introduction

Linked Data (LD) based on Resource Description Framework (RDF) and other semantic technologies are introducing new capabilities for geographic information science. One such capability is user interface interaction between maps and geospatial and geographic LD. However, technical implementations still lag in regard to the ease with which more popular mapping software packages are used by the broader public. We present an implementation of a mapping interface that accesses LD directly as cartographic features. The approach works with mapping results from SPARQL Protocol and RDF Query Language (SPARQL) queries and with the browseable graph or ‘follow-your-nose’ approaches. Though software such as OpenLayers easily maps SPARQL query results, the browseable graph approach does not appear to have been explored in cartographic research literature. Our research contributes to the general development of maps based on linkable graph data models.

The geographic coordinates of large linked databases such as Geonames or OpenStreetMap are often rendered against commonly available Internet base map layers that allow the addition, edit, aggregation, and visualization of vector data through JavaScript-based mapping Application Programming Interfaces (APIs) (Geonames, 2018; OpenStreetMap, 2018). The same capabilities are possible by plotting coordinates resulting from SPARQL queries. Features selected for their specific attributes have coordinates that are plotted over a mapping product such as Leaflet (2018) or Open Layers (2018). However, geographical analysis of a range of data themes requires a wider range of attribute data in association with geospatial coordinates. To enable more reasoning with such data, advanced cartographic functions have been modeled as ontologies in the form of Web Ontology Language (OWL) files (Gould and Mackaness, 2016; Carral et.al., 2013). LD, however, maintain minimal ontology structure in favor of

handling extensive data instances. LD relies on RDF or its extension Resource Description Framework Schema (RDFS) for data structure and vocabulary reasoning, while facing system architecture challenges, particularly in the delivery and rendering of extensive geographical coordinates for geospatial feature objects. Though simpler in design, the LD graphs can be more difficult for users to easily assess because of the large number of instances. Visualization techniques based on non-coordinate attributes can be used with geospatial LD for the visualization of geographic information, but these are mostly not cartographic (EUCLID 2018). The research described here aims to improve user-access to geospatial data to explore geographically themed data with some reasoning capabilities.

## Method

Topographic data in ESRI formats were converted to RDF Triple Notation (N3) to facilitate SPARQL queries that can retrieve an initial set of data to display. For the browse-able graph, data imported from The National Map were converted to GeoJSON, that is more compatible for geospatial data, but not a native format for the LD platform, Apache Marmotta. The data to use in the Marmotta platform are stored in Postgres, a back-end database. A custom-designed servlet operating within the Marmotta platform facilitates data access between user queries originating with the map interface and the data stored in Postgres (Figure 1).

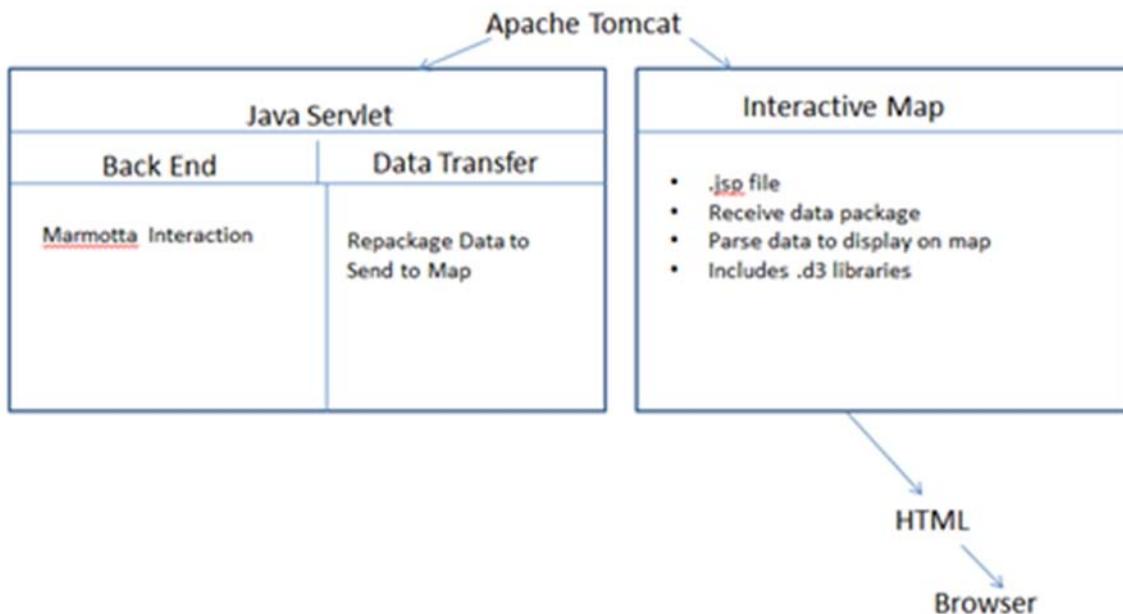


Figure 1: Servlet for exchanging LD between database and interactive map interface.

When a query is delivered to Marmotta from the servlet, HyperText Transfer Protocol (HTTP) methods such as GET are invoked. The data objects are returned in list form

consisting of strings. Upon transfer to the interactive map, a HyperText Markup Language (HTML) file uses Data Driven Document (D3) libraries that support visualization paired with JavaScript code to integrate LD features as path objects for web browser graphics (Bostock, 2017). A Scalable Vector Graphics (SVG) window opens in HTML for the code to correctly plot coordinates and display features (Dahlström et al., 2011). Once the map is opened, JavaScript event listeners wait for interactivity. When a feature is clicked, the code randomly grabs an identified element from the data file and compares it to the clicked feature ID. If the identities match, the retrieved element is saved. If they don't match, a different element is grabbed, and the cycle repeats with multiple datasets. All files are easily linkable because each contains the same value for their primary key, with the exception of the feature geometry data, which is formatted as LD in a compatible way with GeoJSON. (GeoJSON-LD has no standard and is being implemented in the LD community with ad hoc methods.)

Once all the relevant information is gathered, the program combines all the data into a point object and stores it in an array that is appended to a stylized circle object. The circle object is plotted at the recently retrieved coordinate points. This enables the efficient recall of collected information relevant to a particular visual point triggered by clicking on the point. Other geometry types, such as a flow line for hydrographic data, can also be modeled and rendered. On the data-display window, these features appear as a set of properties with literal values or additional feature instances related by properties. Stylistic elements of the data are applied using Cascading Style Sheets (CSS). The cycle begins again when further exploration is initiated by clicking on a feature.

## Results

A manual review of results indicated that the Java servlet to repackage data from the Marmotta/Postgres backend to the interactive map interface to implement map features as triples with corresponding visual data worked correctly. The visual interface directed the user's queries to arrive at specifically relevant data, bringing the satisfaction with the information search for the desired results more quickly than with comparison to queries using GIS interfaces.

Within the map interface itself, the number and order of certain functions do not restrict data gathering, but do interfere with data presentation. Further research about the interaction of graph data structures with visual elements could improve the map design capabilities for LD.

## Conclusions

A prototype system was developed that enables searching and retrieving geographical data using the browse-able graph approach for LD. Linked data from multiple local graphs using a blend of JSON-LD and GeoJSON were created from a Java servlet and passed to the graphical interface. Data linking and the representation of LD were integrated so that users can select a visual feature and retrieve additional information

associated with the feature. Clicking on information retrieved from the initial query triggered additional queries.

## References

Carral, D., Scheider, S., Janowicz, K., Vardeman, C., Krisnadhi, A.A., & Hitzler, P. (2013) An Ontology Design Pattern for Cartographic Map Scaling. In P. Cimiano, O. Corcho, V. Presutti, L. Hollink, & S. Rudolph (Eds.), *The Semantic Web: Semantics and Big Data*. ESWC 2013. Lecture Notes in Computer Science, 7882. Springer, Berlin. [https://doi.org/10.1007/978-3-642-38288-8\\_6](https://doi.org/10.1007/978-3-642-38288-8_6).

Bostock, M. (2017) D3, *Data-Driven Documents*. <https://d3js.org/> Last visited 5/1/2018.

Dahlström, E., Dengler, P., Grasso, A., Lilley, C., McCormack, C., Schepers, D., Watt, J., Ferraiolo, J., 藤沢 淳, Jackson, D. (Eds.). (2011) *Scalable Vector Graphics (SVG) 1.1*. (2nd ed.). [W3C Recommendation]. <http://www.w3.org/TR/SVG11/> Last visited 5/1/2018.

Hollink, S. Rudolph (Eds.), *EUCLID, EDUcational Curriculum for the usage of Linked Data*. ESWC 2013, LNCS 7882, pp. 76–93. Berlin: Springer-Verlag. <http://euclid-project.eu/index.html> Last visited 5/1/2018.

*GeoNames*. <http://www.geonames.org/> Last visited 5/1/2018.

Gould, N., & Mackaness, W. (2016) From taxonomies to ontologies: formalizing generalization knowledge for on-demand mapping. *Cartography and Geographic Information Science*, 43, 3, pp. 208-222.

*Leaflet, an open-source JavaScript library for mobile-friendly interactive maps*. <http://leafletjs.com/> Last visited 5/1/2018.

*OpenLayers*. <http://openlayers.org/> Last visited 5/1/2018.

*OpenStreetMap*. <http://www.openstreetmap.org/> Last visited 5/1/2018.

**Dalia E. Varanka**, Research Scientist, U.S. Geological Survey, Rolla, MO 65401

**Logan J. Powell**, Student Contractor, U.S. Geological Survey, Rolla, MO 65401

**William L. Baumer**, Student Contractor, U.S. Geological Survey, Rolla, MO 65401

## Web-Based Demo to Show a Land Use Code Ontology

Nancy Wiegand

**ABSTRACT:** This lightning talk presents a Web demo to easily view land use codes compiled and matched between Wisconsin jurisdictions. It extends prior work in matching heterogeneous land use coding systems in Wisconsin (Wiegand, 2012; Wiegand, 2016). Here, a Web demo is presented to easily show users the matches and relationships between land use codes from different jurisdictions. The demo is Web-based and directly available without a login (<http://www.ssec.wisc.edu/landuse/>). No software has to be installed or learned. This makes it easy for a user to see a readable ontology and matches. Without this demo, users would have to read OWL code directly, use visualization software that is less useful here, or load an OWL file into an ontology browser. This demo fulfills a need for an easy method for users to view and query an ontology. Although the demo is written for a particular application, namely showing how land use codes match between jurisdictions, the idea is useful for other applications and could be adapted.

The application problem that the demo addresses is that local and regional jurisdictions in Wisconsin develop their own land use codes that do not readily match each other. This makes statewide queries difficult. To solve this problem, various code sets were combined using Semantic Web technology to create a merged ontology that keeps all local codes. Keeping as much detail as possible enables precise query results rather than using standardized statewide codes (e.g., Revenue) that contain only a few general levels. Contrary to that, local codes usually contain many levels of detail and are highly specific to land uses prevalent in the area.

The demo allows selecting a land use code along with one or more jurisdictions of interest (see Figure 1 on next page). In the query result, the relationship of each jurisdiction's code to the query term is given. Such relationships are specifically presented as part of the query result to prevent users from being misled when there is not an exact match. For example, a search for 'Local Group Quarters' returns the result 'Group Quarters, Superclass of Local Group Quarters' for some jurisdictions. In general, when there is no exact match or synonym for a code from a particular jurisdiction, our algorithm first searches for a matching subclass, and if none exists, then goes up the hierarchy to find a matching superclass. If no matching superclass for that jurisdiction exists, the top-level general class is given, such as 'Residential, Superclass of Local Group Quarters'.

It is anticipated that our combined ontology and search algorithm would be part of a statewide land use query system of parcel data and that there would also be mapping so that users could query for a land use statewide and then see the parcels having that land use. A color scheme for subsets or supersets, for example, could be used to represent parcels that have codes that are related but do not exactly match the query term.

**KEYWORDS:** Land use codes, matching, ontology, viewing, Web-based, Semantic Web



Figure 1: Web interface for querying land use codes: <http://www.ssec.wisc.edu/landuse/>.

## References

- Wiegand, N. (2012) Preserving detail in a combined land use ontology. *GIScience LNCS* 7478, pp. 284–297.
- Wiegand, N. Berg-Cross, G., and Zhou, N. (2016) Resolving Semantic Heterogeneities in Land Use and Land Cover, pp. 271-294, in *Land Use and Land Cover Semantics, Principles, Best Practices, and Prospects*, editors Ola Ahlqvist, Dalia Varanka, Steffen Fritz, and Krzysztof Janowicz, CRC Press, Taylor and Francis Group, LLC.

**Nancy Wiegand**, Emeritus Scientist, University of Wisconsin-Madison, Madison, WI 53706

# MAPINS: An Intra-city PM<sub>2.5</sub> Modeling Web Application Using a Scalable Data Management and Analysis System Integrating Public Multi-source Data

**Xin Yu, Yuanbin Cheng, Yijun Lin, Yao-Yi Chiang,  
Dimitrios Stripelis and Jose Luis Ambite**

**ABSTRACT:** Air quality modeling is of great significance for studying the impact of air pollutants on human health and the urban built environment. Existing works mainly focus on applying various machine learning algorithms or physical simulations to generate models for air quality prediction based on different sources of data including geographic data, traffic emission data, remote sensing data, etc. Third-party software/tools are widely developed for air quality modeling as well as visualization. Many studies rely on third-party software for each of the processes. For example, the IBM Statistical Package for the Social Sciences (IBM SPSS) is widely used in the land use regression (LUR) model to handle the stepwise regression [1][2]. CALINE, a software package developed by the California Department of Transportation (Caltrans), is often used to provide a simulated result of the air pollution caused by road traffic (which most studies take as an explanatory variable [3][4]). The use of third-party software maybe plausible, for example, CALINE is a powerful tool when the study is on a large spatial and temporal scale (e.g., study the population impact on air quality). However, our research aims at developing a complete air quality prediction system that serves as easy access for people to be aware of the air quality within their neighborhoods. Thus, the dependency on third-party software/tools can cause “knowledge silos”, that means developers should carry the result of third-party software manually to the next step analysis. This is the case that we do not want it happens during an entire automatic system.

Existing work of evaluating fine scale air quality typically replies on area-specific and expert-selected features (e.g., geographic features) for building an air quality model, which may omit potential important factors and manually reduce the variety of useful spatial data [5]. In this paper, we present an expert-free air quality prediction system, MAPINS. The system applies a data mining approach [6] that utilizes public multi-source data, OpenStreetMap (OSM) and air pollutant data from EPA web service to automatically figure out important geographic features and generate an expert-free method to predict PM<sub>2.5</sub> concentrations at a fine spatial resolution. The fine-scale system can inform people about surrounding air quality so that people can take preventive actions in advance. Thus, our system can serve as a platform that users can easily query the air pollution statue nearby.

In addition, the required data for building an air quality model are often coming from a variety of data sources in heterogeneous formats. The data can be huge and with a high update frequency (i.e., streaming data), which requires a particular big data infrastructure for storage, access, and analytics. As a result, it remains a challenge to integrate all data processing components that start from **data acquisition, pre-processing, modeling, prediction, and visualization** as well as **result dissemination**. To handle large spatial datasets efficiently, we use Apache Spark to build a scalable data management and analysis system to achieve the high performance for air quality prediction. Our system is based on a service-oriented architecture (SOA) to cooperate all the data

processing components together to realize an automatic workflow from gathering raw data to the final display of predicting results.

**KEYWORDS:** Air quality, end-to-end system, SOA, fine scale

## References

- Briggs, D. Collins, S. Elliott, P. Fischer, P. Kingham, S. Lebret, E. Paryl, K. Reeuwijk, H. V. Smallbone, K. and Veen, A. V. E. (1997). Mapping urban air pollution using GIS: a regression-based approach. International Journal of Geographical Information Science. 11, 7 (1997), 699-718.
- Sahsuvaroglu, T. Arain, A Kanaroglou, P. Finkelstein, N. Newbold, B. Jerrett, M. Beckerman, B. Brook, J. Finkelstein, M. and Gilbert, N. L. (2006). A Land Use Regression Model for Predicting Ambient Concentrations of Nitrogen Dioxide in Hamilton, Ontario, Canada. Journal of the Air & Waste Management Association. 56, 8 (2006), 1059-1069.
- Li, L. Lurmann, F. Habre, R., Urman R., Rappaport, E. Ritz, B. Chen, J.-C Filliland, F. D. and Wu, J. (2017). Constrained Mixed-Effect Models with Ensemble Learning for Prediction of Nitrogen Oxides Concentrations at High Spatiotemporal Resolution. Environmental science & technology, 51(17), 9920-9929.
- Wilton, D. Szapiro, A. Gould, T. and Larson, T. (2010). Improving spatial concentration estimates for nitrogen oxides using a hybrid meteorological dispersion/land use regression model in Los Angeles, CA and Seattle, WA. Science of the total environment, 408(5), 1120-1130.
- Liu, C. Henderson, B. H. Wang, D. Yang, X. and Peng, Z. R. (2016). A land use regression application into assessing spatial variation of intra-urban fine particulate matter ( $PM_{2.5}$ ) and nitrogen dioxide ( $NO_2$ ) concentrations in City of Shanghai, China. Science of The Total Environment, 565, 607-615.
- Lin, Y. Chiang, Y.-Y. Pan, F. Stripelis, D. Ambite, J. L. Eckel, S. P. and Habre, R. (2017). Mining public datasets for modeling intra-city  $PM_{2.5}$  concentrations at a fine spatial resolution. In Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (p. 25).

**Xin Yu**, Master student, Spatial Sciences Institute, University of Southern California, Los Angeles, CA 90089

**Yuanbin Cheng**, Master student, Spatial Sciences Institute, University of Southern California, Los Angeles, CA 90089

**Yijun Lin**, Research Programmer, Spatial Sciences Institute, University of Southern California, Los Angeles, CA 90089

**Yao-Yi Chiang**, Associate Professor (Research), Spatial Sciences Institute, University of Southern California, Los Angeles, CA 90089

**Dimitrios Stripelis**, Ph.D. student, Information Sciences Institute, University of Southern California, Los Angeles, CA 90089

**Jose Luis Ambite**, Research Associate Professor, Information Sciences Institute, University of Southern California, Los Angeles, CA 90089

# **Identifying Crime Patterns in Mexico Using Geo-social Mining and Clustering**

**Roberto Zagal-Flores, Miguel- Felix Mata,  
Christophe Claramunt and Edgar Catalan-Salgado**

**ABSTRACT:** Nowadays, institutional crime reporting in the city of Mexico is qualitatively relatively low, which results in a very few insights for further analysis. With the rapid emergence of social networks and communities, novel avenues of research arise for studying the patterns that emerge, especially as citizens are generally keen to report, describe and comment the crimes that occur in the city. However, most of these descriptions are generally informal and per nature imprecise. These textual descriptions vary in length, context and regarding the semantics embedded. A major challenge still to address is how to extract the semantics embedded in social networks, to geo-locate the crimes that occur, and this at different scales and granularities in space and time. The objective of the research presented in this paper is to model and characterize such crime descriptions and analyze the patterns that appear in space and time. The whole approach is experimented in the city of Mexico with social data over a period of five years.

**KEYWORDS:** Geographic knowledge discovery, social perception, spatial data mining.

## **Introduction**

Although crime-based studies have been long based on institutional data recorded from public authorities the emergence of social networks now offer many new research avenues for understanding the way crimes arise and spread in a given city. While data provided by governmental bodies can provide some valuable insights on the spatial and temporal characteristics of crime patterns in an urban environment, social data and networks provide additional data analysis criteria to consider. Crime reports, complaints and reports as they appear in social network provide useful insights when one wants to explore and analyze the categories of crime that happen, where and when, as well as to explore any related contextual and socio-economical figures. Recent evolution of geographical information science, social data and big data sciences provide novel methods and technologies to analyze human patterns in large urban environments (Hardy and Maurushat, 2017). One of the main research objectives when extracting such patterns is to first identify, categorize and integrated such data in a meaningful and comprehensive way. In the specific context of the analysis of crime in the very large city of Mexico City, it appears that the rapid increase of crime has produced new forms of social organization where citizens use social networks as an option to somehow protect themselves from violence (Mata, Torres, Guzmán, Quintero, Zagal, Moreno and Loza, 2016). Citizens use social communities for different complementary purposes: to alert the community of some dangerous zones, to report events and crimes in space and time, crime complaints as well as any additional contextual data. However, the way humans use and interact with social networks is to a large extent very intuitive, imprecise, and overall very much

unstructured. Many specific research questions are opened, amongst many: what semantic characteristics define a crime complaint over a social network? How to precisely relate a given crime description to space and time, and at what level of scale and granularity? How to categorize a crime description, under which ontology? How these categories match different neighborhoods? Are there any pattern and outliers that emerge from such social participation in space and time, and does this provide some useful information at the city scale? Last but not least to which degree the patterns that emerge match institutional crime reports?

The preliminary research and experimental validation presented in this paper introduce and develop a geosocial framework whose objective is to discover additional insights derived from crime reports as extracted from a large-scale social network. The methodological framework is made of several complementary components that successively extract, categorize and cluster crime reports as embedded in a social network. The whole approach is experimented in Mexico City, the figures that appear are cross-related to official crime data as available and given by public authorities. The remainder of the paper is organized as follows. Section 2 introduces related work while Section 3 describes the methodological principles of our approach. Section 4 introduces some preliminary results while Section 5 concludes the paper and outlines further work.

### Related work

Many recent geographical big data and urban planning studies show the importance of informal social data when analyzing many socio-economical patterns as well as exploring the underlying factors that might explain them (Lee and Kang, 2015). Over the past few years, many geographic knowledge discovery methods as well as data cubes have been applied to spatio-temporal database architectures to infer additional knowledge (Han and Miller, 2009). These progresses are indeed still relevant, that is to say that the objective of social data analysis is not to replace conventional methods but rather to provide a different instantaneous, intuitive and complementary solutions to qualitatively analyze and mine crime patterns in the city (Mahboubi, 2013). In fact, social networks become relevant as means for the diffusion and sharing of large information sources when a large number of users report crimes in space and time, as well as providing complaints, opinions and further contextual social, spatial and temporal data (Chakrabarti, 2016).

The analysis of spatial big data has been already applied to the analysis of socio-economical phenomena as suggested in (Blazquez and Domenech, 2017), and where a variety of sources of data sources have been integrate to further derive socio-economical patterns, but social perception has not been so far considered. In a closely related work (Linning, 2015), the authors explored the temporal and spatial dimensions of crime categories and whether specific offences exhibit some micro-spatial patterns over time.

In (Phillips and Lee, 2012), the authors analyze crime datasets in conjunction with socio-economic and socio-demographic factors in order to explore and visualize co-distribution patterns in space and time. Data mining technics based on a Self-Organizing Map (SOM) and clustering have been also applied in (Keyvanipour, Javideh and Ebrahimi, 2011) to

extract significant events from textual police reports in plain text applying a SOM clustering method in the scope of crime analysis. Clustering results then help to perform a crime matching process.

Despite the interest of all these related work, there is still a need to explore novel methods to not only extract crime patterns from social networks but also to qualitatively and quantitatively evaluated them and cross-relate the findings to some available institutional and statistical datasets

## A GeoSocial framework to explore crime-based social data

This section introduced a methodological framework for the discovery of geographical and criminal knowledge inferred from a large social data repository. It combines a spatio-temporal data mining processes with a pipeline machine learning architecture (combination of supervised and unsupervised methods). Social data was collected from Facebook and Twitter accounts related to crime denunciation communities of areas with high crime rates in Mexico City and the State of Mexico. Data collecting process uses a semi-automatic task that analyzes the most frequent hashtags and the accounts associated with them. Some of these communities have more than 800,000 active users (e.g., Denuncia Ecatepec community). Figure 1 shows the different framework phases in detail and as introduced in our previous work (Mata et al., 2016; Zagal, Mata, Claramunt and Catalan, 2017).

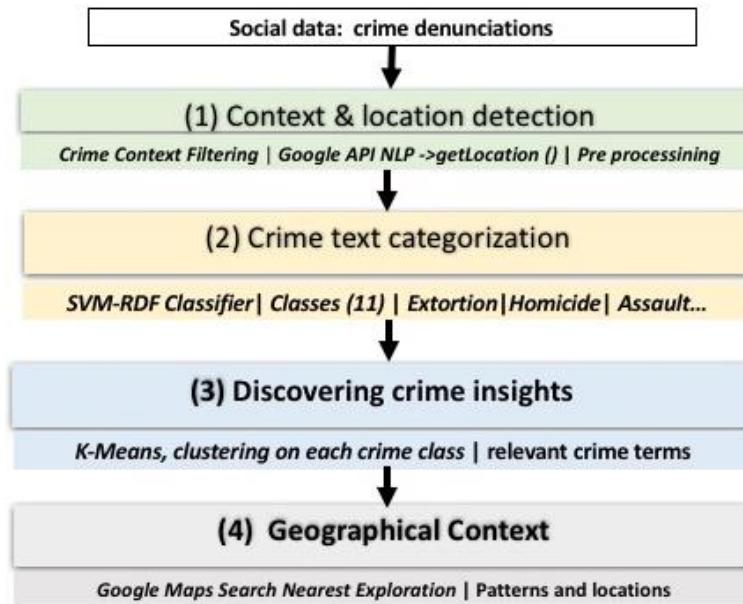


Figure 1: GeoSocial framework for social data insights discovery

The frameworks are as follows:

- 1) Contextual filtering, location detection, and pre-processing.
- 2) Text categorization using 11 types of crimes classes. An SVM classifier was developed with a training dataset of 800 rows classified manually.

- 3) Application of the clustering algorithm K-Means, which allows the discovery of key terms (several hidden places of crimes or phrases that describe *modus operandis* of crime organizations were for instance identified at this step).
- 4) When the classified texts have a spatio-temporal component, it is then possible to analyze the neighborhood of the location and time of a given crime, and eventually to derive some clusters in space and time.

### ***Context and location detection***

The textual descriptions obtained from Facebook and Twitter are filtered using a bag of keywords that conceptualizes crimes such as car theft, extortion, violent assault, etc. The bag of words was designed using the Mexican criminal laws and enriched by words and synonyms discovered in social media.

The next task identifies the location of the crime event, extracting location terms in textual description using a list of regional place names and Google Natural Language API. When a location is not founded, the crime's description is discarded. A gazetteer stores regional place names like mall names, buildings, or well-known places.

Finally, data cleansing is performed using NLTK Python, where text is converted to lower case, words are lemmatized, and punctuation marks and emoticons are eliminated (Zagal et al., 2017).

### ***Crime text categorization***

A support vector machine (SVM) classifier was used to categorize the complaints preprocessed in 11 different crimes classes: home robbery, convenience store robbery, kidnapping, rape, armed bus robbery, murder, extortion, auto theft, carjacking, pickpocketing, and armed robbery to a passerby. SVM was selected for its high performance in textual categorization benchmarks (Bishop, 2006; Abu-Mostafa, Lin and Magdon-Ismail, 2012; Zhang, Liu, Zhang and Almpandis, 2017; Srivastava and Sahami, 2009). The data training contains more than 2000 rows evenly distributed in the crime classes mentioned.

Table 1: Training dataset sample

Crime denunciation (Facebook & Twitter)	Class
AlvaroObregón: I want to denounce that my family and I were assaulted, just in the traffic light that is in front of the market, four guys with guns took advantage of a stop traffic light, then they broke the crystal and threatened my husband. The car stolen is a ...	Carjacking
I hope and my complaint is anonymous, there are some buildings with the number 9 where young people, around 10 persons, are assaulting women close to their house, they use violence and a weapon.	Armed robbery to a passerby
Hello, a few minutes ago they climbed to steal the bus that goes to the subway station on March 18, by the Martín Carrera Avenue at Gustavo a. Madero, there is a street market, there were 5 guys with pistols, only the first one paid as a passenger and the others got on quickly, several passengers took their backpacks. Alert and be careful.	Armed bus robbery

The training dataset primarily focuses on the *modus operandis* description of the crime, and does not give extensive details on the temporal-spatial dimensions in order to enlarge the grade of the generalization during the classification process (Table 1).

In training classification phase, we obtained a performance of 83 percent. It would seem that this performance is not high, however, it allows to obtain a balanced generalization model that avoids overfitting, allowing reach a better performance of 91 percent in the test phase. We tuned the classifier using a Radial Basis Function kernel with the value modified parameters: "gamma" equal to 0.1, and "c" equal to 100, the values were iteratively obtained using the GridSearchCV function of Scikit-learn python.

The parameters value was selected according to training performance and RBF SVM parameters specification (Scikit-learn, 2018) that establishes: The parameter "gama" defines how far the influence of a single training example reaches, we obtained a low values (0.01) that meaning 'far', "gamma" can be seen as the inverse of the radius of influence of samples selected by the model as support vectors. The C parameter trades off misclassification of training examples against simplicity of the decision surface, we obtained a high C (100) that aims at classifying all training examples correctly by giving the model freedom to select more samples as support vectors.

### ***Discovering crime insights***

In order to discover crime insights, we selected K-means clustering algorithm due to its scalability and testing performance (Srivastava and Sahami, 2009). It was implemented using Scikit-learn Python. After location detection and crime text categorization, in this step, K-means uses as an input each set of classified posts, it is executed 11 times, then K-means obtains sets of key words for each crime class. The TF-IDF matrix approach was used to obtain the terms frequency (Sparck Jones, 1972).

Clustering process was implemented using NLTK and Scikit-learn Python. We tuning and tested the algorithm using a rank from 10 to 50 clusters generated, from 5 to 10 iterations. The initial size tested to start the clustering was around of 100 classified posts for each class crime.

### ***Geographical Context***

The goal is to explore the crime patterns that emerge in space and time. For instance, one might explore how crime categories match some specific neighborhoods and/or some given period of time, how crime categories are clustered or not in space and time.

For instance, it appears that carjacking frequently occurs in Mexico City near schools or places with little traffic, and near avenues with a speed limit greater than 40 kilometers. The exploration process is mainly based on an interactive mapping and visualization process whose objective is to present crime categories clusters as well as points of interest. Crime patterns are filtered by locations and time periods. For example, specific buffers are generated around some places of interest (banks, schools, ATMS, malls, churches, hospitals). The whole implementation has been performed using a combination of Google APIs and Java scripts (JSON, jQuery). Correlations between places of interests

and specific crime categories are then explored. Figure 2 illustrates an example of patterns where different extortion crimes are related to a specific point of interest.

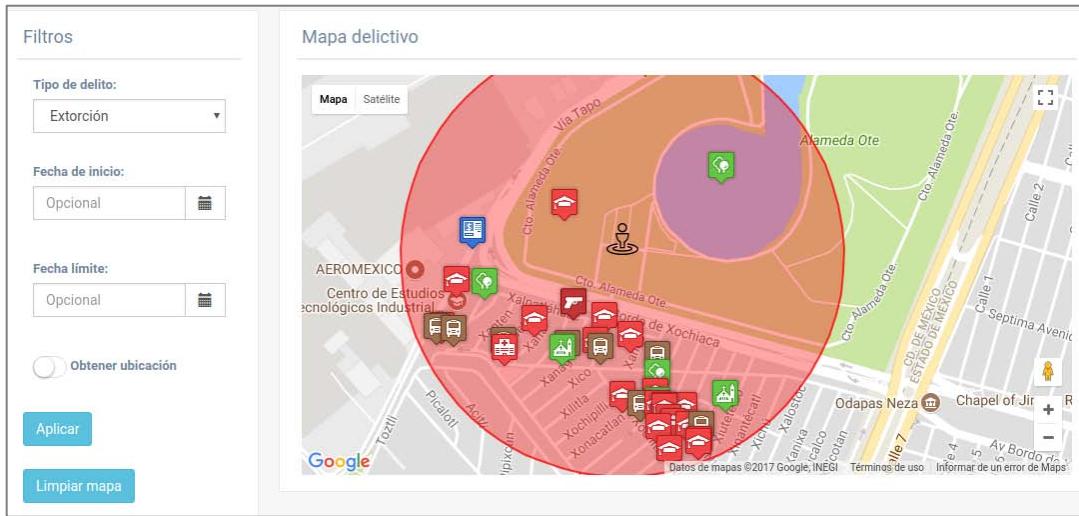


Figure 2: Geographic contextualization of extortion crimes for a place of interest

Overall, we extracted 42,733 posts from Twitter and Facebook from January 2013 to November 2017. At least 50 accounts were used: communities (39), news media (4), and government accounts (7). We detected more than 15 important social communities regarding the recollection and dissemination of high-impact crime reports: kidnapping, carjacking, assault, etc.

In figure 3, we observed that these social communities had an increase of activity from September to October 2017, due to in that months it was recollected an important amount of social data. This activity possibly is related to the increase of rates of murder, car theft and robbery that were reported by the Mexican government in 2017 (SESNP, 2018).

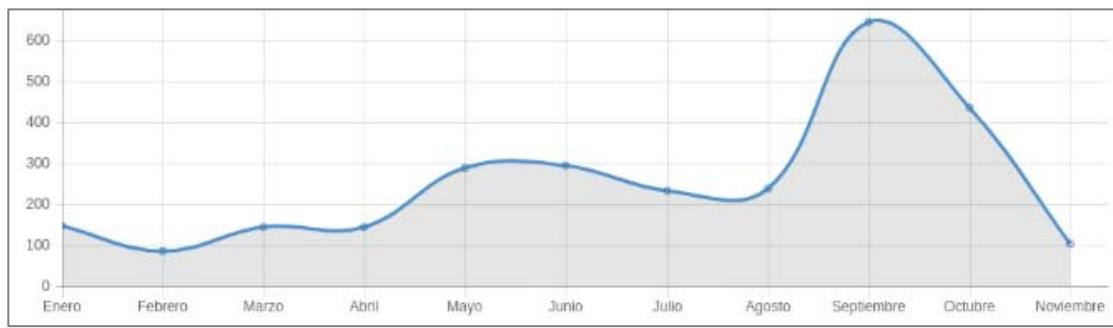


Figure 3: Distribution of recollected texts (from January to November 2017)

## Results

## ***Geographical text categorization***

After the contextualization process, 36696 publications were considered for text categorization; 60% of them have at least one location. From the social data considered (see figure 4), the top crimes were murder (4500), auto theft (more than 2000), carjacking (more than 1000), armed bus robbery (1500), rape (500), and armed robbery of a passerby (250). The results obtained from the classification match relatively well with the official statistics of crimes provided by the national government for the year 2017.

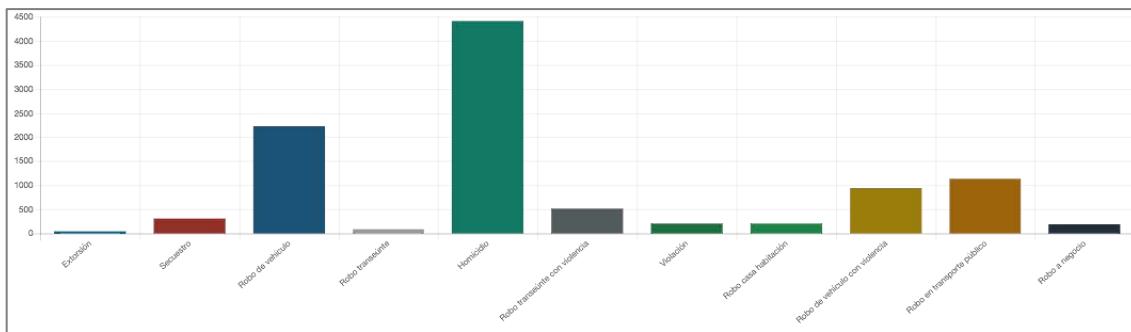


Figure 4. Distributions of recollected texts (texts from left to right: extortion, kidnapping, auto theft, pickpocket, murder, armed robbery to a passerby, rape, home robbery, carjacking, armed bus robbery, and convenience store robbery)

## *Clustering & insights*

Figure 5 shows the results on the execution of the k-means algorithm that uses all posts re-collected as an input. The five highest bars indicate the groups that contain a high number of posts; the green color represents ordinary groups. The "number of plate" group (the highest bar) is a descriptor term related to complaints of "car theft" crimes. The "help" group is another term used by social network complaints, and other interesting group is the term "man" that is related to crime descriptions of *modus operandi*.

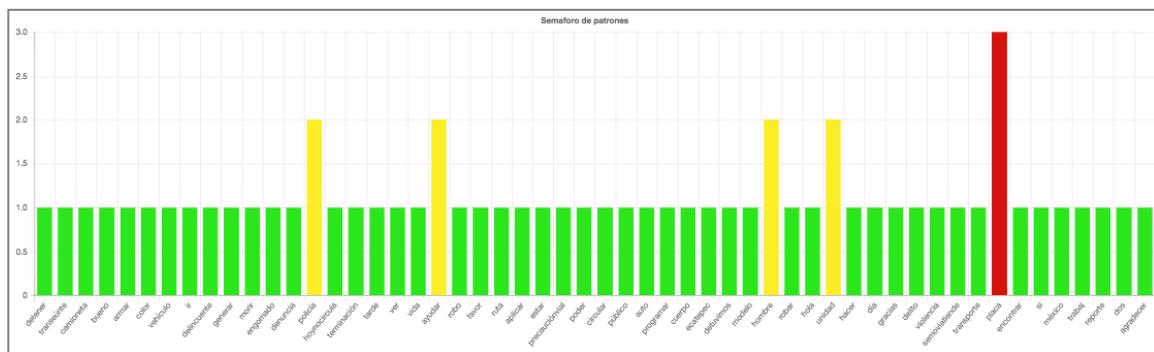


Figure 5: Text clustering using the full set of classified crime posts

Figure 6 shows the most often discovered terms only for posts classified in carjacking and murder classes. For example, we used 965 posts classified in carjacking class, as the

input of the clustering process, then it discovered that “license plate” is the cluster that most groups posts of this class, due to the “number of license plate” is frequently provided when victims report carjacking in social media. Regarding of murder class, the term “police” clusters to 72 posts related to this class, because newspapers constantly report police murders. When geo-located, these patterns allow to discover some places not reported in official statistics where a certain type of crime occurs (for example parks, hospitals and schools).

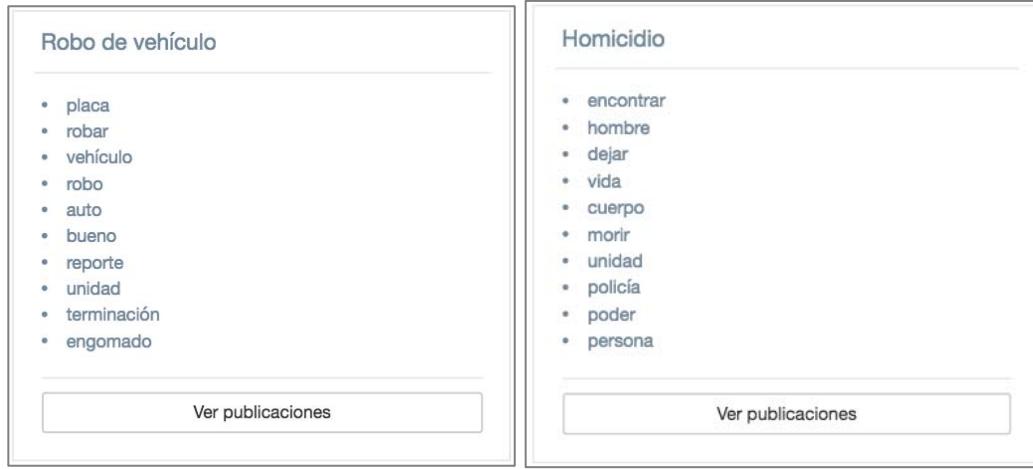


Figure 6: Text clustering results for posts classified in carjacking and murder classes.

### ***Crime social maps***

A sample of 610 crimes, extracted in 2017, are mapped in Fig. 7. There are extracted from 449 posts categorized into 11 crime classes; 161 are correctly classified and geo-located while others are not due to a lack of information to relate them to a specific crime category. Our crime social map contains information of more than 400 classified posts (cf. the map is available at [goo.gl/BX7TJd](http://goo.gl/BX7TJd)), the data corresponds to the mentioned period of extraction (from 2013 to 2017). It appears that the higher number of high-impact crimes is concentrated in the North of Mexico. This social perception coincides with the official statistics provided by the government of Mexico (SESNP, 2018).

### ***Geographical Context***

Using the previous data set, we obtained the geographical characterization of the three most frequent crimes: murder, carjacking and armed bus robbery as well as the closest and most common places where a crime has occurred. Murder and carjacking generally occur near bus stations, while armed bus robbery and carjacking occur near avenues with a speed limit greater than 40 kilometers per hour. See Table 2.

Table 2: Geographical contextualization of crimes

Crime	Geographic context
Murder	Schools, bus stations, convenience stores
Carjacking	Bus stations, avenues
Armed bus robbery	Avenues, convenience stores

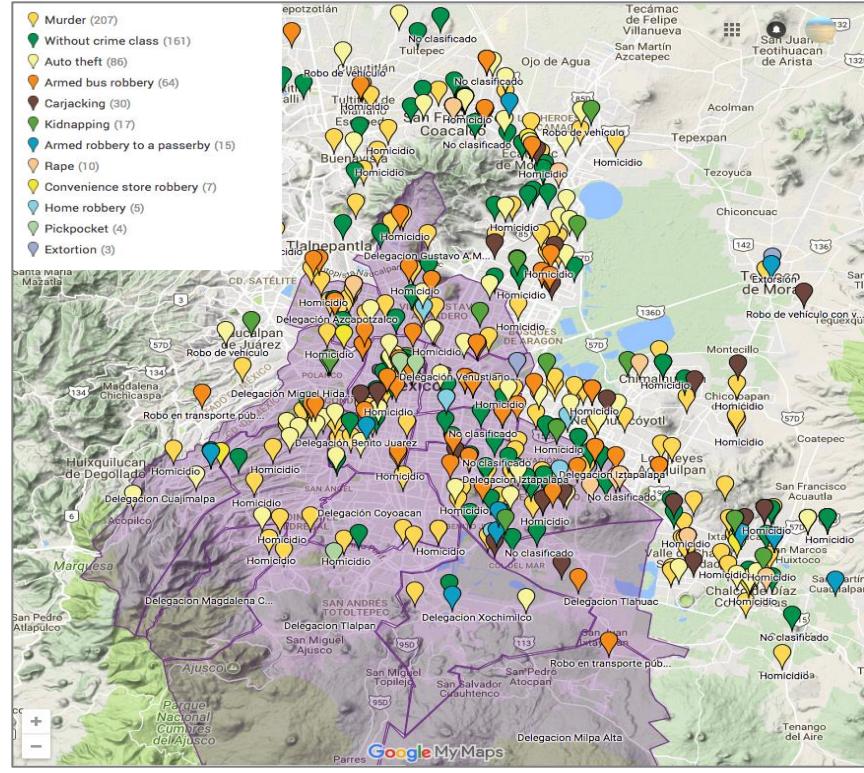


Figure 7: Social crime map of Mexico City and the State of Mexico

## Conclusions

The preliminary research presented in this paper introduced a methodology approach for geographic knowledge discovery applied to crime descriptions as they appear in a series of social media as delivered by Facebook and Twitter. The whole approach is based on the application of a machine-learning architecture pipeline, context and location detection, clustering and geographical analysis. Overall, this exploratory approach based on human perceptions as expressed by social media complements crime figures provided by local governments. While official data mostly describe and locate crime categories in space and time, our approach provides much more capabilities when exploring the relationships between the crime that happen and the geographical space, particularly when relating crimes with places of interest as well as additional contextual information as they appear in crime textual descriptions (e.g., key terms such as warnings). Overall this work establishes a preliminary foundation to integrate complementary heterogeneous data sources: institutional data, social media data and geographical data.

Future work will consider additional spatial analysis experiments such as the exploration of additional spatio-temporal clustering methods and exploratory visualization techniques. Amongst different directions to explore we would like to develop a crime forecasting algorithm potentially based on probabilities to predict crime events in space and time. Finally, we would like to further analyze the social patterns that appear from our study with previous reports on social security of crime perception in Mexico City (INEGI, 2017).

**Acknowledgments:** The authors of this paper thank God, CONACYT project number 1051, the Laboratorio de Cómputo Móvil-UPIITA, COFAA-IPN, SIP-IPN project 20181759, Shanghai Maritime University and Instituto Politécnico Nacional (IPN) for their support.

## References

- Hardy, K., & Maurushat, A. (2017). Opening up government data for Big Data analysis and public benefit. *Computer law & security review*, 33(1), 30-37.
- Han, J., & Miller, H. J. (2009). Geographic data mining and knowledge discovery. CRC Press.
- Chakrabarti, A. S. (2016). Cross-correlation patterns in social opinion formation with sequential data. *Physica A: Statistical Mechanics and its Applications*, 462, 442-454.
- Lee, J. G., & Kang, M. (2015). Geospatial big data: challenges and opportunities. *Big Data Research*, 2(2), 74-81.
- Blazquez, D., & Domenech, J. (2017). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*.
- Mata, F., Torres-Ruiz, M., Guzmán, G., Quintero, R., Zagal-Flores, R., Moreno-Ibarra, M., & Loza, E. (2016). A Mobile Information System Based on Crowd-Sensed and Official Crime Data for Finding Safe Routes: A Case Study of Mexico City. *Mobile Information Systems*, 2016.
- Zagal-Flores, R., Mata, F., Claramunt, C., & Catalan-Salgado, E. (2017). Discovering geographical patterns of crime localization in Mexico City. *WEB 2017 : The 5<sup>th</sup> International Conference on Building and Exploring Web Based Environments*.
- Mahboubi, H., Bimonte, S., Deffuant, G., Chanet, J. P., & Pinet, F. (2013). Semi-automatic design of spatial data cubes from simulation model results. *International Journal of Data Warehousing and Mining (IJDWM)*, 9(1), 70-95.
- Bishop, C. M. (2006). Machine learning and pattern recognition. *Information Science and Statistics*. Springer, Heidelberg.
- Abu-Mostafa, Y. S., Lin, H. T., & Magdon-Ismail, M. (2012). Learning from Data: A Short Course: AMLbook. *View Article PubMed/NCBI Google Scholar*.
- Zhang, C., Liu, C., Zhang, X., & Almanidis, G. (2017). An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications*, 82, 128-150.
- Srivastava, A. N., & Sahami, M. (2009). *Text mining: Classification, clustering, and applications*. Chapman and Hall/CRC.

SESNSP. (2018) Secretariado de seguridad pública.  
<http://secretariadoejecutivo.gob.mx/incidencia-delictiva/incidencia-delictiva-fuero-comun.php>. Last visited 01/30/2018

INEGI. (2017) Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública 2017,  
<http://www.beta.inegi.org.mx/proyectos/enchogares/regulares/envipe/2017/> Last visited 01/30/2018

Linning, S. J. (2015). Crime seasonality and the micro-spatial patterns of property crime in Vancouver, BC and Ottawa, ON. *Journal of Criminal Justice*, 43(6), 544-555.

Phillips, P., & Lee, I. (2012). Mining co-distribution patterns for large crime datasets. *Expert Systems with Applications*, 39(14), 11556-11563.

Keyvanpour, M. R., Javideh, M., & Ebrahimi, M. R. (2011). Detecting and investigating crime by means of data mining: a general crime matching framework. *Procedia Computer Science*, 3, 872-880.

Scikit-learn.org. (2018). *RBF SVM parameters — scikit-learn 0.19.1 documentation*.  
[http://scikit-learn.org/stable/auto\\_examples/svm/plot\\_rbf\\_parameters.html#sphx-glr-auto-examples-svm-plot-rbf-parameters-py](http://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html#sphx-glr-auto-examples-svm-plot-rbf-parameters-py) Last visited 01/10/2018.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1), 11-21.

**Roberto Zagal-Flores**, PHD student, Instituto Politécnico Nacional, UPIITA-IPN, Ciudad de México, México

**Miguel Felix Mata-Rivera**, Professor, Instituto Politécnico Nacional, UPIITA-IPN, Ciudad de México, México

**Christophe Claramunt**, Professor, Shanghai Maritime University, Haigang Ave,Shanghai, P. R. China. christophe

**Edgar Catalan-Salgado**, Professor, Instituto Politécnico Nacional, ESCOM-IPN, Ciudad de México, México

Community Resilience in Maricopa County, Arizona, USA:  
The Analysis of Indoor Heat-related Death and Urban Thermal Environment  
**Qunshan Zhao, Heather Fischer, Wei Luo and Elizabeth A. Wentz**

**ABSTRACT:** The SEER (Social, Economic, and Environmental Resilience) Knowledge Exchange is an effort at Arizona State University to integrate resilience data collected from community stakeholders, social media, citizen science, and local and federal authoritative organization to identify and mitigate resilience threats to Maricopa County of Arizona, USA. This paper focuses on a case study of SEER that aims to understand the communitive factors leading to indoor heat-related death in Maricopa County, AZ. The authoritative data we have include daytime land surface temperature and vegetation coverage (NDVI) from remotely sensed images, cooling center locations, demography data from U.S. census, parcel and house age, tree numbers, tree canopy coverage, urban park, as well as the census-tract level indoor heat-related death data. With all of these local and federal authoritative data, we attempt to understand what demographic, residential, and urban infrastructure factors influence the indoor heat-related death and how we can reduce indoor heat-related health issues in Maricopa County by GIS and regression analysis. The research results show a negative relationship between indoor heat-related death and outdoor land surface temperature. More vegetation coverage could cool down the neighborhood and reduce the potential indoor heat-related death. The elderly, the poor, and children are more vulnerable to urban heat. The research results will provide a guideline for the next phase of urban thermal environment enhancement and urban green infrastructure improvement in the Maricopa County, and help mitigate urban heat for vulnerable populations and reduce the happen of indoor heat-related death in the future. In terms of future research, a citizen science project will recruit heat vulnerable population to track their heat exposure both qualitatively and quantitatively. Volunteers will be asked to carry air temperature sensors and GPS sensors with them for a week, combining with a detailed social survey to better understand the complex factors that lead individuals and families to need utility assistance or become more vulnerable to heat. This case study along with other aspects of the Knowledge Exchange is used to characterize the social, economic, and environmental vulnerabilities of the individuals in Maricopa County.

**KEYWORDS:** community resilience, indoor heat-related death, authoritative data, GIS

**Author names and affiliations:**

**Qunshan Zhao**, Postdoctoral Research Associate, Spatial Analysis Research Center, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85287

**Heather Fischer**, Postdoctoral Research Associate, Spatial Analysis Research Center, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85287

**Wei Luo**, Postdoctoral Research Associate, Spatial Analysis Research Center, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85287

**Elizabeth A. Wentz**, Professor, Spatial Analysis Research Center, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85287

# Deep Convolutional Neural Networks for Map-Type Classification

Xiran Zhou, Wenwen Li, Samantha T. Arundel and Jun Liu

**ABSTRACT:** Maps are an important medium that enable people to comprehensively understand the configuration of cultural activities and natural elements over different times and places. Although massive maps are available in the digital era, how to effectively and accurately access the required map remains a challenge today. Previous works partially related to map-type classification mainly focused on map comparison and map matching at the local scale. The features derived from local map areas might be insufficient to characterize map content. To facilitate establishing an automatic approach for accessing the needed map, this paper reports our investigation into using deep learning techniques to recognize seven types of map, including topographic map, terrain map, physical map, urban scene map, the National Map, 3D map, nighttime map, orthophoto map, and land cover classification map. Experimental results show that the state-of-the-art deep convolutional neural networks can support automatic map-type classification. Additionally, the classification accuracy varies according to different map-types. We hope our work can contribute to the implementation of deep learning techniques in cartographical community and advance the progress of Geographical Artificial Intelligence (GeoAI).

**KEYWORDS:** Deep learning, deep convolutional neural network, map-type classification

## Introduction

Maps are an important medium that enable people to comprehensively understand the configuration of cultural activities and natural elements over different times and places. Map features, such as text and geographical features, are used to benefit the representation and communication in cartography and the GIScience community (Lloyd and Bunch, 2003). In the last two decades, Internet and spin-off techniques have significantly changed the nature of map generation and the use of maps (Hurst and Clough, 2013). Due to the advent of web-based service technologies, cyberinfrastructure, and volunteered geographic information (Li, Yang and Yang, 2010), a number of online platforms and tools such as Google Maps, Bing Maps, and Wikimapia are available for map creation, visualization, and geospatial analysis. Currently, maps are not used by geographical domain experts to conduct geospatial computing and analysis. The information included in maps allows the public to better facilitate daily activities such as ridesharing, delivery, and transportation network analysis, to name just a few.

Although a great number of maps are available in the digital era, how to effectively and accurately access the required map remains a challenge today. Three problems have left this challenge unsolved. First, a majority of maps available from the Internet lack map elements like map title, direction indicators, or legends, leaving the reader to rely only on

the map frame itself in order to interpret the content. Second, new techniques enable the creation of immense map repositories containing maps with diverse themes, configurations and designs. This diversity increases the difficulty in accessing appropriate maps. Third, unlike the objects in a photograph or image, which are easily characterized, it is impossible to precisely characterize maps that are short their defining map elements. For example, topographic maps and road maps may both contain road networks and streets.

To our knowledge, no literature with respect to map-type classification has yet been reported. Previous works partially related to map-type classification mainly focused on map comparison and map matching at the local scale (Power, Simms and White, 2001; Li and Huang, 2002; Fritz and See, 2005; Zhu, Y., et al., 2017). The local map scales defined by these approaches include pixels, pixel blocks (or superpixels), and polygons (or map objects) (Stehman and Wickham, 2011). However, features derived from local map areas might be insufficient to characterize map content, since overlaps can always be observed in different types of maps. For instance, it is possible to observe water in both ocean maps and topographic maps. An automated approach ensures the availability of needed maps and is essential to facilitate the role of maps in geographical analysis and public activities. Thus, this paper reports our investigation into using deep learning techniques to recognize different types of map.

## Method

### *Dataset*

The datasets for map-type classification are selected from a benchmark called *deepMap*. We created this benchmark to provide datasets for studying automatic map classification with deep learning techniques. All data in *deepMap* are collected from the online maps of ArcGIS, Google Maps, the National Map of the USGS, and other online search engines. *deepMap* offers three types of benchmark datasets: a map text dataset, a text-labeled map dataset, and a map dataset. In this paper, we use the map dataset to conduct map-type classification with Deep Convolutional Neural Networks (DCNNs). The dimensionality of each image in the map dataset is  $256 \times 256 \times 3$ , where  $256 \times 256$  denotes image size and 3 refers to the RGB channels of an image. *deepMap* contains seven available map categories: 1) topographic map/terrain map/physical map, 2) urban scene map, 3) the National Map, 4) 3D map, 5) nighttime map, 6) orthophoto map, and 7) land cover classification map (Figure 1 next page). Each map category contains around 200 maps in total, and the total number of maps in the map dataset is 1812.

### *Deep convolutional neural networks*

Previous machine learning (ML) approaches, or “shallow ML,” have foundered when handling complex functions and features, and generally require substantial labor in training data to obtain satisfactory results (LeCun, Bengio and Hinton, 2015). Deep learning (DL) approaches enable computers to spontaneously access highly valuable information through unsupervised learning, and discover the high-level representations of

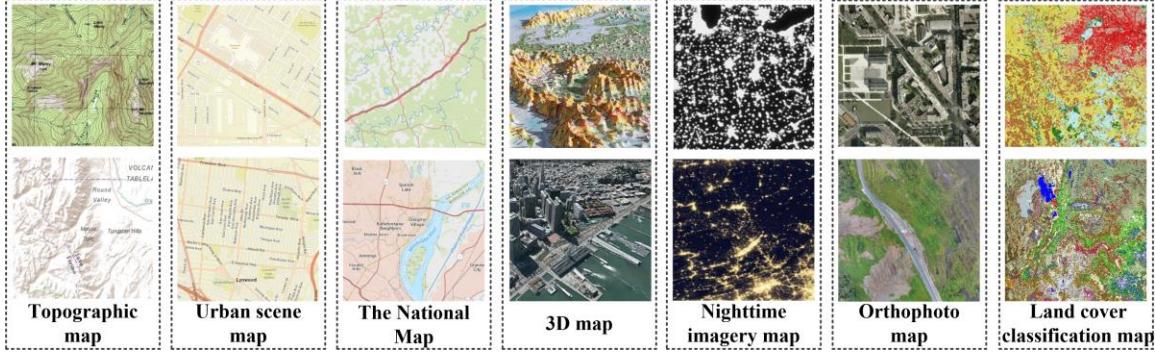


Figure 1: Illustration of the seven types of maps found in the *deepMap* dataset.

data based on a multi-layered processing framework. In the last five years, a large number of DCNNs have produced impressive image classifications. Thus, we intend to apply DCNNs to classify map-types based on the map content when metadata and other auxiliary information (map title, map legend, etc.) are not available. Figure 2 shows a general architecture of a DCNN.

A DCNN is a class of multi-layered neural networks designed to exploit features of image or speech signals, which generally consist of two parts: feature generation and classification. The input image is an image that has RGB channels. Feature generation includes a number of convolutional layers and pooling layers (or unsampling layers) to produce a feature map that includes high-level representation of the input image. Using the resulting convolutional layer, the classification includes fully-connected or densely-connected layer(s) and a classifier (e.g. softmax) to classify the input image as one of the predefined classes.

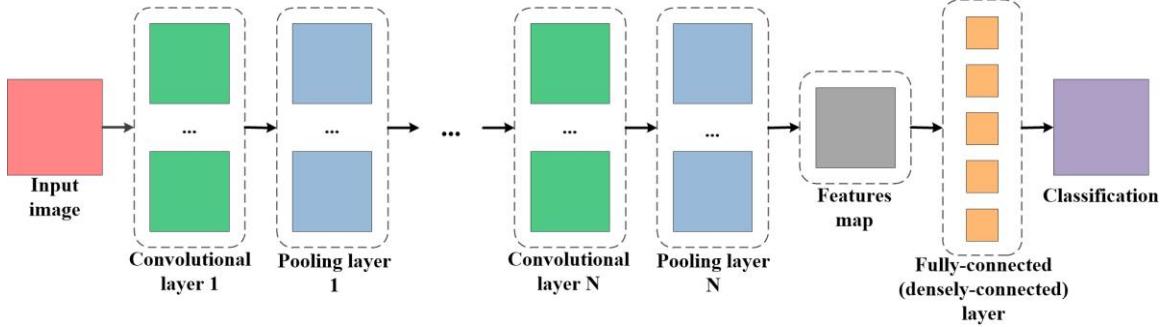


Figure 2: General architecture of a deep convolutional neural network.

### ***Implementation details***

The default size of input image varies according to different DCNNs. Thus, two data augmentation approaches: image rotation and image rescaling are used to preprocess the image. First, all maps in *deepMap* are rescaled to the size fitting for a DCNN. Then, we randomly divided recalled maps into training dataset and test dataset. For the training

dataset, image rotation creates three new maps by rotating the original one through 90, 180, and 270 degrees, respectively.

We have selected state-of-the-art DCNN models to test their performance of map-type classification. Each DCNN model is listed below.

- *AlexNet* (Krizhevsky, Sutskever and Hinton, 2012)
- *VGG Net* (Simonyan and Zisserman, 2014)
- *GoogleNet or Inception* (Szegedy, et al., 2015)
- *ResNet* (He, et al., 2016)
- *Inception-ResNet* (Szegedy, et al., 2017).

## Results and Discussions

Each DCNN has some influential improvements and designs. The objective of our experiments is to test whether and how much these improvements and designs effectively facilitate automatic map-type classification. Since DCNNs are sensitive to the quality and amount of training data, we make classifications according to different ratios of training data and test data. Table 1 lists the details of the experimental design and experimental results.

Table 1: Experimental design and results for comparing various (D)CNN methods.

CNNs & DCNNs	<i>Experimental results</i>	
	<i>Group 1: 60% data used for training &amp; 40% data used for testing</i>	<i>Group 2: 80% data used for training &amp; 20% data used for testing</i>
AlexNet	71% ~ 78%	77% ~ 83%
VGG Net-19	73% ~ 80%	78% ~ 84%
Inception V4	82% ~ 87%	88% ~ 94%
ResNet V2-152	82% ~ 86%	89% ~ 93%
ResNet-Inception V2	88% ~ 92%	95% ~ 99%

The experimental results have shown that the classification accuracies generated by these CNNs and DCNNs ranged from around 70% to 98%. Moreover, increasing the volume of training data would significantly raise the performance of DCNNs in map-type

classification. This phenomenon supports the claim that it is critical to prepare large-scale well-labeled data to feed a neural network for enhancing its capability to distinguish different classes (Bengio, Courville and Vincent, 2012).

In detail, although AlexNet has been replaced by many later DCNNs in image classification, this pioneering CNN still enables the production of an acceptable result in map-type classification. The VGG method resulted in higher accuracy than did AlexNet, which supports the claim that the depth of a neural network is much more crucial than its spatial dimensions (Szegedy, et al., 2015). Moreover, GoogleNet or Inception organizes a very deep architecture of DCNN, which markedly improves classification accuracy and computational efficiency. However, the very deep network may produce higher accuracy results, but training is very difficult for a DCNN with a very deep architecture. ResNet proposed residual blocks to revolutionize the trade-off between efficient training and deep architecture in DCNNs. The results produced by ResNet prove that this strategy is also useful to facilitate map-type classification. To maintain the advantages of deep neural network and computational efficiency, Inception-ResNet integrates two compelling networks, inception network (Inception) and deep residual network (ResNet), as a unified and simplified architecture. This integrated DCNN produced the highest accuracy in image classification when a large-scale training dataset is available. Generally speaking, DCNNs can support the automated classification of the majority of map types in *deepMap*. However, the classification accuracy varies according to different map types. Some types of maps, such as the National Map's topographic maps, are difficult to distinguish without a well-labeled dataset.

Besides deep learning techniques, knowledge of map design and map generation are potential means to facilitate automatic map-type classification. Recently, the significance of transferring knowledge has been shown to substantially improve the performance of DCNNs by some edge-cutting models like NASNet (Zoph et al. 2017) and PNASNet (Liu et al., 2017). In the future, a DCNN that supports the transfer of knowledge, and high-level features of different maps will be our focus. We hope our work can contribute to the implementation of deep learning techniques in the cartographical community, and advance the progress of GeoAI.

## References

- Fritz, S. and See, L. (2005). Comparison of land cover maps using fuzzy agreement. *International Journal of Geographical Information Science*, 19(7), 787-807.
- He, K., et al. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition* (pp. 770-778).
- Hurst, P. and Clough, P. (2013). Will we be lost without paper maps in the digital age?. *Journal of Information Science*, 39(1), 48-60.
- Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).
- LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436.

- Li, W., Yang, C. and Yang, C. (2010). An active crawler for discovering geospatial web services and their distribution pattern—A case study of OGC Web Map Service. *International Journal of Geographical Information Science*, 24(8), 1127-1147.
- Li, Z. and Huang, P. (2002). Quantitative measures for spatial information of maps. *International Journal of Geographical Information Science*, 16(7), 699-709.
- Liu, C., Zoph, B., Shlens, J., Hua, W., Li, L. J., Fei-Fei, L., Yuille, A., Huang, J., & Murphy, K. (2017). Progressive neural architecture search. *arXiv preprint arXiv:1712.00559*.
- Luchetta, S. (2017). Exploring the literary map: An analytical review of online literary mapping projects. *Geography Compass*, 11(1), e12303.
- Lloyd, R. and Bunch, R. L. (2003). Technology and Map-Learning: Users, Methods, and Symbols. *Annals of the Association of American Geographers*, 93(4), 828-850.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Stehman, S. V. and Wickham, J. D. (2011). Pixels, blocks of pixels, and polygons: Choosing a spatial unit for thematic accuracy assessment. *Remote Sensing of Environment*, 115(12), 3044-3055.
- Szegedy, C., et al (2015, June). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-5).
- Szegedy, C., et al (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI* (Vol. 4, p. 12).
- Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2017). Learning transferable architectures for scalable image recognition. *arXiv preprint arXiv:1707.07012*.
- Zhu, Y., et al. (2017). A similarity-based automatic data recommendation approach for geographic models. *International Journal of Geographical Information Science*, 31(7), 1403-1424.
- Zhu, Y., Zhu, A. X., Feng, M., Song, J., Zhao, H., Yang, J., ... & Yao, L. (2017). A similarity-based automatic data recommendation approach for geographic models. *International Journal of Geographical Information Science*, 31(7), 1403-1424.

**Xiran Zhou**, PhD student, School of Geographical Sciences & Urban Planning, Arizona State University, Tempe, AZ 85281

**Wenwen Li**, Associate Professor, School of Geographical Sciences & Urban Planning, Arizona State University, Tempe, AZ 85281

**Samantha T. Arundel**, Research Geographer, Center of Excellence in Geographic Information Science, U.S. Geological Survey, Rolla, MO 65401

**Jun Liu**, Associate Researcher, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong Province, P.R. China 518055