Cédric Grueau
Jorge Gustavo Rocha (Eds.)

Communications in Computer and Information Science 582

Geographical Information Systems Theory, Applications and Management

First International Conference, GISTAM 2015
Barcelona, Spain, April 28–30, 2015
Revised Selected Papers

Springer

Communications in Computer and Information Science 582

Commenced Publication in 2007
Founding and Former Series Editors:
Alfredo Cuzzocrea, Dominik Ślęzak, and Xiaokang Yang

Editorial Board

More information about this series at http://www.springer.com/series/7899

Cédric Grueau · Jorge Gustavo Rocha (Eds.)

Geographical Information Systems Theory, Applications and Management

First International Conference, GISTAM 2015
Barcelona, Spain, April 28–30, 2015
Revised Selected Papers

Springer

Editors
Cédric Grueau
Escola Superior de Tecnologia de Setúbal
Setúbal
Portugal

Jorge Gustavo Rocha
Universidade do Minho
Braga
Portugal

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-319-29588-6 ISBN 978-3-319-29589-3 (eBook)
DOI 10.1007/978-3-319-29589-3

Library of Congress Control Number: 2016930051

Printed on acid-free paper

This Springer imprint is published by SpringerNature
The registered company is Springer International Publishing AG Switzerland

Preface

The present book includes extended and revised versions of a set of selected papers from the First International Conference on Geographical Information Systems Theory, Applications and Management (GISTAM 2015), held in Barcelona, Spain, during April 28–30, 2015, which was sponsored by the Institute for Systems and Technologies of Information, Control and Communication (INSTICC) in cooperation with the ACM SIGSPATIAL, the Global Spatial Data Infrastructure Association (GSDI), the Canadian Institute of Geomatics and Cartographic, the Geological Institute of Catalonia (ICGC), and the International Society for Photogrammetry and Remote Sensing (ISPRS). GISTAM 2015 was also technically co-sponsored by the IEEE Geoscience and Remote Sensing Society.

The purpose of the International Conference on Geographical Information Systems Theory, Applications and Management was to create a meeting point for researchers and practitioners addressing new challenges in geospatial data sensing, observation, representation, processing, visualization, sharing, and managing, in all aspects concerning information communication and technologies (ICT) as well as management information systems and knowledge-based systems.

GISTAM 2015 received 45 paper submissions from 25 countries in all continents, of which 24 % were presented at the conference as full papers, and their authors were invited to submit extended versions of their papers for this book. In order to evaluate each submission, a double-blind review was performed by the Program Committee. Finally, only the ten best papers were included in this book.

We would like to highlight that GISTAM 2015 also included four plenary keynote lectures, given by internationally distinguished researchers, namely: Robert Laurini (INSA, University of Lyon, France), José Bioucas Dias (Telecommunications Institute, Portugal), Jordi Corbera (Geological and Cartographic Institute of Catalonia, Spain), and Ed Parsons (Google UK Ltd., UK). We must acknowledge the invaluable contribution of all keynote speakers, who, as renowned researchers in their areas, presented cutting-edge work, thus contributing to enriching the scientific content of the conference.

We must thank the authors, whose research and development efforts are recorded here. We also thank the GISTAM local chair Joaquín Huerta, whose competence was essential for ensuring the technical quality of the conference and whose collaboration was very much appreciated. The knowledge and diligence of the reviewers were essential for the quality of the papers presented at the conference and published in this book. Finally, a special thanks to all members of the INSTICC team, whose involvement was fundamental for the success of this conference.

April 2015

Cédric Grueau
Jorge Gustavo Rocha

Organization

Conference Chair

Cédric Grueau	Polytechnic Institute of Setúbal/IPS, Portugal

Program Chair

Jorge Gustavo Rocha	University of Minho, Portugal

Local Chair

Joaquín Huerta	Jaume I University, Spain

Program Committee

Ana Paula Afonso	Universidade de Lisboa, Portugal
Masatoshi Arikawa	The University of Tokyo, Japan
Pedro Arnau	Universitat Politècnica de Catalunya, Spain
Thierry Badard	Laval University, Canada
Rex G. Cammack	University of Nebraska in Omaha, USA
Manuel Campagnolo	Instituto Superior de Agronomia, Portugal
Keith Clarke	University of California, Santa Barbara, USA
Tonie M. van Dam	Université du Luxembourg, Luxembourg
Maria Luisa Damiani	Università degli Studi di Milano, Italy
Jeff Dozier	University of California Santa Barbara, USA
Suzana Dragicevic	Simon Fraser University, Canada
Ana Paula Falcão	Instituto Superior Técnico, Portugal
Manfred M. Fischer	Vienna University of Economics and Business, Austria
Ana Fonseca	Laboratório Nacional de Engenharia Civil (LNEC), Portugal
Efi Foufoula-Georgiou	University of Minnesota, USA
Lianru Gao	Chinese Academy of Sciences, China
Georg Gartner	Vienna University of Technology, Austria
Peter Gerstoft	University of California San Diego, USA
Luis Gomez-Chova	Universitat de València, Spain
Michael Gould	Universitat Jaume I, Spain
Kingsley E. Haynes	George Mason University, USA
Haosheng Huang	Vienna University of Technology, Austria
Andrew Hudson-Smith	University College London, UK
Karsten Jacobsen	Leibniz Universität Hannover, Germany

Ingensand Jens	University of Applied Sciences Western Switzerland, Switzerland
Bin Jiang	University of Gävle, Sweden
Simon Jirka	52 North, Germany
Harry D. Kambezidis	National Observatory of Athens, Greece
Marinos Kavouras	National Technical University of Athens, Greece
Waldo Kleynhans	Council for Scientific and Industrial Research, South Africa
Alexander Klippel	The Pennsylvania State University, USA
Andreas Koch	University of Salzburg, Austria
Wei-Shinn Ku	Auburn University, USA
Jun Li	Sun Yat-Sen University, China
Christophe Lienert	Canton of Aargau, Department of Construction, Traffic and Environment, Switzerland
Chang-Tien Lu	Virginia Tech, USA
Yannis Manolopoulos	Aristotle University, Greece
Andre Marçal	Universidade do Porto, Portugal
Paulo Marques	Instituto de Telecomunicações/ISEL, Portugal
Janet Mersey	University of Guelph, Canada
Richard Milton	University College London, UK
Lan Mu	University of Georgia, USA
Daniele Perissin	Purdue University, USA
Mathieu Roche	Cirad, France
Armanda Rodrigues	Universidade Nova de Lisboa, Portugal
Anne Ruas	IFSTTAR, France
Markus Schneider	University of Florida, USA
Yosio Edemir Shimabukuro	Instituto Nacional de Pesquisas Espaciais, Brazil
Francesco Soldovieri	Consiglio Nazionale delle Ricerche, Italy
Uwe Stilla	Technische Universität München, Germany
Laura Toma	Bowdoin College, USA
Michael Vassilakopoulos	University of Thessaly, Greece
Miguel A. Vengazones	GIPSA-Lab, CNRS, France
Jan Oliver Wallgrün	The Pennsylvania State University, Germany
Ouri Wolfson	University of Illinois at Chicago, USA
Xiaojun Yang	Florida State University, USA
May Yuan	University of Texas at Dallas, USA
Shuqing Zhang	Northeast Institute of Geography and Agroecology, CAS, China

Additional Reviewer

Xiang Xu	Guangdong Key Laboratory for Urbanization and Geo-Simulation, China

Invited Speakers

Robert Laurini	INSA, University of Lyon, France
José Bioucas Dias	Telecommunications Institute, Portugal
Jordi Corbera	Geological and Cartographic Institute of Catalonia, Spain
Ed Parsons	Google UK Ltd., UK

Contents

Reasoning Geo-Spatial Neutral Similarity from Seismic Data Using Mixture and State Clustering Models

Avi Bleiweiss(✉)

Platform Engineering Group, Intel Corporation, Santa Clara, CA, USA
avi.bleiweiss@intel.com

Abstract. Conventionally, earthquake events are recognized by guided and well established geographical region confines. However, explicit regional schemes are prone to overlook patterns manifested by cross-boundary seismic relations that are regarded vital to seismological research. Rather, we investigate a statistically motivated system that clusters earthquake impacted places by similarity in seismic feature space, and is hence impartial to geo-spatial proximity constraints. To facilitate our study, we have acquired hundreds of thousands recordings of earthquake episodes that traverse an extended time period of forty years. Episodes are split into groups singled out by their affiliated geographical place, and from each, we have extracted objective seismic features expressed in both a compact term-frequency of scales format, and as a discrete signal representation that captures magnitude samples spaced in regular time intervals. Attribute vectors of the distributional and temporal domains are further applied towards our mixture model and Markov chain frameworks, respectively, to conduct clustering of presumed unlabeled, shake affected locations. We performed comprehensive cluster analysis and classification experiments, and report robust results that support the intuition of geo-spatial neutral similarity.

Keywords: Earthquake · Seismic · Mixture model · Hidden Markov model · Expectation-maximization · Clustering · k-nearest neighbors

1 Introduction

Modern seismological exploration of disseminating earthquake sites and magnitudes rests on both the advancement in instrumental seismometry and the analysis of macroseismic effects, including geological structures, population, and the landscape [12]. To describe the severity of shaking, seismic effects are commonly assigned an intensity scale set by different yet fairly correlated standards, traditionally in the range of one to ten. In recent years, the development of methods to quantitatively and objectively analyze scale data, coupled with the emerging of online systems to generate unprecedented volumes of both real-time and archival earthquake data, had sparked renewed interest in research to

C. Grueau and J.G. Rocha (Eds.): GISTAM 2015, CCIS 582, pp. 1–21, 2016.
DOI: 10.1007/978-3-319-29589-3_1

assess global seismic intensity distribution. One indispensable resource for practitioners in the field is the United States Geological Survey (USGS) [27] science organization, fully devoted to furnish impartial information on the health of our ecosystem. Amongst the many services, USGS provides a large web based repository of geo-spatially rich data for expressing earthquake events that are dynamically collected as they occur, and furthermore allows for this knowledge base to be programmatically accessible for software development. Figure 1 shows a high level, distributed earthquake scale around the globe, based on USGS data we acquired that reproduces an extent of four decades, from 1975 till 2014.

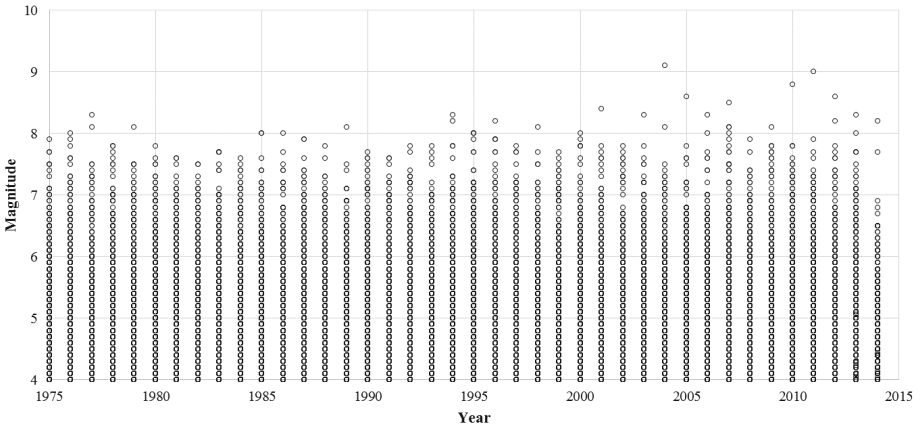


Fig. 1. Earthquake events: showing magnitude scale as a function of time, tracking forty years of activity from 1975 till 2014.

In our work, we investigate a discovery [23] method that extracts a statistical relation model of earthquake bound geographical locations from a large data set of hundreds of thousands of recorded seismic events, and incorporates both information retrieval [16] and unsupervised machine learning [7] techniques. Information retrieval (IR) is rapidly becoming the dominant form of data source access. Amongst multitude disciplines, IR encompasses the field of grouping a set of documents that enclose non structured content, to behave similarly with respect to relevance to information needs. Our work closely leverages IR practices by realizing a seismic bound place after a text document, composed of a collection of intensity scales and represented in a compact histogram of term frequencies format. For a broader context, we contrast this distribution feature form with a classic, discrete seismic signal constructed of a series of shake magnitudes over time that spans a course of several tens of years. Furthermore, we are interested in uncovering objectively the underlying cluster nature of hundreds of geographical sites, without resorting to any prior knowledge of the erupting physical location, nor to constraining geo-spatial proximity as prescribed by the Flinn-Engdahl regionalization scheme [29]. To this extent, we use both finite

mixture [18] and Markov chain [22] models, recognized for providing effective and formal statistical framework to cluster high dimensional data of continuous nature.

Finite mixture models are widely used in the field of cluster analysis [9,10], and apply to a growing application space including web content search, gene expression linking, and image segmentation. They form an expressive set of classes for multivariate density estimation, and the entire observed data set of scale histograms is represented by a mixture of either continuous or discrete, parametric distribution functions. An individual distribution, often referred to as a component distribution, constitutes thereof a cluster. Traditionally, the likelihood paradigm provides a mechanism for estimating the unknown parameters of the mixture model, by deploying a method that iterates over the maximum likelihood. One of the more broadly used and well behaved technique to guarantee process convergence is the Expectation-Maximization (EM) algorithm [6] that scales well with increased data set size. Upon completion, the likelihood function reflects the conformity of the model to the incomplete observed data. While not immediately applicable to our work, noteworthy is the research that further extends the empirical likelihood paradigm to a model, whose component dimension is unknown. Hence, both model fitting and selection must be determined from the data simultaneously, by using an approximation based on any of the Akaike Information Criterion (AIC) [1], the Bayesian Information Criterion (BIC) [25], or the sum of AIC and BIC plus an entropy term [20].

The discrete Hidden Markov Model (HMM) [4,22] is a probabilistic framework that formalizes a reasoning about a series of observations over time, to recover a set of states. The model is extensively used in many application domains including speech recognition, biological sequence analysis, and stochastic natured financial economics. HMM is described by a set of parameters that are estimated to maximize the probability of an observation. Much like the mixture model, it employs the maximum likelihood estimation principle and commonly uses the Baum-Welch algorithm [3], an analog to the EM method. In our work, we characterize a time progression of earthquake events, occurring in a prescribed geographical location, as a discrete seismic signal comprised of shake scale samples. A seismic signal thus forms an observation vector, and a collection of these temporal feature vectors are applied to HMM, deriving for each a log-likelihood measure. Unlike the mixture model, the grouping of observation vectors in HMM is not implicit, hence we follow HMM to perform hierarchical agglomerative clustering [13,17] on log-likelihood values.

The main contribution of our work is a novel, statistically driven system that combines IR and unsupervised learning techniques to discover instinctive cluster patterns from presumed unlabeled seismic data, and best match earthquake bound geographical locations by objective similarity in feature space. In contrast to a more constraining approach that prescribes physical regionalization boundaries. Figure 2 provides an overview of the sequence of logical steps that constitute our learning framework. The remainder of this paper is organized as follows. We overview the motivation for selecting seismic feature representations,

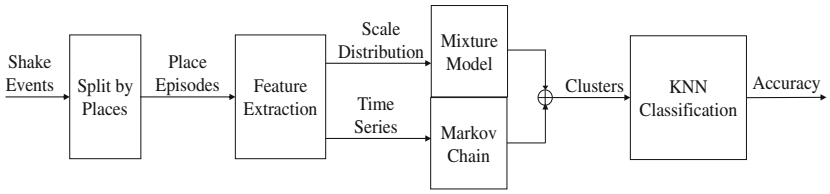


Fig. 2. The input to our learning framework is a set of separable earthquake events. Events are split by physical geographical location to form a collection of variable length, place episodes. From the place data, we follow by extracting both scale distribution and time series seismic features that feed into a mixture model and a Markov chain, respectively. Generated clusters of shake bound locations are analyzed and cross validated in a k-nearest neighbor (KNN), baseline classifier.

leading to our compact formats of intensity scale distribution and a time series signal, in Sect. 2. Section 3 reviews algorithms and provides theory to multivariate cluster analysis, discussing both the normal mixture model and Markov chain foundations, and the role of their respective EM method in estimating model parameters. Whereas in Sect. 4, we present our evaluation methodology of seismic cluster analysis and classification, and report quantitative results of our experiments. We conclude with a discussion and future prospect remarks, in Sect. 5.

2 Seismic Features

We acquired seismic data from the USGS [27] science organization. USGS provides real-time earthquake data in a well-structured format, GeoJSON [11], readily parsed by most programming languages. GeoJSON uses the popular JavaScript Object Notation (JSON) to encode a diverse set of geographic data structures. A GeoJSON object may represent any of a geometry, a feature or a collection of features. Typically, an earthquake event is characterized by a geometrical bounding box and a set of seismic features (Table 1). The three dimensional volume of eruption is defined by the minimum and maximum extent of each of the latitude, longitude, and depth attributes, and a rather extensive set of seismic properties are specified, although many of them appear either unavailable or partially missing in the data frames we gathered. Most relevant features to our work include the magnitude, magnitude type, place, and time. The magnitude value is measured and recorded by a seismograph that responds to distinct seismic waves traveling through the ground, who are excited by relative motion of the earth. Whereas magnitude type identifies the method or algorithm to calculate the scale of the event. Most commonly used scales comprise of local (M_l), also referred to the Richter scale, surface-wave (M_s), body-wave (M_b), and moment (M_w) metrics. Moment scale is directly related to the faulting process and is considered a more consistent measure of earthquake size, unlike the rest that are accuracy limited by an upper bound. Nonetheless, all magnitude types

Table 1. Features extracted from a GeoJSON object that describes geometry and selected seismic properties of an earthquake event. Showing for each the value range and measurement units or data type.

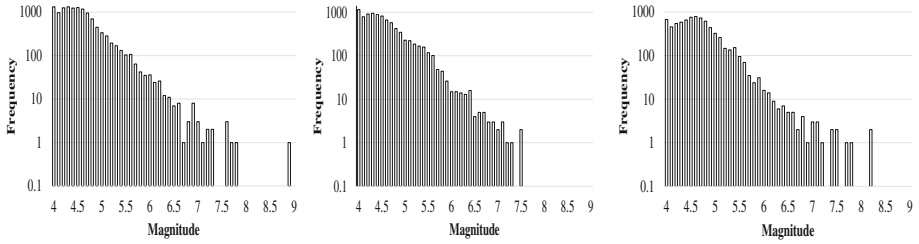
(a) Bounding Box.

Dimension	Range	Units
latitude	$[-90.0^\circ, +90.0^\circ]$	degrees
longitude	$[-180.0^\circ, +180.0^\circ]$	degrees
depth	$[0, 1000]$	kilometer

(b) Seismic Properties.

Feature	Range	Units
mag	$[-1.0, 10.0]$	scale
mag type	M_l, M_s, M_b, M_w	string
place	Flinn-Engdahl region	string
time	date-time	milliseconds

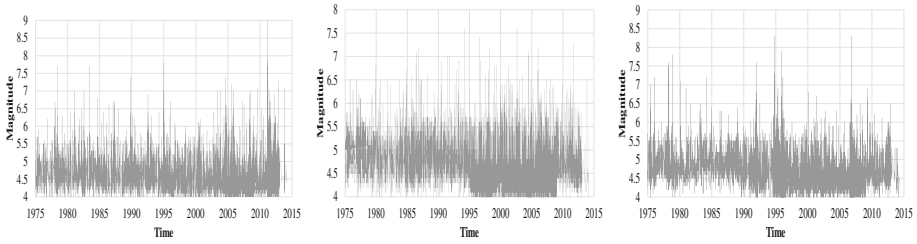
yield approximately the same value for a given earthquake event. The place property is a named geographic location closest to the event, either a city or a region enumerated in the Flinn-Engdahl seismic and geographical, globe partitioning scheme [8], along with the time of a shake occurrence, reported in milliseconds.



(a) Honshu, Japan. (b) Fiji region. (c) Kuril Islands.

Fig. 3. Scale distribution feature vector: showing in log scale the number of magnitude occurrences extracted from events associated with a place data point, for three geographical locations.

Our seismic dataset comprises several hundreds of thousands earthquake events that track an extended time period of several tens of years. The process of extracting features from this large seismic collection proceeds in several stages. First, we split the dataset into groups, each embedding all event occurrences in an identical place or region, chronologically. Our transformed dataset represents now a compilation of distinct places drawn out from our raw data, and totals several thousands data points. Let $P = \{p_1, p_2, \ldots, p_n\}$ be our observed, place subjected seismic data, with each place data point, p_i, retaining a different event count. Next, we derive from each data sample, p_i, two domain feature vectors to provide for unified dimensionality. An unnormalized, scale distribution vector $D \in \mathbb{N}^{|V|}$, with $|V|$ the number of possible magnitude values, and a time series vector $S \in \mathbb{R}^d$ of a sampling dimensionality, d.



(a) Honshu, Japan. (b) Fiji region. (c) Kuril Islands.

Fig. 4. Time series feature vector: resampled irregular raw signal using the year-week sampling mode, for three geographical locations. A sample of no event is assigned a zero magnitude.

D formalizes a term frequency description that assigns each vector element a count of unique magnitude occurrences, accumulated in the events prescribed to a place data point, p_i. This is modeled after the *bag of words* [2] representation, a simple and one of the more effective text retrieval methods, founded on the premise that the respective order of events to emerge in a location, is ignored. In our work, we tend to events who record a scale in the $[4.0, 9.9]$ range and sampled in 0.1 increments, hence $|V|$, the dimensionality of D, amounts to 60 elements. Figure 3 outlines scale distribution feature vectors extracted from three distinct, place data points. The location compact format of bag of scale words is passed on to our mixture model to perform seismic place clustering, and follows efficient similarity calculations, directly from the well known Vector Space Model [24].

The raw, time series vector we extract is an irregular periodicity formulation of magnitudes, dispersed sequentially along the course of our event capturing time frame of forty years. S is then further resampled with regular time intervals consisting of year-week, monthly and bi-monthly formats, thus leading to a time series feature vector of discretized dimensionality depicted in Table 2. Our year-week sample index, $[0, 53]$, follows the US rule, and for a place of multiple events, excited in the same week, we compute a weekly mean of all magnitudes to ensure a single scale is identified with a week. Whereas a week of no event defaults to the value of zero intensity. The monthly and bi-monthly sampling modes arise from a direct decimation of the year-week signal by a factor of four and eight, respectively. Time series vectors, resampled in the year-week mode for three geographical places are further illustrated in Fig. 4. Subsequently, we use a hidden Markov chain (HMM) to model the durational and spectral variability of our generated seismic signal, S, that constitutes an observation vector.

3 Place Clustering

Clustering procedures based on finite mixture models provide a flexible approach to multivariate statistics. They become increasingly preferred over heuristic methods, owing to their robust mathematical basis. Mixture models standout

Table 2. Time series vector: listing our three uniform resampling modes and for each the corresponding feature dimensionality.

Year-week	Monthly	Bi-monthly
2120	530	265

in admitting clusters to directly identify with the components of the model. To model our system probability distribution of scale count features, we deploy the well established, Normal (Gaussian) Mixture Model (GMM) [18,19], known for its parametric, probability density function that is represented as a weighted sum of Gaussian component densities. GMM parameters are estimated from our incomplete training data, composed of bags of intensity scale words, using the iterative Expectation-Maximization (EM) [6] algorithm. Correspondingly, for our place bound, seismic signal features we exploit the Hidden Markov Model (HMM) [4,22], using the Baum-Welch algorithm [3] to repeatedly recalibrate model parameters, and follow this process to construct an agglomerative hierarchy of seismic aware clusters, employing an efficient dynamic tree cutting technique.

3.1 Normal Mixture Model

Let $X = \{x_1, x_2, \ldots, x_n\}$ be our observed collection of seismic bound places, each represented as an intensity scale, term frequency vector $I \in \mathbb{N}^d$. An additive mixture model, defines a weighted sum of k components, whose density function is formulated by Eq. 1:

$$p(x|\Theta) = \sum_{j=1}^{k} w_j p_j(x|\theta_j), \tag{1}$$

where w_j is a mixing proportion, signifying the prior probability that an observed place x, belongs to the j^{th} mixture component, or cluster. Mixing weights satisfy the constraints $\sum_{j=1}^{k} w_j = 1$, and $w_j \geq 0$. The component probability density function, $p_j(x|\theta_j)$, is a d-variate distribution, parameterized by θ_j. Most commonly, and throughout this work, $p_j(x|\theta_j)$ is the multivariate normal (Gaussian) density (Eq. 2), characterized by its mean vector $\mu_j \in \mathbb{R}^d$ and a covariance matrix $\Sigma_j \in \mathbb{R}^{dxd}$. Hence, $\theta_j = (\mu_j, \Sigma_j)$, and the mixture parameter vector $\Theta = \{\theta_1, \theta_2, ..., \theta_k\}$.

$$\frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma_j|}} \exp\left(-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\right) \tag{2}$$

Seismic places, distributed by mixtures of multivariate normal densities, are members of clusters that are centered at their means, μ_j, whereas the cluster geometric feature is determined by the covariance matrix, Σ_j. For efficient processing, our covariance matrix is diagonal, $\Sigma_j = diag(\sigma_{j1}^2, \sigma_{j2}^2, \ldots, \sigma_{jd}^2)$, and thus clusters are of an ellipsoid shape, each nonetheless of a distinct dimension. To fit

the normal mixture parameters onto a set of training feature vectors, we use the maximum likelihood estimation (MLE) principle. Furthermore, in regarding the set of seismic places as forming a sequence of n independent and identically distributed data samples, the likelihood corresponding to a k-component mixture, becomes the product of their individual probabilities:

$$L(\Psi|X) = \Pi_{i=1}^{n} \sum_{j=1}^{k} w_j p_j(x_i|\theta_j), \tag{3}$$

where $\Psi = \{\Theta, w_1, w_2, \ldots, w_k\}$. However, the multiplication of possibly thousands of fractional probability terms, incurs an undesired numerical instability. Therefore, by a practical convention, MLE operates on the log-likelihood basis. As a closed form solution to the problem of maximizing the log-likelihood, the task of deriving Ψ analytically, based on the observed data X, is in many cases computationally intractable. Rather, it is common to resort to the standard, expectation-maximization (EM) algorithm, considered the primary tool for model based clustering.

To add more flexibility in describing the distribution $P(X)$, the EM algorithm introduces new independences via k-variate hidden variables $Z = \{z_1, z_2, \ldots, z_n\}$. Hidden variables mainly capture uncertainty in cluster assignments, and are estimated in conjunction with the rest of the parameters. The combined observed and hidden portions form the complete data set $Y = (X, Z)$, where $z_i = \{z_{i1}, z_{i2}, \ldots, z_{ik}\}$ is an unobserved vector, with indicator elements

$$z_{ic} = \begin{cases} 1, & \text{if } x_i \text{ belongs to cluster } c \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

EM is an iterative procedure, alternating between the expectation (E) and maximization (M) steps. For the hidden variables z_i, the E step estimates the posterior probabilities w_{ic} that a place object x_i belongs to a mixture cluster c, given the observed data and the current state of the model parameters

$$w_{ic} = \frac{w_c p_c(x_i|\mu_c, \Sigma_c)}{\sum\limits_{j=1}^{k} w_j p_j(x_i|\mu_j, \Sigma_j)}. \tag{5}$$

Then the M step maximizes the joint distribution of both the observed and hidden data, and parameters are fitted to maximize the expected log-likelihood, based on the conditional probabilities, w_{ic}, computed in the E step. The E step and M step are iterated until convergence or up to a set limit of iterations, after which a scale distribution feature vector, x_i, is assigned to a cluster, corresponding to the highest conditional or posterior probability of its membership. EM typically performs well once the observed data reasonably conforms to the mixture model, and by ensuring robust selection of random values assigned to starting parameters, the algorithm warrants convergence to either a local maximum or a stationary value.

3.2 Hidden Markov Model

The Hidden Markov Model (HMM) [4,22] formulates an effective statistical framework to describe time varying processes of physical systems. HMM is a stochastic model of a signal that at regularly spaced time samples undergoes state transitions conforming to a set of probabilities identified for each state. HMM models the joint probability of a collection of the random variables $O = \{o_1, o_2, \ldots, o_T\}$ and $Q = \{q_1, q_2, \ldots, q_T\}$, over time T. O comprises a set of discrete event observations in a time series feature vector. An observation takes one of M possible symbols $\in \{v_1, v_2, \ldots, v_M\}$, expressed by the magnitudes $\in \{4.0, 4.1, \ldots, 9.9\}$ along with the value zero to mark a no-event element, thus making the vocabulary size $M = 61$. Q is hidden, with each its elements set to one of N admissible states $\in \{1, 2, \ldots, N\}$. Under the discrete Markov chain, there are two conditional independence assumptions about these random variables that make related algorithms tractable. Namely, the t^{th} hidden variable only depends on the $(t-1)^{st}$ hidden variable, and the t^{th} observation solely rests on the t^{th} state. These hypotheses resonate well with our seismic signal, constructed of loosely coupled and independent place events. We further assume that the underlying hidden Markov chain, defined by $P(Q_t|Q_{t-1})$, is time homogeneous and represented as a stochastic transition matrix $A = \{a_{ij}\} \in \mathbb{R}^{NxN}$, where $a_{ij} = P(Q_t = j|Q_{t-1} = i)$. Time $t = 1$ is deemed a special case specified by the initial state distribution $\pi_i = P(Q_1 = i)$. Respectively, the probability of an observation symbol at time t for state j is expressed by the emission matrix $B = \{b_j(v_t)\} \in \mathbb{R}^{MxN}$, where $b_j(v_t) = \{P(O_t = v_t|Q_t = j)\}$. Parametrically, an HMM is compactly represented as $\lambda = (A, B, \pi)$, and our goal is to solve the HMM learning problem for each of our observed, place constructed seismic signals, by maximizing the probability of an observation vector O, $P(O|\lambda)$, and iteratively estimating the model parameters.

Akin to the EM algorithm used for mixture models, we adopted the Baum-Welch (BW) method [3] to find the maximum likelihood estimation of the HMM parameters, for each of our generated, shake signal vectors. The method starts by choosing arbitrary values for the model parameters. It then proceeds to compute the forward probability, $\alpha_i(t)$, for the partial observation $\{o_1, \ldots, o_t\}$ ending in state i at time t, and the backward probability, $\beta_i(t)$, for the complementary sequence $\{o_{t+1}, \ldots, o_T\}$ that started on state i, at time $(t+1)$. Time T is bound to the resampling mode set for the time series vectors, matching the sizes depicted in Table 2. Both $\alpha_i(t)$ and $\beta_i(t)$ are calculated efficiently using recursion. The algorithm then creates two auxiliary variables: $\gamma_i(t)$ (Eq. 6) as the probability of being in state i at time t, normalized over the entire observed symbols

$$\gamma_i(t) = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N} \alpha_j(t)\beta_j(t)}, \tag{6}$$

and $\xi_{ij}(t)$ (Eq. 7) representing the joint probability of being successively in state i at time t and in state j at time $(t+1)$, normalized for the integrated feature

vector

$$\xi_{ij}(t) = \frac{\alpha_i(t)a_{ij}\beta_j(t+1)b_j(o_{t+1})}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i(t)a_{ij}\beta_j(t+1)b_j(o_{t+1})}. \tag{7}$$

From γ and ξ, the definition of intuitive update rules to the model parameters ensues, as shown in Eqs. 8, 9, and 10, respectively

$$\pi_i = \gamma_i(1), \tag{8}$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1}\xi_{ij}(t)}{\sum_{t=1}^{T-1}\gamma_i(t)}, \tag{9}$$

$$b_j(k) = \frac{\sum_{t=1}^{T}\delta_{O_t,v_k}\gamma_j(t)}{\sum_{t=1}^{T}\gamma_j(t)}. \tag{10}$$

In the BW algorithm, the steps of computing forward and backward probabilities, calculating γ and ξ, and updating model parameters, repeat finite times or until convergence is reached. For each of the seismic signal vectors, the procedure returns the log-likelihood value that we further use for hierarchical clustering.

3.3 Agglomerative Merge

Unlike a mixture model, there is no implicit clustering directly derived from HMM. Therefore, the log-likelihood values we computed for each of our observed seismic vectors, serve as input for further feature matching grouping. We opted for hierarchical clustering [13,17] over flat data structures, with the former intended for more detailed data analysis, and found agglomerative grouping more intuitive in our design compared to the divisive approach. The clustering algorithm starts with each individual geographical place as its own cluster, and successively combines clusters that are most similar. This process builds a tree topology from bottom-up and is repeated until it reaches the root node that merges all of our seismic places. For n geographical places we compute an nxn matrix of similarity coefficients, and update the matrix as the hierarchy is constructed. We chose the Euclidean distance as the similarity metric, and applied a subset of the most commonly used linkage methods [14] that determine how clusters are merged. Similarity functions must obey monotonicity to warrant the operation of merging does not increase similarity, and furthermore be agnostic to the merge order. Linking procedures along with their corresponding formulas are further listed in Table 3. The single linkage measures the distance between nearest neighbors of the combined clusters, whereas the complete procedure evaluates the two farthest member points. In average mode, the mean distance of all inter-group pairs is computed, and for the Ward minimum variance method [28],

Table 3. Dissimilarity formulas in merging clusters A and B, for selected linkage methods (d - distance, c - centroid).

Linkage method	Cluster dissimilarity
Single	$\min_{a \in A, b \in B} d[a, b]$
Complete	$\max_{a \in A, b \in B} d[a, b]$
Average	$\frac{1}{\lvert A \rvert \lvert B \rvert} \sum_{a \in A} \sum_{b \in B} d[a, b]$
Ward	$\sqrt{\frac{2\lvert A \rvert \lvert B \rvert}{\lvert A \rvert + \lvert B \rvert}} \lvert\lvert c_A - c_B \rvert\rvert_2$

notably is its tendency to join clusters with a small number of observations, and be strongly biased towards producing clusters of roughly the same size.

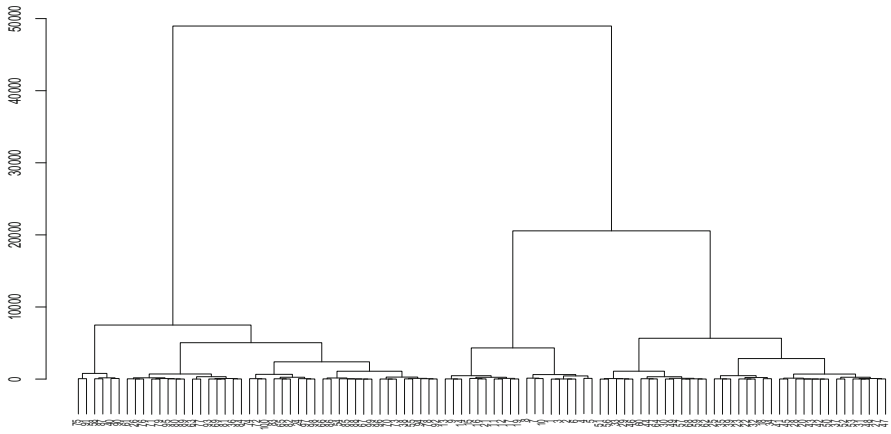


Fig. 5. Agglomerative clustering dendrogram shown for the top one hundred seismic locations of the highest event count. This process uses the Ward linkage method, known for producing a more evenly cluster distribution.

Our hierarchical clustering implementation exploits a dynamic tree cut method that expands on the work by Langfelder et al. [15], and detects a set of coherent groups, each with its correlated seismic features of shake affected places. We use an adaptive branch height approach to generate a user defined, number of clusters. The algorithm respects the order of merges encountered in building our tree, and for each similarity measure it traverses the tree in a top-down manner, until the number of clusters desired becomes stable. Starting at the root node that represents a single cluster, the search descends the tree nodes comparing for each its similarity measure to a provided adaptive threshold. The subtrees of a successful horizontal cut are then explored down to their leaf nodes to extract their corresponding geographical seismic places. Agglomerative clustering is typically visualized as a dendrogram, shown in Fig. 5 for our top one

hundred seismic places of the highest event count. The dendrogram depicted resulted from employing the Ward linkage method, known to form more evenly distributed clusters. Graphically, each merge is represented by a horizontal line, and the y coordinate of the horizontal line is the similarity measure of the two clusters that were merged.

4 Empirical Evaluation

To validate our system in practice, we have implemented a software library that realizes the cluster analysis of seismic places in several stages. After collecting and cleaning the archived earthquake data, our library commences with extracting both static and dynamic, location based feature vectors. They take the formulation of scale distribution and temporal signals, successively fed into our mixture and Markov chain models, respectively. Our features are regarded as unlabeled, and follow either an implicit or explicit clustering. Constructed groups of places are then contextually contrasted against a standardized, seismic regionalization scheme [8].

4.1 Experimental Setup

Our work exploits the R programming language [21] to acquire the raw earthquake data and further clean it to serve useful in our software environment. We have managed to retrieve from USGS a total of 326,267 recorded events occurred in a forty years interval that started on the first year-week of 1975 till the fourteenth year-week in 2014. Shake events are spread across 3,247 geographical places, however 1,300 of those are affected by a single incident, and additional 1,579 sites enumerate under 100 episodes. To reason statistically for conducting cluster analysis, this leaves out then 368 places of sustainable feature vectors. For reference, Tables 4 and 5 lists top five places of highest event count and of largest magnitudes, respectively.

Table 4. Top five places of highest seismic event count.

Place	Event count
Honshu, Japan	12293
Fiji islands region	8887
Kuril islands	7584
Vanuatu islands	6750
Tonga islands	6064

The Flinn-Engdahl scheme defines 50 geo-spatial regions and lists succinctly a total of 757 unique locations across. On the other hand, our captured event

Table 5. Top five places of highest magnitude, showing for each statistical summarization of scale distribution.

Place	Min	Max	Mean	SD
Northern Sumatra	4	9.1	4.60	0.45
Honshu, Japan	4	9	4.61	0.44
Bio-Bio, Chile	4	8.8	4.62	0.45
Southern Sumatra	4	8.5	4.77	0.47
Southern Peru	4	8.4	4.66	0.49

recordings exposed dozens of affiliated place names that are not registered in the standard seismic sites. Secondly, and particularly in recent years, name descriptions appear extremely verbose and embed excessive orientation information and absolute distance in kilometers from the set location. This disparity against the Flinn-Engdahl listings required both an additional pass of earthquake data cleanup to tidy up name strings, and to properly correlate the recorded data with the standard representation, our implementation extends the source directory of Flinn-Engdahl model by 32 sites. Thus bringing the total number of seismic places to 789, all distributed and abide by the originally specified, fifty geographical regions.

4.2 Experimental Preview

To set the stage for reporting experimental results that promote our geo-spatial neutral intuition for clustering seismic events, in this section we preview in contrast our rather rigorous approach to grouping shake events, while strictly abiding by geo-location proximity rules. In our seismic dataset, every shake event record incorporates a three-dimensional, absolute geo-location entity formulated by the tuple $\tau = \{latitude, longitude, depth\}$. This set of coordinates describes the epicenter of the shake volume of eruption that is defined by the geometrical bounding-box section of the GeoJSON object. Notably, the components of τ are however non-homogeneous, with latitude and longitude measured in angles, and depth stated in kilometers (Table 1). For the purpose of conducting explicit geo-spatial bound clustering, we have extracted verbatim the geo-location elements from each event instance, and constructed a collection of disjoint coordinate sets, τ_i, that are unlabeled and entirely detached to bear no dependence on any seismic place name. On this large opaque list of 326,267 physical 3D positions, we then ran the k-means clustering algorithm, along with deploying the more impartial, cosine similarity distance measure, since the epicenter coordinate system is non-Euclidean, and hence L_1-norm and L_2-norm metrics are unsuitable.

Figure 6 provides visualization of k-means clustering applied to our anonymous collection of shake epicenter positions, and showing planar projections of each of the possible unordered coordinate pairs from the {latitude, longitude, depth} set. As expected, group regions are fairly distinct with succinct borders

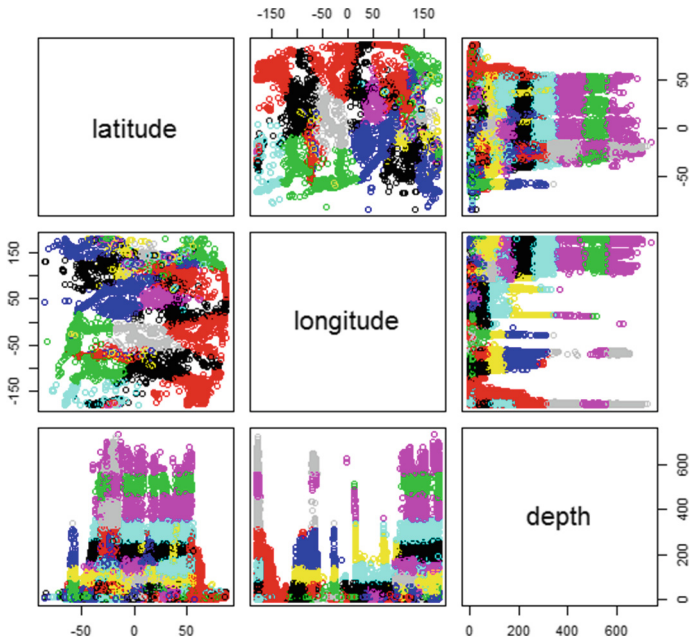


Fig. 6. *k*-means clustering on disjoint epicenter locations. Showing planar projections of each of the unordered coordinate pairs from the {latitude, longitude, depth} set.

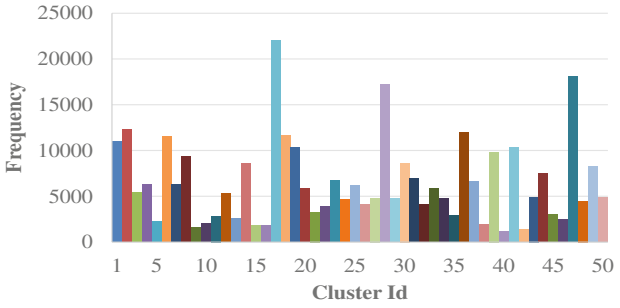


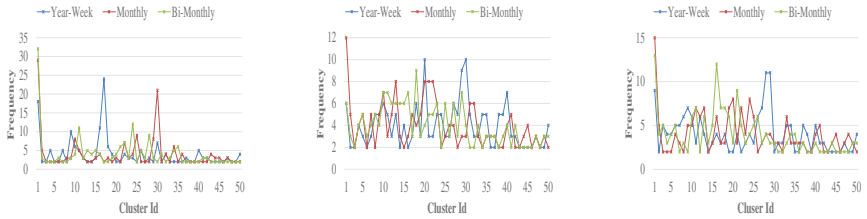
Fig. 7. *k*-means clustering on disjoint epicenter locations. Showing membership distribution of coordinate sets across 50 logical groups.

and are rendered for the most part with a solid color, signifying little to no occurrences of outliers. Epicenter member distribution across the fifty logical clusters has a mean of 6,525 locations per group, and only ten bins exceed a member count of 10,000 coordinate sets (Fig. 7). Overall, our results for generating seismic buckets merely based on geo-spatial proximity constraints, concur with our assertion for being of limited scale and lacking broader universal context.

4.3 Experimental Results

Our cluster analysis process is completely anonymous and assumes no prior knowledge of earthquake event locations. It solely relies on automatic feature extraction from recorded data, and incorporates statistical methods that facilitate the search of unsolicited seismic patterns, to discover global relations of earthquake occurrences that are not necessarily bound to geo-spatial proximity. In our software, both the number of seismic places to select from our earthquake data and the number of generated clusters are system level, user settable parameters. For our reported experiments we use consistently the recorded data of the top 200 geographical sites that underwent each at least 300 seismic events, and further split the locations into 50 logical clusters. First, we derive the implicit groups of geo-spatial proximity nature, by simply looking up an experimental place name from the extended and manually constructed Flinn-Engdahl directory structure, incorporated into our software. This distribution of places into already defined regional clusters, serves a useful comparative reference in analyzing our generic statistical approach, composed of a mixture model, whose components directly entail the partitions of places, and a Markov chain that follows hierarchical clustering and a dynamic tree cutting procedure. To match the Flinn-Engdahl scheme for analysis, both the number of components and the number of subtrees are set in our software to fifty, respectively.

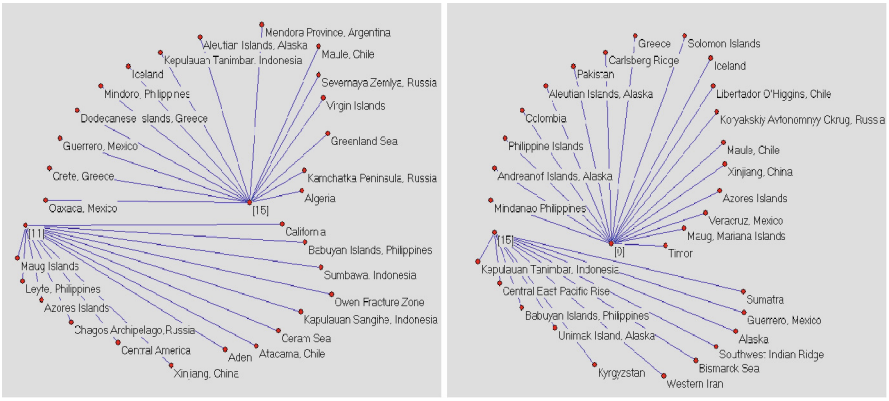
Fig. 8. Seismic place distribution across 50 clusters: bottom row is the geographically based Flinn-Engdahl model, middle row depicts the mixture model results, and the top row shows the Markov chain outcome. A lighter grey color implies a higher membership place count and a black stripe identifies an empty group.



(a) Single linkage. (b) Complete linkage. (c) Average linkage.

Fig. 9. Place membership distribution in clusters for the Markov chain model, shown for the agglomerative single, complete, and average linkage methods, and parameterized for each by year-week, monthly and bi-monthly resampling modes.

Unless otherwise noted, for the Markov chain model we apply the year-week resampling mode to the time series, feature vector, and report agglomerative clustering results using the single linkage method. Figure 8 shows cluster distribution of seismic places in Flinn-Engdahl, mixture model, and Markov chain formulations. A grey stripe represents a group, and the lighter its intensity the higher the membership place count. In excluding empty clusters, identified by black stripes, populated location collections total 40, 38, and 50 for our three clustering paradigms, respectively. Our analysis experiments exploit 200 places spread across 50 relational arrays, or four seismic sites per cluster, on average, and Table 6 provides complementary statistical summarization of cluster place membership, emphasizing single member groups of no association in the 1-Place column. The Markov chain approach stands out in both finding seismic relations for at least a pair of places, and moreover, it employs the full set of fifty groups.



(a) Mixture model. (b) Markov chain.

Fig. 10. Graphically visualized two cluster networks for each the mixture model and Markov chain clustering frameworks.

Figure 9 further depicts an interpretation of place membership distribution for the Markov chain model, reviewing the single, complete and average linking methods, each parametrized by the seismic signal resampling modes, including year-week, monthly and bi-monthly intervals. Results pertaining to the Ward linkage method are intentionally precluded to avoid reporting any bias towards even place divisions. Partition allocations for the complete and average link functions appear on a fairly equal scale and show a convincing behavior resemblance. Whereas on first inspection, the simple similarity method differs strikingly from the rest and has three peaks that stand apart for groups of about 20 to 30 place members. However, in barring the outliers and rescaling the remaining member counts, a clear indication of equivalence ensues. As the place term frequency in a cluster varies orthogonally to any of modifying the linkage method or the resampling mode, notably is the strong inclusive correlation often observed across classes generated out-of-order by different linkage methods. Figure 10(b) shows for network node 0 a super group created by the single linkage, and Fig. 11 presents for that node the subgroups produced by both the complete and average similarity methods that are fully contained in the aforementioned super class.

Table 6. Statistical measures of distributed, cluster place membership for the Flinn-Engdahl, mixture model, and Markov chain clustering paradigms. The 1-Place column identifies single member groups of no relations.

Model	Min	Max	Med	SD	1-Place
Flinn-Engdahl	0	24	2	4.48	8
Mixture Model	0	16	3	4.06	4
Markov Chain	2	24	3	4.07	0

Apart from the intuition of seismic similarity resulting from geo-spatial proximity, as prescribed in the Flinn-Engdahl model, we are interested in patterns that relates places by their closeness in feature space, for both the scale frequency and time series representations that feed into our mixture model and Markov chain, respectively. Figure 10 shows graphically the networks of two cluster nodes and their contextual place descendants, for each of our clustering frameworks. Emanating from a statistical process of grouping unlabeled earthquake bound locations, an immediate observation of the cluster content identifies seismic behavior similarities in geographical places that are both close and far apart physically. For example, cluster id 15 (Fig. 10(a)), originated in the composition of scale distribution features, incorporates European, African, North and South American, and Asian countries including Iceland, Greece, Algeria, Alaska, Mexico, and the Philippines. Similarly, cluster id 11 (Fig. 10(a)) has California from North America, Aden in Africa, Eastern Europe Russia, and Asian Indonesia. Corollary, assemblies of time series features (Fig. 10(b)) configure sites of different continents and show little resemblance to the Flinn-Engdahl

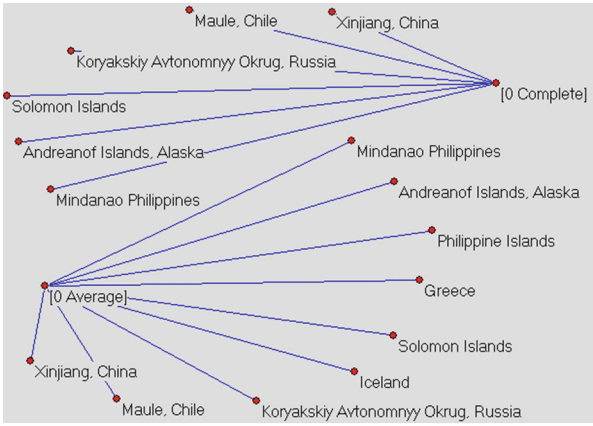


Fig. 11. Graphically visualized cluster networks produced by using the complete and average linkage methods.

geo-spatial regional scheme. The discovery of unsolicited seismic patterns promotes less dependence on a constraint physical partition profile and encourages more flexible and autonomous ecological relations, founded on objective and informative macroseismic effects [12].

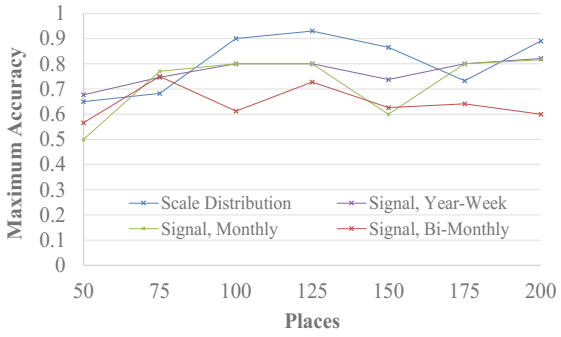


Fig. 12. Classification maximum accuracy as a function of ascending number of places, split into five clusters and parametrized by seismic feature type.

Our software is flexible to let the user set both the number of quake affected places and the number of clusters to generate, in each of the mixture model and Markov chain formulations. Constructing groups composed of a larger site count, enable us to perform classification and measure system level accuracy. We use the holdout method for cross validation and deploy the k-nearest neighbor (KNN) [5] baseline model that computes a Euclidean-squared distance between a randomly selected, test seismic vector against the remaining training feature

vectors, in either a distribution scale or a signal based representation. We then apply a normalized majority rule to ten nearest samples to a test feature vector, and derive a seismic score. This score is further accumulated and averaged for each cluster, and the matching cluster corresponds to the highest average scoring, cluster id. Figure 12 shows classification maximum accuracy for a five-cluster partition as a function of a non descending place count, and parametrized by a seismic feature type. Scale distribution features depict a slightly higher accuracy compared to the seismic signal form, mostly ascribed to sparseness of the latter due to samples of no seismic action, and evidently, a coarser mode of time series resampling results in a mild decline of accuracy rate.

To the best of our knowledge and based on literature published to date, we are unaware of seismic analysis systems with similar goals to even-handedly contrast our results against. The seismo-surfer [26], developed for seismic data management and mining employs the k-means algorithm for clustering. By specifying n geo-places and k clusters, k-means time complexity is $\mathcal{O}(kn)$ for each iteration, however the number of iterations to converge can be very large. Conversely, in our experiments both the EM and BW algorithms ran efficiently well under 100 iterations to convergence, along with setting the likelihood delta threshold to $1e^{-10}$. On the other hand, the computational complexity of a bottom-up hierarchical clustering is $\mathcal{O}(n^2 \log n)$, but this is often traded off with the construction process to terminate early once the desired number of clusters is reached. Another key architectural difference is the localized spatio-temporal nature of queries into the seismo-surfer database, as our design tends to seek more broader shake relations that span the universe mostly unconditionally.

5 Conclusions

We have demonstrated the apparent potential in deploying information retrieval and unsupervised machine learning methods, to accomplish the discovery of geo-spatial free similarity of earthquake bound places. By disregarding any prior location knowledge from presumed unlabeled seismic data, our proposed system is generic and scalable and relies entirely on objective closeness metrics in feature space that removes dependency on a more constraining regional scheme. For each of our distribution and signal typed feature renditions, both cluster analysis and classification results affirm the presence of seismic relational patterns that cross continent boundaries, suggesting similarity of impartial macroseismic effects.

Our system is exclusively inspired by a mixture and a state based clustering models that reason parametric distribution functions and a time series of intensity scale observations, respectively. As such, our method is more adaptable and far-reaching to challenge a wider gamut of multi geographical contexts. In contrasting our collaborative technique against a more compulsory concept that abides by a constraining geo-spatial physical closeness, and learns centroids from a collection of disjoint three-dimensional coordinate sets with no linkage to a specific geo-location, we have confirmed the limited extensibility of the latter and its lacking of an inclusive macrocosm perspective.

The data we acquired comprised of a large number of hundreds of thousands earthquake events, recorded in an extended period of time of four decades, and affected a few thousands sites around the world. However, only a few hundreds of places, each bearing at least several hundreds of shake occurrences, are statistically reasoned and pertinent to our probabilistic system approach. Advancing the growth of the seismic training set is imperative to our work and directly affects classification robustness. Yet using geographical locations that endured under one hundred seismic events is a suboptimal choice for our system, giving rise to highly sparse feature vectors. Alternatively, we contend that by coalescing locations of a small event count into a macro seismic site, based on geo-spatial proximity considerations, our training collection size is likely to increase further and proportionally let us gain a more stable classification process.

A direct progression of our work is to assume no foregoing knowledge of the number of seismic clusters to generate, and discover both the model fitting and the selection dimension directly from the incomplete seismic training set, using a combination of Akaike and Bayesian information criteria. We look forward to further incorporate the three dimensional geometrical data provided in a GeoJSON object, and possibly detect seismic similarity along either a latitude or a longitude extent perspective. Lastly, the flexibility of our software allows us to pursue a higher level, inter-cluster network study to better understand second order set of seismic relations.

Acknowledgements. We would like to thank the anonymous reviewers for their insightful and helpful feedback on our work.

References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: International Symposium on Information Theory, Budapest, pp. 267-281 (1973)
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press Series/Addison Wesley, Essex (1999)
3. Baum, L.E.: An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In: Symposium on Inequalities, Los Angeles, pp. 1-8 (1972)
4. Baum, L.E., Petrie, T.: Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Stat. **37**(6), 1554–1563 (1966)
5. Cormen, T.H., Leiserson, C.H., Rivest, R.L., Stein, C.: Introduction to Algorithms. MIT Press/McGraw-Hill Book Company, Cambridge (1990)
6. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. **39**(1), 1–38 (1977)
7. Duda, R.O., Hart, P.E., Stork, D.G.: Unsupervised learning and clustering. In: Pattern Classification, pp. 517–601. Wiley, New York (2001)
8. Flinn-Engdahl Seismic and Geographic Regionalization Scheme (2000). http://earthquake.usgs.gov/learn/topics/flinn_engdahl.php
9. Fraley, C., Raftery, A.E.: Bayesian regularization for normal mixture estimation and model-based clustering. J. Class. **24**(2), 155–181 (2007)

10. Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis and density estimation. J. Am. Stat. Assoc. **97**(458), 611–631 (2002)
11. GeoJSON Format for Encoding Geographic Data Structures (2007). http://geojson.org/
12. Hough, S.E.: Earthquake intensity distribution: a new view. Bull. Earthq. Eng. **12**(1), 135–155 (2014)
13. Johnson, S.C.: Hierarchical clustering schemes. J. Psychom. **32**(3), 241–254 (1967)
14. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
15. Langfelder, P., Zhang, B., Horvath, S.: Defining clusters from a hierarchical cluster tree: the dynamic tree cut library for R. J. Bioinform. **24**(5), 719–720 (2007)
16. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
17. Manning, C.D., Schutze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (2000)
18. Mclachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
19. Mclachlan, G.J., Basford, K.E.: Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York (1988)
20. Ngatchou-Wandji, J., Bulla, J.: On choosing a mixture model for clustering. J. Data Sci. **11**(1), 157–179 (2013)
21. R Project for Statistical Computing (1997). http://www.r-project.org/
22. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**(2), 257–286 (1989)
23. Rajaraman, R., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press, New York (2011)
24. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975)
25. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
26. Theodoridis, Y.: SEISMO-SURFER: a prototype for collecting, querying, and mining seismic data. In: Manolopoulos, Y., Evripidou, S., Kakas, A.C. (eds.) PCI 2001. LNCS, vol. 2563, pp. 159–171. Springer, Heidelberg (2003)
27. United States Geological Survey (USGS) (2004). http://earthquake.usgs.gov/earthquakes/feed/v1.0/
28. Ward, J.H.: Hierarchical grouping to optimize an objective function. Am. Stat. Assoc. **58**(301), 236–244 (1963)
29. Young, J.B., Presgrave, B.W., Aichele, H., Wiens, D.A., Flinn, E.A.: The Flinn-Engdahl regionalization scheme: the 1995 revision. Phys. Earth Planet. Inter. **96**(4), 223–297 (1995)

Web-Based Geoinformation System for Exploring Geomagnetic Field, Its Variations and Anomalies

Andrei V. Vorobev(✉) and Gulnara R. Shakirova

Ufa State Aviation Technical University,
Ufa, Russian Federation
gims@geomagnet.ru

Abstract. In the modern World, specialists in many scientific and applied spheres consider parameters of geomagnetic field, its variations and anomalies as one of the key factors, which can influence on systems and objects of various origins. The estimation of the influence requires an effective approach to analyze the principles of distribution of geomagnetic field parameters on the Earth's surface, its subsoil and in circumterrestrial space. The approach causes a complicated problem to be solved, which is concerned with modeling and visualization of parameters of geomagnetic field, its variations and anomalies. The most effective and obvious solution to this problem is supposed to be a geoinformation system, because of the geodata-centric character of the problem itself. In this paper the authors suggest the solution, which is based on modern geoinformation and web technologies and provides the mechanisms to calculate, analyze and visualize parameters of geomagnetic field and its variations.

Keywords: Geoinformation systems · Geomagnetic field · Geomagnetic variations · Geomagnetic anomalies · 2D/3D-visualization

1 Introduction

Geomagnetic field is well-known as the magnetic force field that surrounds the Earth. To simplify the description of the geomagnetic field it can be defined as a large bar magnet placed at the center of the Earth, with its south end oriented toward the north magnetic pole. The main goal of the Earth's magnetic field is to deflect most of the solar wind. Otherwise the charged particles of the solar wind would strip away the ozone layer of the Earth and all the systems and objects at the planet would be subjected to the influence of harmful ultraviolet radiation [1].

Geomagnetic field is shaped somewhat like a comet, which tail stretches for hundreds of thousands of kilometers in the direction opposite to the Sun and accumulates magnetic energy [2]. This area is called the magnetosphere and its shape is formed in response to the dynamic pressure of the solar wind (Fig. 1).

At any point on the Earth's surface the geomagnetic field is a combination of several magnetic fields generated by various sources. These fields are superimposed on and interact with each other [3].

C. Grueau and J.G. Rocha (Eds.): GISTAM 2015, CCIS 582, pp. 22–35, 2016.
DOI: 10.1007/978-3-319-29589-3_2

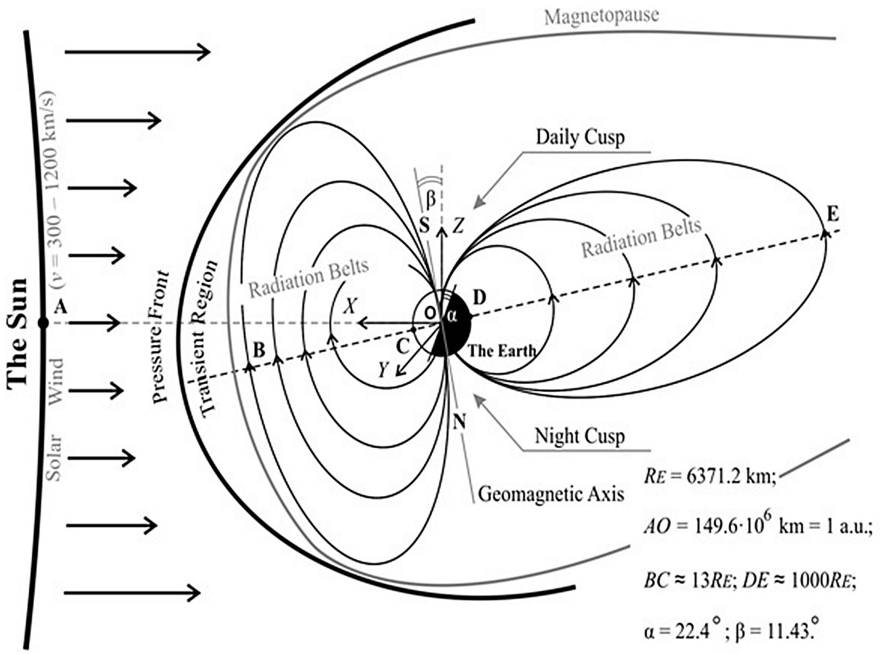


Fig. 1. A structure of geomagnetic field.

However, more than 90 % of the geomagnetic field measured is generated in the Earth's outer core (e.g. internal sources of the planet). This part of the geomagnetic field is known as the main field. It varies slowly in time and can be described by mathematical models [3].

Geomagnetic field is not stable and changes with periodicities from about 0.3 s to hundreds of years. These changes are often referred to as geomagnetic variations, which can arise from both sources external to the Earth or internal to the Earth. For example, sometimes the energy in magnetosphere tail is released in explosions. They heat up plasma, and cause powerful electric currents. At this moment the magnetosphere is filled to capacity with hot plasma, while its electric currents embrace the entire near-Earth space. These phenomena are referred to as magnetic storms [3].

One more important type of geomagnetic field change is caused by the object movement in anisotropic magnetic field. The parameters of geomagnetic field (which influence at the object) can vary significantly.

It is well known that some components of geomagnetic variations or their combinations can influence on biological, technical, geological and other objects and systems in common and on human in particular [4, 5]. There are a lot of known cases, when geomagnetic variations and anomalies affect the performance of equipment, upset radio communications, blackout radars, disrupt radio navigation systems, and endanger living organisms. Also there are some studies that describe correlations between human behavior and geomagnetic activity that might support some causal relations. Many

animals use the magnetic field like we use GPS to navigate. So any changes of geomagnetic field can cause some troubles for them.

Today the problem of monitoring of geomagnetic field and its variations parameters is partially solved by various ways. First, there are traditional magnetic measurements, which have routinely been carried out on the ground and over the oceans since 16th century. Next, a global view of geomagnetic field and how it changes is provided several by several magnetometry satellites in near-Earth orbit. And finally, geomagnetic field can be studied by a number of magnetic observatories (The magnetic observatory is a scientific organization, which is specialized on parametric and astronomical observations of the Earth's magnetosphere). The registered information about magnetic field and ionosphere state is regularly sent to the International centers in Russia, USA, Denmark and Japan. In these centers the information is registered, analyzed and partially available to the broader audience with some delay. Today there are about 100 geomagnetic observatories, and one third of them are in Europe [2].

Today monitoring, registration, visualization, analysis, forecast and identification of geomagnetic variations is a relevant sophisticated fundamental scientific problem with strong applied character.

All the data measured and collected about geomagnetic field is distributed in various sources and archives. There is still no integrated information space to get any data about geomagnetic field at any point of the Earth's surface at any moment of time. The obvious way to solve the problem is to implement innovative information technologies there. In particular the most expectations are about using geoinformation systems to solve the problem. In this paper the authors suggest an approach to study, monitoring, analyze and visualize geomagnetic field, its variations and anomalies, which is based on modern Web and geoinformation technologies.

2 Information Technologies to Explore Geomagnetic Field

In spite of the wide variety of specialized geoinformation systems there are no advanced hard- and software, which provide a calculation, geospatial connection, visualization and analysis of geomagnetic field, its variations and anomalies.

All the known information resources can be divided into two main groups. First one is represented by a complex of online databases from a number of magnetic observatories. These databases can be used by the following way. A user chooses an observatory and the period and gets a file with result to upload or to online plot it on screen. To use this data it is necessary to do something special to make a layer between these uploaded files located somewhere and user applications. However this approach is quite enough just to look through data.

Another solution is represented by geomagnetic calculators. To obtain necessary data a user enters coordinates of the place and gets parameters of normal magnetic field in the point. The great disadvantage here is that a user has to know the exact coordinates (no place name or something else). Sometimes it is a problem.

An example of modern geomagnetic calculator is the service, which is provided by NOAA (National Oceanic and Atmospheric Administration) and available at http://www.ngdc.noaa.gov/geomag-web. However the calculation results are out of limits of

permissible errors. It is takes no much time to ensure about incorrect work of some tools, absence of visualization tools and multilingual support, bad geolocation and non-informative interface.

It is important to mention, that due to low-efficiency, limited functionality and incorrect work of the known solutions the topicality, scientific and applied interest to such a solution development continuously increases. The necessary thing here is a set of mathematical models, which can describe the main field as a number of equations, based on the spatiotemporal parameters of the point of the Earth's surface.

3 Mathematical Modeling of Geomagnetic Field and Its Variations

The full vector of the Earth's magnetic field intensity in any geographical point with spatiotemporal coordinates is defined as follows [6]:

$$\boldsymbol{B}_{ge} = \boldsymbol{B}_1 + \boldsymbol{B}_2 + \boldsymbol{B}_3,$$

where $\boldsymbol{B}_1$ is an intensity vector of geomagnetic field of intraterrestrial sources; $\boldsymbol{B}_2$ is a regular component of intensity vector of geomagnetic field of magnetosphere currents, which is calculated in solar-magnetosphere coordinate system; $\boldsymbol{B}_3$ is a geomagnetic field intensity vector component with technogenic origin.

Normal (undisturbed) geomagnetic field is supposed as a value of $\boldsymbol{B}_1$ vector with excluding a component, which is caused by rocks magnetic properties (including magnetic anomalies). So this component is excluded as a geomagnetic variation:

$$\boldsymbol{B}_0 = \boldsymbol{B}_1 - \Delta\boldsymbol{B}'_1,$$

where $\boldsymbol{B}_0$ is undisturbed geomagnetic field intensity in the point with spatiotemporal coordinates; $\Delta\boldsymbol{B}'_1$ is component of intraterrestrial sources geomagnetic field, which represents magnetic properties of the rocks.

Solving the problem of $\boldsymbol{B}_0$ parameters analytical estimation, it is helpful to represent the main field model by spherical harmonic series, depending on geographical coordinates.

The scalar potential of intraterrestrial sources geomagnetic field induction U [nT·km] in the point with spherical coordinates r, θ, λ is defined by the expression (1).

$$U = R_E \times \\ \times \sum_{n=1}^{N} \sum_{m=0}^{n} \left(g_n^m \cos(m\lambda) + h_n^m \sin(m\lambda)\right) \left(\frac{R_E}{r}\right)^{n+1} P_n^m \cos(\theta), \tag{1}$$

where r is a distance from the Earth's center to observation point (geocentric distance), [km]; λ is a longitude from Greenwich meridian, [degrees]; θ is a polar angle (collatitude, $\theta = (\pi/2) - \varphi'$, [degrees], where φ' is a latitude in spherical coordinates, [degrees]); R_E is an average radius of the Earth, R_E = 6371.03, [km]; $g_m^n(t)$, $h_m^n(t)$ are

spherical harmonic coefficients, [nT], which depend on time; p_m^n are Schmidt normalized associated Legendre functions of degree n and order m.

In specialized literature the expression (1) is widely known as a Gaussian and generally recognized as an international standard for undisturbed state of geomagnetic field.

The amount of performed spherical harmonic analysis is significant. However a problem of spherical harmonic optimal length of still acute.

Thus, the analyses with great amount of elements prove Gauss's hypothesis about convergence of spherical harmonic, which represents a geomagnetic potential. As usual in spherical harmonic analyses the harmonics are limited by 8–10 elements. But for sufficiently homogeneous and highly accurate data (for example, as like as in satellite imaging) the harmonics series can be extended up to 12 and 13 harmonics. Coefficients of harmonics with higher orders by their values are compared with or less than error of coefficients definition.

Due to the main field temporal variations the coefficients of harmonic series (spherical harmonic coefficients) are periodically (once in 5 years) recalculated with the new experimental data.

The main field changes for one year (or secular variation) are also represented by spherical harmonics series, which are available at http://www.ngdc.noaa.gov/IAGA/vmod/igrf11coeffs.txt.

Schmidt normalized associated Legendre functions p_m^n from expression (1) in general can be defined as an orthogonal polynomial, which is represented as follows (2).

$$\begin{aligned} P_n^m(\cos(\theta)) &= 1 \cdot 3 \cdot 5 \ldots \sqrt{\frac{\varepsilon_m}{(n+m)!(n-m)!}} \times \\ &\times \sin^m \theta [\cos^{n-m} \theta - \frac{(n-m)(n-m-1)}{2(2n-1)} \cos^{n-m-2} + \\ &+ \frac{(n-m)(n-m-1)(n-m-2)(n-m-3)}{2 \cdot 4(2n-1)(2n-3)} \cos^{n-m-4} \theta - \ldots], \end{aligned} \tag{2}$$

where ε_m is a normalization factor ($\varepsilon_m = 2$ for $m \geq 1$ and $\varepsilon_m = 1$ for $m = 0$); n is a degree of spherical harmonics; m is an order of spherical harmonics.

4 Geomagnetic Pseudostorm Effect

Here it is supposed to enter the term geomagnetic pseudostorm, which is intended to represent real geomagnetic field influence on the object in conditions of its non-zero speed and undisturbed geomagnetic field anisotropy [6]. Let us describe some main parameters of geomagnetic pseudostorm effect [6].

Range of geomagnetic pseudostorm is a difference between maximal and minimal values of geomagnetic field induction in area of the object, which is moving in anisotropic magnetic field during the time period or at the distance:

$$\boldsymbol{B}_{\text{GMPS}} = \boldsymbol{B}_{0\ \max} - \boldsymbol{B}_{0\ \min},$$

where $\boldsymbol{B}_{0\ \max}$ and $\boldsymbol{B}_{0\ \min}$ are maximal and minimal values of geomagnetic field induction, [nT] in area of the object, which is moving in anisotropic magnetic field.

Frequency spectrum of geomagnetic pseudostorm is a function of distribution of geomagnetic pseudostorm amplitude spectrum in frequency area for continuous and discrete variants, which is defined by the following expressions:

$$B^*(f) = \int_{-\infty}^{+\infty} B_0(t) e i^{-2\pi f t} dt \text{ or } B^*(f) = \frac{1}{M} \sum_{t=0}^{M-1} B_0(t) e^{\frac{-2\pi f t}{M}},$$

where B^* is a frequency spectrum of geomagnetic pseudostorm; $\boldsymbol{B_0}$ is a value of geomagnetic field induction in the point with spatiotemporal coordinates; M is a quantity of registered values with constant discretization step by time.

Constant component of geomagnetic pseudostorm is a vector of harmonics superposition vertical shift, which represent frequency spectrum of geomagnetic pseudostorm:

$$B_{//} = \frac{1}{M} \sum_{t=0}^{M-1} B_0(t),$$

where M is a quantity of registered values with the discretization step.

Intensity of geomagnetic pseudostorm is a physical quantity, which is numerically equal to the speed of undisturbed geomagnetic field characteristic change in time relatively to the frame of reference, which is connected to the moving object and depends on the object speed:

$$I_{\text{GMPS}} = \frac{\partial B_0}{\partial t},$$

where $\boldsymbol{I}_{\text{GMPS}}$ is GMPS intensity, [nT/s];

Potentiality of geomagnetic pseudostorm (geomagnetic induction gradient) is a vector, which is oriented in three-dimensional space and points to the direction of the fastest increase of undisturbed geomagnetic field induction absolute value. The vector by its absolute value is equal to the increase speed of $\boldsymbol{B_0}$ in the geographical direction, [nT/rad; nT/rad; nT/km] and depends on the object position.

$$G_B = \nabla B_0(\theta, \lambda, r) = \text{grad}\ B_0(\theta, \lambda, r) = \left(\frac{\partial B_0}{\partial \theta}, \frac{\partial B_0}{\partial \lambda}, \frac{\partial B_0}{\partial r} \right),$$

where $\boldsymbol{B_0}$ is an induction (intensity) of geomagnetic field in the point with spatiotemporal coordinates:

$$B_0^2(\theta, \lambda, r)[\mathrm{nT}] = \left[\frac{1}{r}\frac{\partial U}{\partial \theta}\cos(\varphi - \phi') - \frac{\partial U}{\partial r}\sin(\varphi - \phi')\right]^2 +$$
$$+ \left[-\frac{1}{r \cdot \sin\theta}\frac{\partial U}{\partial \theta}\right]^2 + \left[-\frac{\partial U}{\partial r}\cos(\varphi - \phi') - \frac{1}{r}\frac{\partial U}{\partial \theta}\sin(\varphi - \phi')\right]^2.$$

So the analysis of geomagnetic field induction gradient distribution allows defining an area of possible maximal intensity of geomagnetic pseudostorm of in the geographical region. So, the parameter G_B must be taken into account in developing aerospace navigation maps and flight paths [6].

Next to study the geomagnetic pseudostorm effect there is an example of the flight route AA-973 of «American Airlines» from New York (JFK) to Rio de Janeiro (RIO). The flight path is represented as an array of spatial coordinates, which describe the airplane position, taken during flight in equal time intervals. The array allows calculating geomagnetic field parameters for each set of spatial coordinates.

The results of amplitude and frequency analysis of flight data and parameters of geomagnetic field are represented on Fig. 2. Here are some special points (Fig. 2(a)):

- t_1–t_2 – takeoff time;
- t_2–t_4 – flight on cruise speed at the altitude of 11033 m;
- t_4–t_5 – landing time;
- t_3 – passing the equator.

5 G-Service to Explore Geomagnetic Field

To take all advantages and recover all disadvantages of existing projects for exploring geomagnetic field the authors have suggested and developed special web-service, which is based on a set of mathematical models, defined at previous sections.

The service is called GIMS Calculator (or G-Service) and provides a set of instrumental tools to calculate parameters of normal geomagnetic field in the user-defined point [6]. The service is available at URL: http://www.geomagnet.ru.

On the logical and programming levels the web application is a set of complex procedures, which provide the realization of geospatial data visualization and analysis of geomagnetic field parameters in the point with spatiotemporal coordinates.

With the lower abstraction the web application is a special class of web page, which is developed according to three-tier client-server architecture. The visible and adapted to the user (after rendering) markup of the page is realized via W3C-standardized markup language HTML (specifically, its XML-type modification XHTML) [6].

The page design is performed in traditional table style: each region of the page is the table cell of various levels. However, the table-like layout of the page is not the only design solution here: there is also block-type layout via HTML-elements div, which logically structure the page by its semantics. For example, one block HTML-element stands for the region with map [7], another – for the region with spatiotemporal coordinates of the point, and the last – for the region with geomagnetic field parameters calculation results.

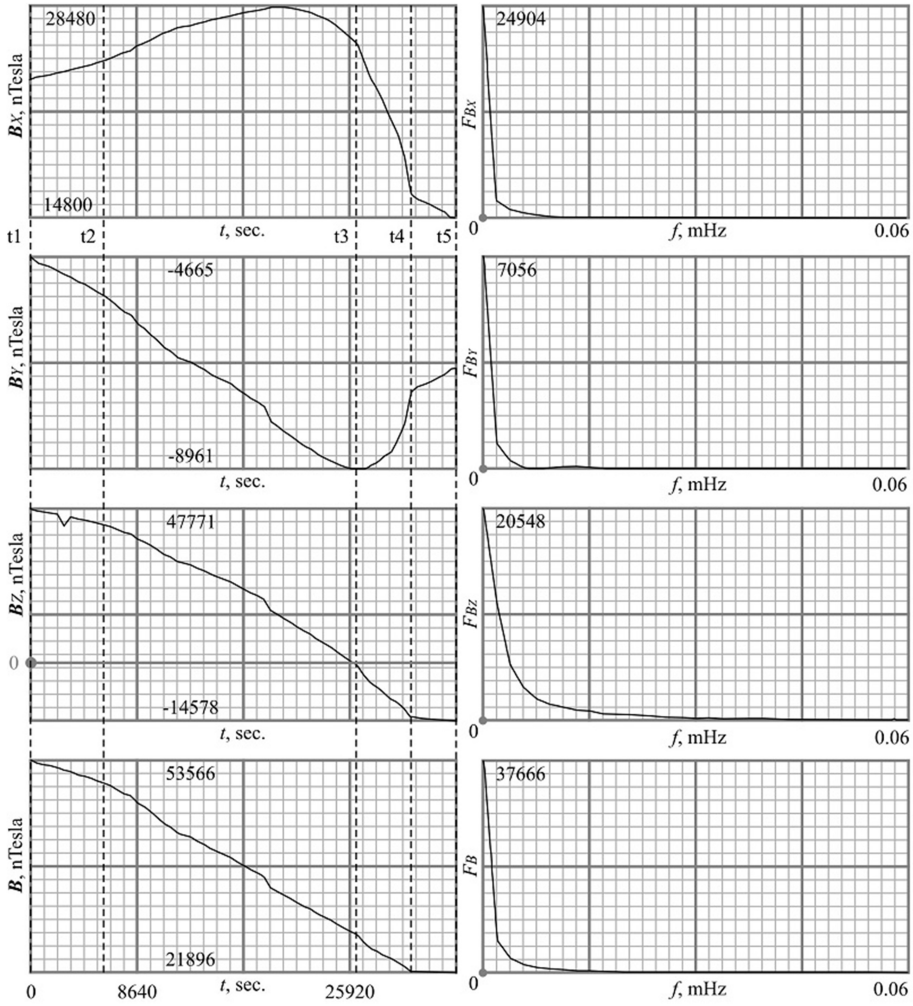


Fig. 2. Experimental data analysis results.

The great feature of the service is its integration with a map: parameters of the point and the map are placed on the one screen. User can enter coordinates into input fields or pick up a point directly on map.

User window of service "GIMS-calculator" is logically divided into two functional areas (panels). Left panel (Fig. 3) is supposed for loading and rendering the Earth's surface maps fragments in either scheme or photo. Right panel is supposed for representing the input parameters/initial conditions, calculation results and "GIMS-calculator" functionality control. (The initial conditions are defined as spatiotemporal default coordinates: 54.7249° N, 55.9425° E, 0.172 km amsl) [6].

Calculation of parameters of geomagnetic field is based on spatiotemporal coordinates of the point on the Earth's surface. To define a point a user can apply one of the following approaches:

- Geolocation. It is the simplest way to define the current geographical position of the user. Geospatial coordinates of user location are defined by IP address of device, which is used for accessing the Internet. This possibility allows the user to get the point without its searching on the map or filling the appropriate input fields. This feature increases its efficiency and speed of the research.
- Pick up a point on the map. By the map rendered in service window a user can choose any point he is interested in. A user can move through the map (using keyboard or mouse) and click at the point. All necessary spatiotemporal parameters of the point are calculated automatically and immediately displayed on screen.
- Enter coordinates. It is a good way to calculate parameters of geomagnetic field and see the point on map with high accuracy (up to a few meters). A user enters the coordinates into the input fields, provided by the service. After that the service displays the point on the map and the parameters of normal geomagnetic field there.
- Enter address. It is a function often referred to as geocoding. To find the point a user enters address of the place he is interested in. The address can be represented at any level (city name, address with city and street names, full address including building number, etc.). Also the address is represented in special information window, which is connected with the point on the map. The solution of this task is based on special Google API geocoding service (exactly by its realization – reverse geocoding). Geocoding is a transformation from address full form "postal code, country, city, street, building number" or short form into the coordinates set "north latitude, east longitude". Reverse geocoding provides address description by the coordinates.

A geocoding service is also an asynchronous function: it supposes status check, callback function and array of results. Geocoding realization in API is based on the object google.maps.Geocoder. The resulting array contains multiple representation of address (because of reversal geocoding) with various detailing levels. First element of the array (with index equal to 0) is the most detailed address, which is used to be represented in information window.

User-defined spatiotemporal coordinates put the center of the map visible fragment relatively to the geographical point, which is defined by them. The point is outlined by the marker with geolocation results.

An important feature of "GIMS-calculator" is data representation in one of the two formats: DD (decimal degrees) and DMS (degrees – minutes – seconds). Depending on the chosen format a user gets the appropriate input mask. Also the application supposes the automatic transformation of coordinate systems via checking the appropriate radio button.

The altitude of the point is also calculated automatically (but it can be corrected by a user) on the basis of latitude and longitude input values. An altitude value is represented in International System of Units or Imperial and US customary measurement systems.

As in previous case, the direct and indirect transformations are available. To calculate an altitude (or elevation) of the point the system uses special Google API service

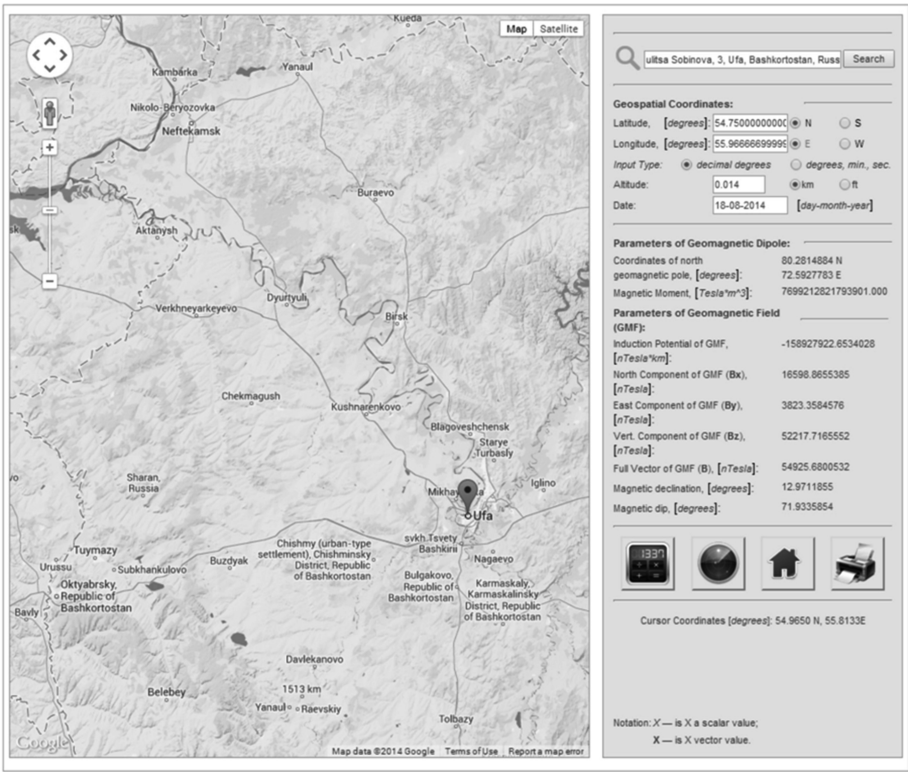


Fig. 3. "GIMS-calculator" user interface.

"ElevationService". It is an object with an asynchronous interaction: after sending a request to the server a user (or a web page) does not wait for its response and keeps performing all existing operations (or start new ones) in background mode. The system also takes into account the metric system and represents the results in the appropriate form.

After picking the point a user can calculate parameters of geomagnetic field there. By default the "GIMS-calculator" represents parameters of geomagnetic field in international system of units. So this data can be analyzed without any preliminary calculations.

The main parameters of geomagnetic field to be calculated are the following:

- north component of geomagnetic field induction vector;
- vertical component of geomagnetic field induction vector;
- magnetic declination and dip;
- scalar potential of geomagnetic field induction vector.

It is important to mention, that the "GIMS-calculator" calculates the parameters of geomagnetic field and its variations depending on the date, which user enters as an input parameter. By default this parameter is set to current date, but a user can change it to any other.

To calculate the parameters of geomagnetic field and its variations the system actualizes matrices of spherical harmonic coefficients. It calculates the current epoch and the difference between the result value and the moment for geomagnetic field analysis. And the resulting difference is a normalization multiplier for the harmonic coefficients matrices.

To visualize the results of calculation the system provides a set of contours, where each contour is a curve along which the parameter of geomagnetic field has a constant value. A user chooses a parameter to be visualized and the system renders on the map a set of contours, which represent a distribution of magnetic field on the Earth's surface (Fig. 4).

It is important to mention that "GIMS Calculator" is based on the authors-suggested algorithm of calculation of parameters of geomagnetic field. The algorithm supposes special scheme to minimize the possible error because of the rounding. The error of the calculation is less than 2 %.

Also to extend the functionality of the developed Web GIS "GIMS-calculator" there were programmed an option of generating electronic report about the research results with file or printer form and a possibility of three-dimensional modeling.

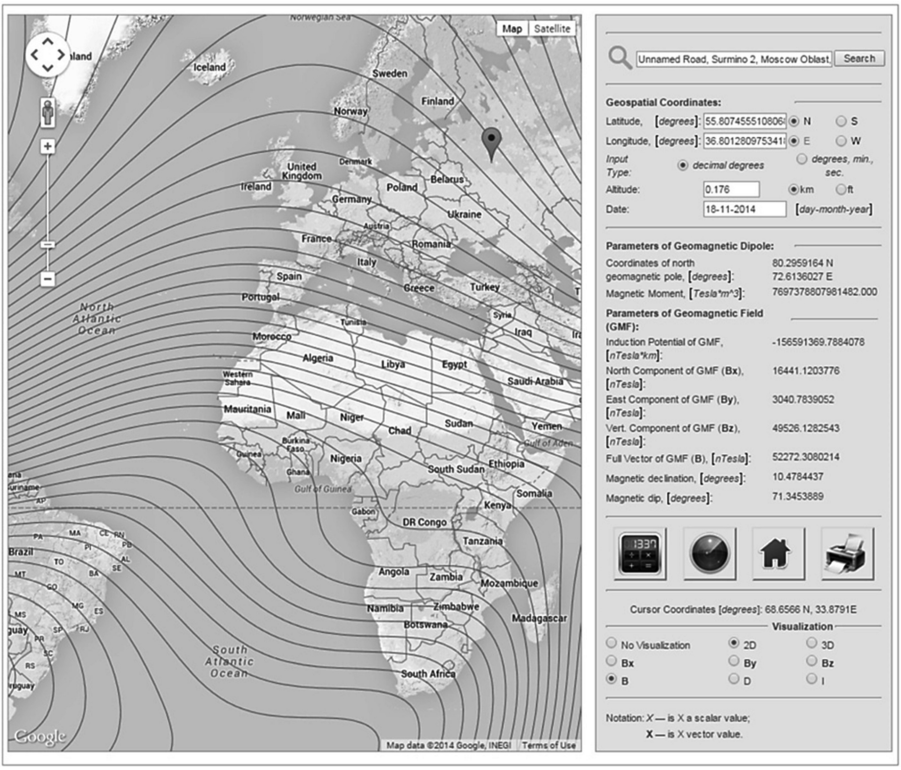


Fig. 4. Contours in "GIMS-calculator".

6 Three-Dimensional Visualization of Geomagnetic Field

An effective solution for complicated problem of modeling and visualization of geomagnetic field and its variations parameters is a key to understand the principles of geomagnetic field parameters distribution on the Earth's surface, its subsoil and in circumterrestrial space.

Three-dimensional representation of calculation of parameters of GMF and its variations is one of the main aspects in solution of visualization problems of both geospatial data and parameters of geomagnetic field, its variations and anomalies. It is obvious, that in this case geoinformation system provides much more information than any other system or technology. And it is even more important due to the dynamic properties and multilevel scale ability.

Today a problem of geographical and attributive spatial data three-dimensional visualization is usually solved via web applications of special type, which are known as virtual globes. It is important to mention, that virtual globes technology is based on the Earth's surface representation as a sphere with applied graphical layers.

Virtual globes integration with applications is provided by special API, which is a set of programming functions for creation, visualization and manipulation of three-dimensional spatial data.

Programming interface is used by an application as a set of local or remote functions. These functions can be used with the special possibilities of interpreter, which is already used or additionally loaded on user computer.

Usually three-dimensional geomodeling represents information on two levels: geographical and attributive. A geographical description of geospatial data supposes three-dimensional visualization of the Earth's surface with variable zoom and detail parameters. And the attributive component of the data is represented as a set of numerical values, which correspond to values of GMF parameters for spatial coordinates with an appropriate step.

To perform three-dimensional geomodeling the "GIMS-calculator" applies a technology of virtual globes or geobrowsers (Google Earth API). It is important to mention, that geobrowsers technology is based on the Earth's surface representation as a sphere with applied graphical layers (Fig. 5).

Three-dimensional geomodeling is provided by "GIMS-calculator" just similar to two-dimensional representation. A user chooses a parameter to be visualized, and the system renders data on the globe.

The "GIMS-calculator" defines layers to be represented on the globe in KML format. Keyhole Markup Language is one of the most popular formats of geodata representation, which is supposed as XML-oriented description of three-dimensional model of the Earth surface and the objects on it. The description on KML is a set of geographical and attributive data.

Each KML layer in necessity can be overlaid on any other layer (for example, data about seismic, volcanic activities, medical statistics, geological maps, etc.). It provides an effective tool for complex analysis of various parameters, correlation and principles definition. Numerical value of the parameter (physical value), which is distributed along the one contour, is available via picking the appropriate line with mouse cursor.

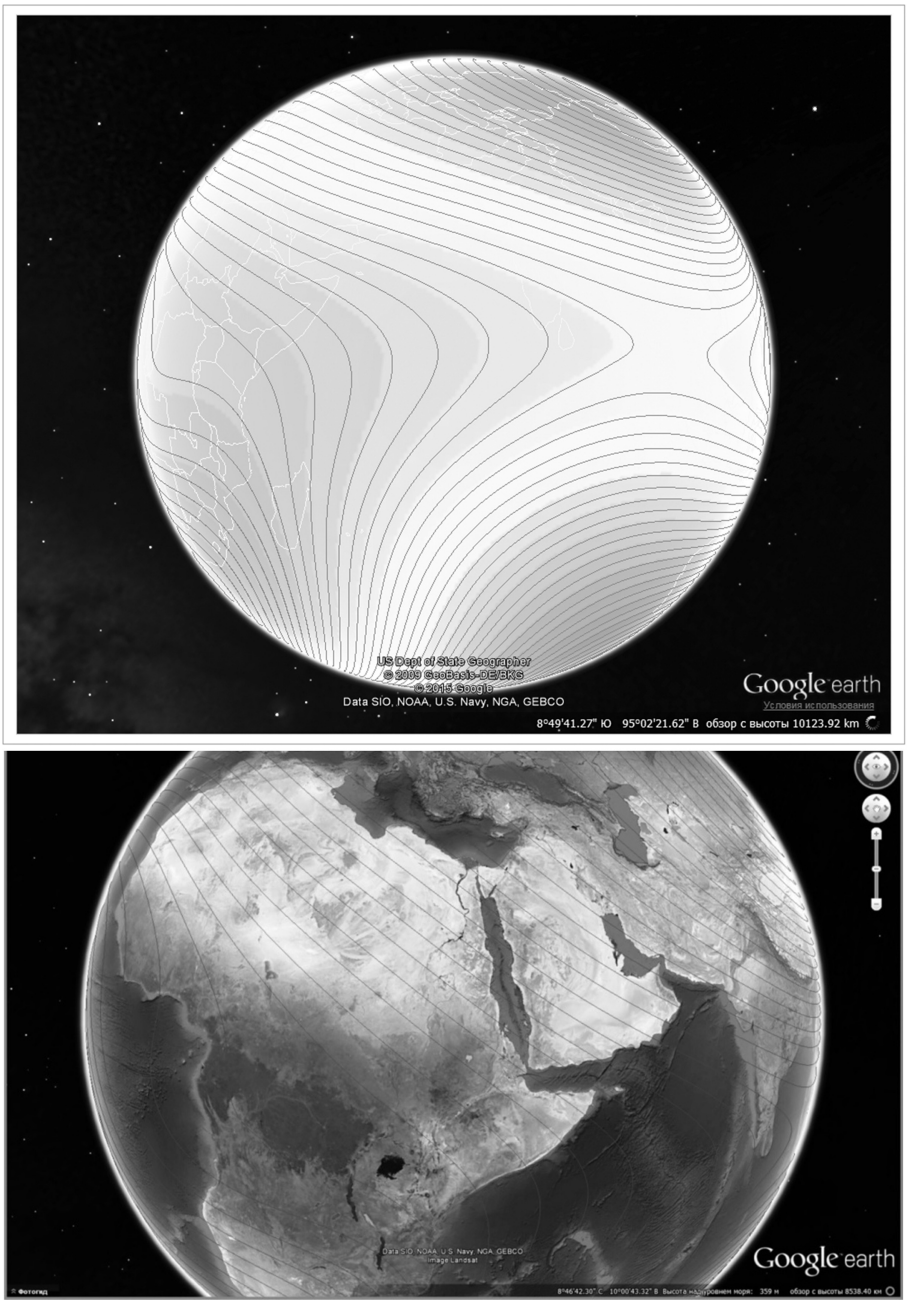


Fig. 5. Three-dimensional geomodeling in "GIMS-calculator".

Various methods of visual data representation (color outlining, gradient, etc.) significantly increase the model informativeness. Active layers (country borders, cities, rivers, etc.) managing keeps the key points of the model with decreasing the probability of possible error.

7 Conclusion

- Geomagnetic field is a complex structured natural matter with ambiguous field characteristics, which is distributed in the Earth (and near-Earth) space and interacts with both astronomical objects and objects/processes on the Earth's surface, subsoil and in near-Earth space. Geomagnetic field and its variation can influence on systems and objects of various origins. The estimation of the influence requires an effective approach to analyze the principles of distribution of geomagnetic field parameters on the Earth's surface, its subsoil and in circumterrestrial space. The approach causes a complicated problem to be solved, which is concerned with modeling and visualization of geomagnetic field and its variations parameters. The most effective and obvious solution to this problem is supposed to be a geoinformation system.
- Web-based geoinformation system "GIMS-Calculator" provides the complex calculation, analysis and 2D/3D-visualization of geomagnetic field and its variations parameters. Geomagnetic field and its variations models, which are represented and described by "GIMS-Calculator", meet the requirements of specialists in various areas. They effectively provide formatting and structuring the data about the Earth magnetosphere parameters and their further analysis.

Acknowledgements. The reported study was supported by RFBR, research projects No. 14-07-00260-a, 14-07-31344-mol-a, 15-17-20002-d_s, 15-07-02731_a, and the grant of President of Russian Federation for the young scientists support MK-5340.2015.9.

References

1. Campbell, W.H.: Introduction to Geomagnetic Fields, 2nd edn. Cambridge University Press, Cambridge (2003)
2. Mandea, M., Korte, M.: Geomagnetic observations and models. IAGA Special Sopron Series, vol. 5. Springer, Heidelberg (2011)
3. Merrill, R.T., McElhinny, M.W., McFadden, P.L.: The Magnetic Field of the Earth. Academic Press, Cambridge (1996)
4. Chizhevskii, A.L.: Earth Echo of Sun Storms. Mysl, Moscow (1976). (in Russian)
5. Vernadsky, V.I.: The Biosphere and the Noosphere. Iris Press, Moscow (2004). (in Russian)
6. Vorobev, A.V., Shakirova, G.R.: Pseudostorm effect: computer modelling, calculation and experiment analyzes. In: Proceedings of the 14th SGEM GeoConference on Informatics, Geoinformatics and Remote Sensing, vol. 1, pp. 745–751 (2014)
7. Haklay, M., Singleton, A., Parker, C.: Web mapping 2.0: the neogeography of the geo web. Geogr. Compass **2**, 2011–2039 (2008)

Identifying Local Deforestation Patterns Using Geographically Weighted Regression Models

Jean-François Mas(✉) and Gabriela Cuevas

Centro de Investigaciones en Geografía Ambiental,
Universidad Nacional Autónoma de México,
Antigua carretera a Pátzcuaro 8701,
Col. Ex Hacienda de San José de la Huerta,
58190 Morelia, Michoacán, Mexico
{jfmas,gcuevas}@ciga.unam.mx
http://www.ciga.unam.mx

Abstract. This study aimed at identifying drivers and patterns of deforestation in Mexico by applying Geographically Weighted Regression (GWR) models to cartographic and statistical data. We constucted a nation-wide multidate GIS database incorporating digital data about deforestation from the Global Forest Change database (2000–2013); along with ancillary data (topography, road network, settlements and population disribution, socio-economical indices and government policies). We computed the rate of deforestation during the period 2008–2011 at the municipal level. Local linear models were fitted using the rate of deforestation as dependent variable. In comparison with the global model, the use of GWR increased the goodness-of-fit (adjusted R2) from 0.20 (global model) to 0.63. The mapping of GWR models' parameters and its significance, anables us to highlight the spatial variation of the relationship between the rate of deforestation and its drivers. Factors identified as having a major impact on deforestation were related to topography, accessibility, cattle ranching and marginalization. Results indicate that the effect of these drivers varies over space, and that the same driver can even exhibit opposite effects depending on the region.

Keywords: Deforestation · Drivers · Geographically weighted regression · Mexico

1 Introduction

Mexico, with a total area of about two million square kilometres, is a megadiverse country, but it presents high rates of deforestation [9]. Various studies have attempted to assess land use/cover change (LUCC) over the last decades [21] but there have been few attempts to assess the main causes of deforestation at national level [5,10,24]. Given the complexity of the Mexican territory, the processes of change and its factors are expected to be different depending on

C. Grueau and J.G. Rocha (Eds.): GISTAM 2015, CCIS 582, pp. 36–49, 2016.
DOI: 10.1007/978-3-319-29589-3_3

the region. Geographically Weighted Regression (GWR) has been applied in exploring spatial data in the social, health and environmental sciences. The goal of this study is to evaluate the spatial patterns of deforestation with respect to drivers reported to influence LUCC using GWR models.

2 Materials and Methods

2.1 Material

In order to elaborate the geospatial database, the following data were used:

- Maps of tree cover, forest loss and forest gain from the Global Forest Change 2000–2013 database at 30 m resolution [14]. The map of tree cover in 2000 estimates canopy closure for all vegetation taller than 5 m in height and is encoded as a percentage per cell. The map of forest cover loss (2000–2013) is a binary map (loss/no loss) which indicates deforestation, defined as a change from a forest to non-forest state. An additional map (loss year) indicates the year forest loss occurred and is encoded as either 0 (no loss) or else a value in the range 1–13. The map of forest cover gain 2000–2012 indicates a non-forest to forest change within the study period and is encoded as either 1 (gain) or 0 (no gain). There is no map of forest gain year because forest gain was not allocated annually.
- Maps of ancillary data (digital elevation model, slope, roads maps, human settlements, climate, soils, municipal boundaries) [15].
- Socio-economic data from the National Institute of Geography, Statistics and Informatics (INEGI for its Spanish acronym) organized by municipality (Population census for 2005 and 2010) [16,17].
- Index of marginalization 2010 by municipality from the National Council of Population [8].
- Government policies (rural and cattle-rearing subsidies, and protected areas) [7,28].

GIS operations were carried out with Q-GIS [25]. Statistical analysis and graphs were created using R [26]. Geographically weighted regressions were carried out using the packages spgwr Bivand and Yu, [4] and GWmodel [13,19] in R.

2.2 Deforestation Rate Computing and Database Elaboration

In this study, forest area and forest change were estimated at municipality level (2456 municipalities) using the Global Forest Change 2000–2013 database [14]. A map of 2000 forest area was obtained thresholding the tree cover map at 10 %. For each municipality, the 2007 forest area was estimated updating 2000 forest area taking into account forest gain and loss during 2000–2006. As forest gain was allocated over the entire period 2000–2012, we computed a gain area proportional to the 2008–2011 period duration with respect to the 2000–2012 period. Deforestation rate was computed as the average annual proportion of 2007 forest

deforested during 2008–2011 for each municipality. In order to determine which ancillary variables are most likely to be indirect drivers of deforestation, we calculated, for each municipality, various indices describing resources accessibility, population, economic activities, and governmental policies:

- Road density (km of road per km^2 taking into account dirt and paved roads),
- Population density in 2010 (people per km^2),
- Settlements density (number of settlements per km^2),
- Index of marginalization, which takes into account incomes, level of schooling and housing conditions [8],
- Cattle density (heads per km^2),
- Goat density (heads per km^2),
- Average slope (degrees),
- Average elevation,
- Amount of governmental subsidies for agriculture (PROCAMPO, thousand of Mexican pesos per km^2),
- Amount of governmental subsidies for cattle ranching (PROGAN, thousand of Mexican pesos per km^2),
- Proportion of municipality within protected areas.

As the dependant variable is the proportion of 2007 forest cleared during 2008–2011, the variance of this proportion is likely to decrease as it approaches 0 or 1, which is a problem because the regression analysis assumes the error terms to have constant variance. To remove this problem, the proportion was transformed using the Eq. 1 to produce a variance-stabilised rate of deforestation [11].

$$TP = \arcsin[\sqrt{p}] \tag{1}$$

where TP is the transformed rate of deforestation and p is the original proportion.

Explanatory variables were also transformed using logarithm, square, square-root and exponential transformation in order to improve linearity.

2.3 Statistical Analysis

GWR is a local spatial statistical technique for exploring spatial non-stationarity [11]. It supports locally modelling of spatial relationships by fitting regression models. Regression parameters are estimated using a weighting function based on distance in order to assign larger weights to closer locations. Different from the usual global regression, which produces a single regression equation by summarizing the overall relationships among the explanatory and dependent variables (for the whole Mexican territory in that case), GWR produces spatial data that express the spatial variation in the relationships among variables. Maps that present the spatial distribution of the regression coefficient estimates along with the level of significance (e.g. t-values) have an essential role in exploring and interpreting spatial non-stationarity. Fotheringham et al., [11] provide with a

full description of GWR, and Mennis [22] gives useful suggestions to map GWR results.

Collinearity amongst the explanatory variables of a regression model is a well known problem which can lead to a loss of precision in the coefficient estimates [3]. This issue can be more pronounced in local regression models than in global ones due to smaller samples used to calibrate local regression and because some localities may exhibit high collinearity levels when others do not due to spatial heterogeneity [19]. The first stage of the study was a global correlation analysis between explanatory variables using the Spearman coefficient in order to discard highly correlated variables. Local collinearity between explanatory variables was assessed computing diagnostics as local correlations between pairs of explanatory variables, local variance inflation factors (VIFs) and local variance decomposition proportions (VDPs) for each explanatory variable and local condition numbers (CN). According to Lu et al. [19], local collinearity problems likely occur in the local regression model when local correlation values are greater than 0.8 for a given explanatory variable pair, when VIFs and VDPs are respectively greater than 10 and 0.5 for a given explanatory variable and when CN values are greater than 30 for the entire set of explanatory variables. Local correlation between the dependant variable (deforestation rate) and each explanatory variable was also calculated and mapped in order to detect explanatory variables with a high explanatory power. Explanatory variables were selected or discarded in order to reduce collinearity, keeping the ones with higher correlation with the deforestation rate and, as possible, trying to conserve variables describing different aspects of deforestation causes (accessibility, topography, human activities, public policies...).

2.4 Geographically Weighted Regression (GWR)

Due to the uneven distribution and size of the municipalities, the weighting function used an adaptive kernel which selects a proportion of the observations (k-nearest neighbours) assigned to each municipality and calculates the weights using a Gaussian model. The optimal size of the bandwidth (in this case the proportion of observations) was found by minimising the Akaike Information Criterion (AIC) [1], which is a model fit diagnostic that takes into account the model parsimony (trade-off between model complexity and prediction accuracy). A map was elaborated for each explanatory variable showing the value of the regressions coefficients (color scaling of the symbol) and statistical significance (hatched mask layer).

3 Results

3.1 Deforestation and Drivers Assessment

As shown in Fig. 1, the rate of deforestation varies greatly over space. The coastal floodplains of the Gulf of Mexico and the southern part of the country exhibits

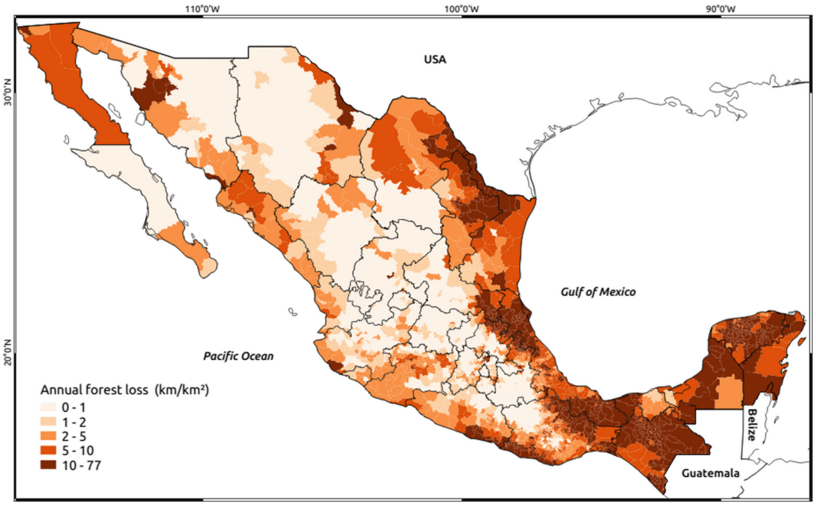


Fig. 1. Rate of deforestation during the period 2008–2011.

high rates of deforestation. It is worth mentioning that this rate of deforestation is a gross rate which does not include forest gain. It is also a relative rate, expressed as a proportion of previously existing forest that has been deforested during the period of study. While absolute rates of deforestation are expected among the municipalities with the largest areas of forests, high relative rates can occur when small area of forest are cleared in scarcely forested municipalities. Figures 2, 3 and 4 show the spatial distribution of some of the explanatory variables (road density, cattle density and marginalization).

3.2 Explanatory Variables Selection

The pairs of variables population and human settlements densities and, amount of cattle-ranching subsidies and cattle density present a global Spearman correlation of 0.71 and 0.99 respectively. For this reason, population density and cattle-ranching subsidies were removed from the set of explanatory variables. However, with the nine remaining variables there is still a problem of collinearity (Fig. 5). Therefore, more variables were discarded from the analysis. Global diagnostic give a poor information to carry out the process of selection. For example, slope, which has a low global correlation with deforestation (Table 1) exhibits high local values of correlation (Fig. 6). In most of the cases the relation is negative but in some cases, which correspond mainly to municipalities with large flooded areas or with a high proportion of anthropic cover, the relation is negative (more deforestation in steeper sloping areas). However, slope was dropped out the set of explanatory variables because it presented local collinearity. The analysis of local correlation between the rate of deforestation and the explanatory variables along with collinearity indices enable us to identify the

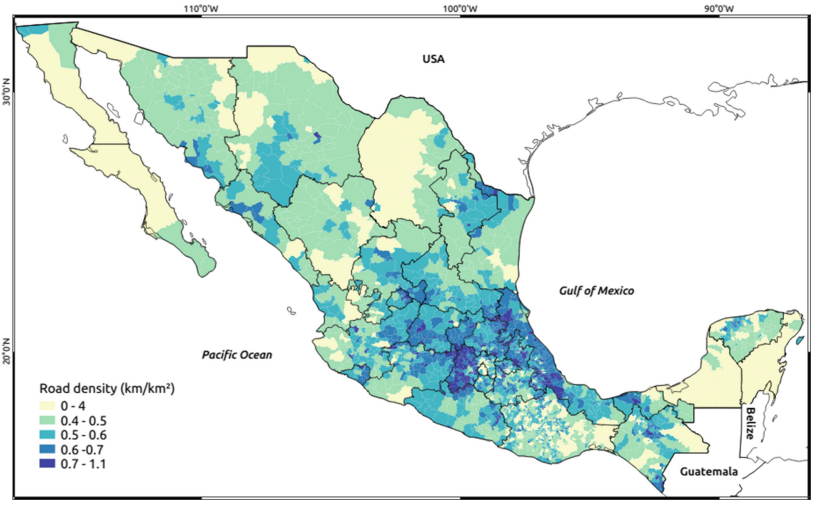


Fig. 2. Road density (km per km^2).

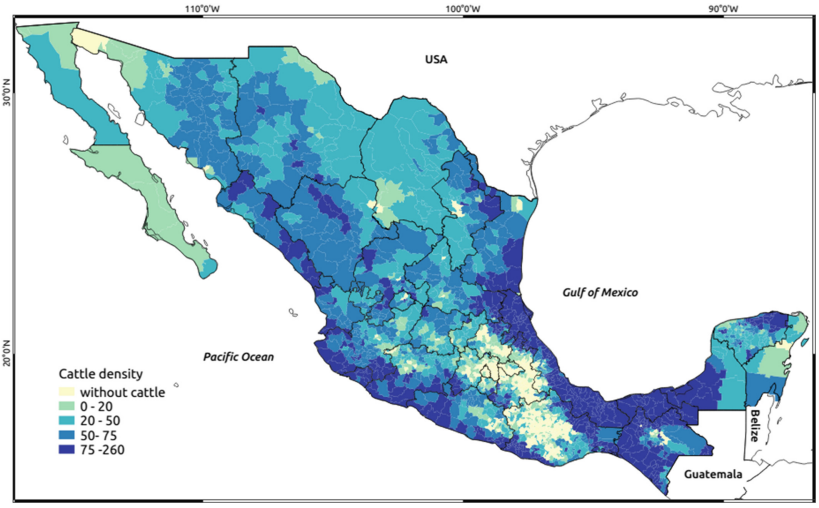


Fig. 3. Cattle density (heads per km^2).

variables with low explanatory power and high collinearity with other explanatory variables.

In order to decrease collinearity at tolerable levels, the number of explanatory variables was finally reduced to five: Index of marginalization, cattle density, elevation, road density, subsidies for agriculture and, protected areas. Figure 7 shows that the Condition Number is inferior to 30 in almost the entire territory.

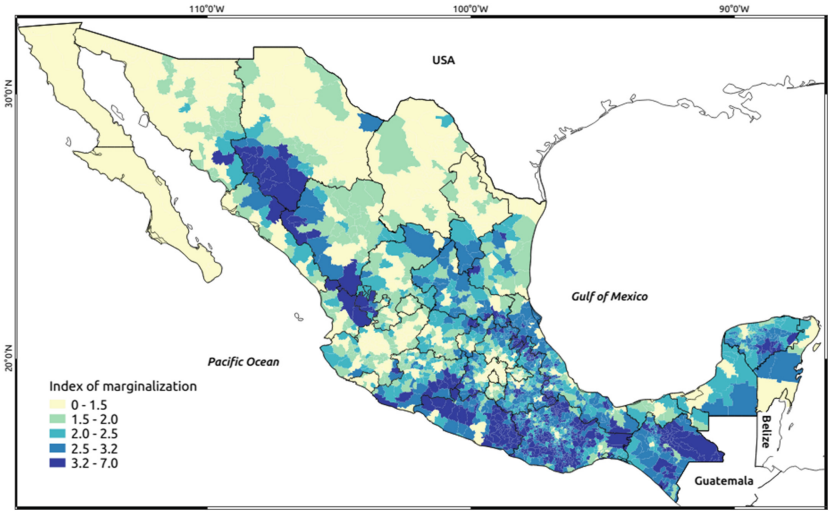


Fig. 4. Index of marginalization 2010 (CONAPO).

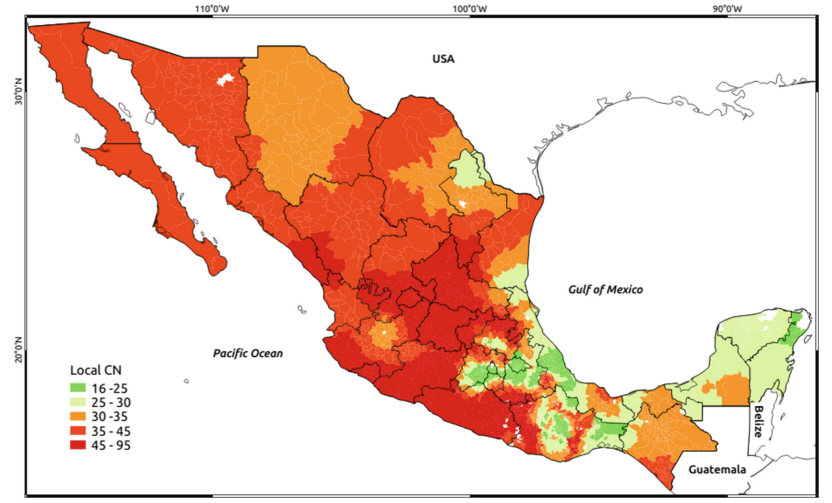


Fig. 5. Local condition number (CN) using the nine input explanatory variables. A value above 30 (red tones) indicates collinearity (Color figure online).

3.3 Geographically Weighted Regression (GWR)

For comparison purpose, a global model was fitted and obtained an adjusted-R^2 of 0.20 (Table 2). A GWR was fitted using the five selected explanatory variables and a weighting function based on a 7 % of the observations (167 neighbours). The use of GWR increased the strength in the relationship in terms of the goodness-of-fit (adjusted R^2) to 0.63.

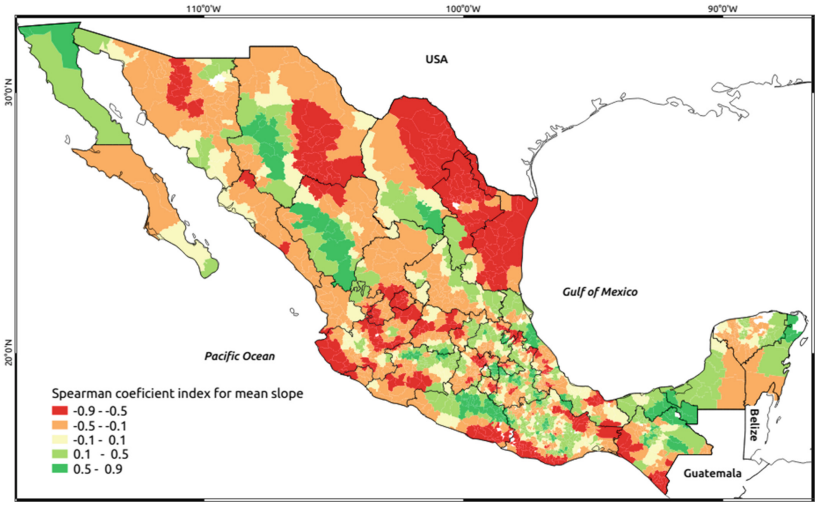


Fig. 6. Local Spearman correlation between the rate of deforestation and the slope. Red and green tones indicates negative and positive relationship respectively (Color figure online).

Table 1. Global coefficient of correlation (Spearman) between the rate of deforestation and explanatory variables.

Explanatory variable	Correlation
Road density	0.05
Population density	0.06
Settlements density	0.11
Marginalization index	0.25
Cattle ranching subsidies	0.36
Goat density	−0.16
Slope	−0.05
Elevation	−0.49
Subsidies for agriculture	0.11
Subsidies for cattle ranching	0.36
Protected areas	−0.06

Road density presents a positive relationship with deforestation in the north of the country and in the southern part, which is an expected behaviour as roads are often reported as a deforestation driver (Fig. 8). In the center of the country, which presents the higher road density values (Fig. 2), this relation is no significant or even negative, likely due to the fact that municipalities with the highest road densities are likely already almost totally deforested and therefore

Table 2. Global model summary.

	Estimate	Std. error	t value	Pr(>t)
(Intercept)	−0.0078	0.0074	−1.05	0.2924
Road density	0.0098	0.0114	0.86	0.3890
Subsidies for agriculture	0.0001	0.0000	3.40	0.0007
Protected areas	−0.0244	0.0067	−3.65	0.0003
Marginaization	0.0139	0.0016	8.78	0.0000
Cattle density	0.0007	0.0000	20.48	0.0000

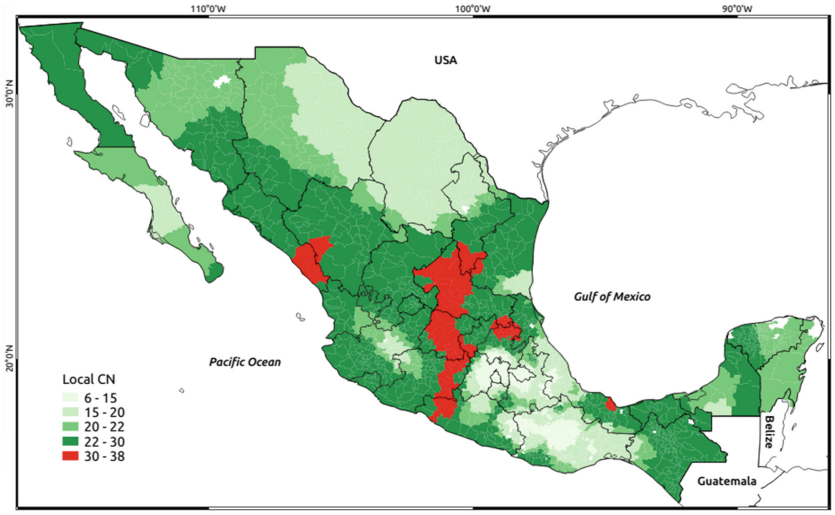


Fig. 7. Local condition number using the five inputs variables. Values above 30 (red), which indicates collinearity, appear only in a small part of the center of the country (Color figure online).

present low rates of deforestation during the period 2008–2011. Cattle density presents a positive relationship with deforestation in all the country (Fig. 9). The marginalization index presents a significant positive relationship with the rate of deforestation in the eastern part of Mexico (Fig. 10). Many studies have associated poverty and deforestation [27]. There are many regions with high level of marginalization where the relationship between marginalization and deforestation is not significant (and in some cases negative). Previous researches have reported that the most conserved natural areas in Mexico are often located in poor rural areas and/or community lands [2,6,10,12,18].

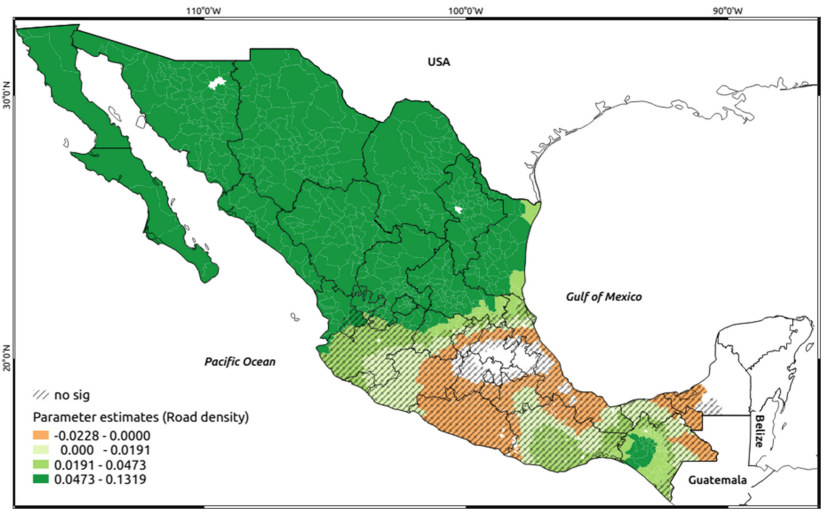


Fig. 8. Distribution of the local parameter estimates associated with road density.

4 Discussion

Some limitations of this study related with input data and with the way information is summarized at municipality level have been identified:

– Change data are based only on a drastic change of land cover (forest loss), they do not consider cover degradation. This factor has to be considered during the

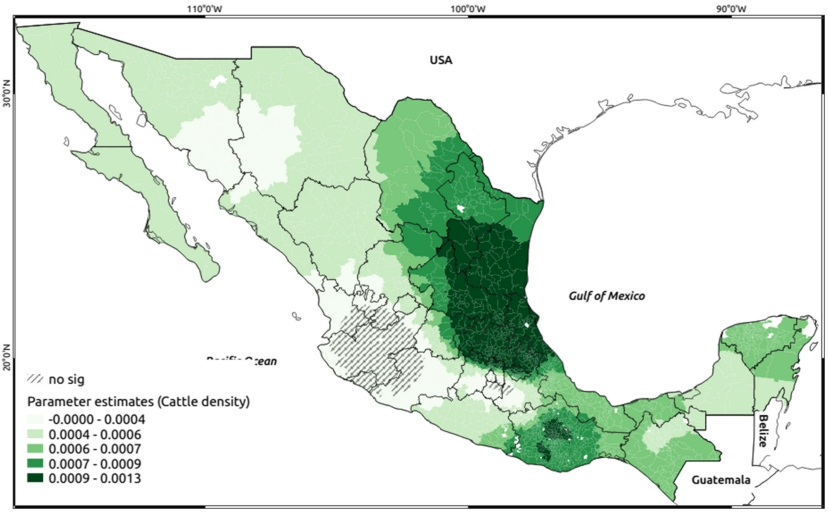


Fig. 9. Distribution of the local parameter estimates associated with cattle density.

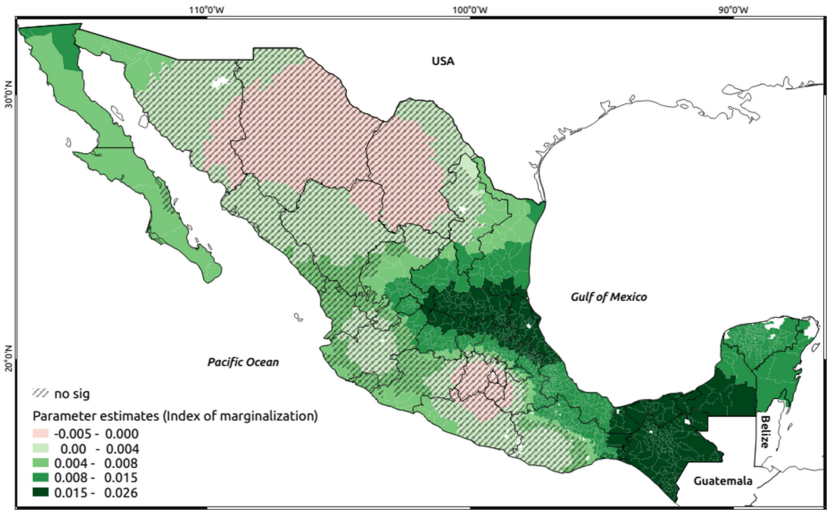


Fig. 10. Distribution of the local parameter estimates associated with the marginalization index.

results interpretation. In Mexico, the rate of deforestation has been decreasing during the last decades and most of the change processes are related with vegetation cover degradation rather than deforestation. Moreover, as the study used a rate of deforestation as dependant variable it does not take into account large extensions of scrublands in the north of the country.

– The rate of deforestation shows change from 2008 to 2011, but the drivers variables (population, marginalization, government subsidies) are from a particular date at different times of the period depending on data availability. The temporal issue cannot be totally addressed due to the lack of information. Additionally, in some cases, it could be interesting to calculate rates of change of these indices. For instance deforestation may be more related to the increase of population density than to population itself.
– Another limitation is the averaging of indices at municipality level which may end up with a figure that does not reflect the actual situation over much of the area. For instance, a municipality with both flat and steep slope areas will present the average value corresponding to moderate slope. Moreover, deforestation can occur in small regions which present very different features from the average figure. This effect, known as the modifiable areal unit problem (MAUP) [23] was evaluated in the case of municipal data for Mexico [20]. The evaluation concluded that, in most of the cases, MAUP did not make large difference to the results. However, counterintuitive relationship between deforestation and slope in some areas (Fig. 6) is maybe due to MAUP. A way to decrease this problem could be to calculate the indices taking into account only the forested area. For instance, average slope of municipality forest area

is used to explain deforestation instead of the slope average over the entire municipality.
– Finally, as depicted by the moderate R^2 of the model, the set of explanatory variables we used did not allow to explain the dependent variable in a satisfactory manner for the entire territory. More drivers have to be taken into account for future analysis.

Other limitations are related with the method used and the deforestation process itself: Deforestation is a complex process that depends on interacting environmental, social, economic and, cultural drivers. Some of them cannot be used into the model because they are unable to be mapped. Moreover, the GWR uses municipality information to explain deforestation but is unable to take into account shifting effects (deforestation in a given municipality is due to the actions from inhabitants from other municipalities) and effects at different scale (as the GWR use the same bandwidth for all the explanatory variables). It is worth noting that some drivers cannot act with very fuzzy spatial pattern or no pattern at all (e.g. global economy effect such as import/export of agriculture goods).

It is likely that the effect of a driver on a given region is related to the time such driver has been shaping the landscape and that different drivers have affect at different temporal and spatial scales, which makes the interpretation of the results difficult. Considering the rate of deforestation during different past periods of time will enable us to analyze the dynamic of deforestation in its temporal and spatial dimensions.

In this study, a special attention has been paid to evaluate and avoid collinearity removing the offending explanatory variables. However, the strategy of removing an explanatory variable is not ideal, particularly when only a local collinearity effect is present, because it limits the number of useful explanatory variables in the regression model due to the high correlation existing between spatial variables. An alternative strategy is to use models with a locally-compensated ridge term [13]. Other alternative GWR models are robust models to identify and reduce the effect of outliers and, mixed models which are able to manage some explanatory variables as constant over space (or stationary) and other as local (non-stationary). In future researches, alternative deforestation rates will be also computed, new explanatory variables such as land tenure will be integrated into the model and, a workshop will be organized to carry out deep interpretation of the results.

5 Conclusions

The results we obtained clearly show the advantages of a local approach such theGWR models over a global one, to evaluate drivers' effect on LUCC processes as deforestation over such a diverse and complex territory as Mexico. The GWR model enabled us to describe spatial relationships between drivers and deforestation. Local models gave a much better explanation of deforestation patterns than the average changes identified by global models such as global regression models.

GWR models can be useful in different types of LUCC study. The exploration of space can help account for differences between regions not captured by standard global measures and thus explain causes of LUCC in different areas. However, the interpretation of the maps of regression estimates is not an easy task and need to be supported by a profound knowledge of the area and of its history. In this regard, GWR offers an attractive visual tool to motivate discussion among specialists from different fields of knowledge (e.g. geographers, anthropologists, economists) and to communicate scientific results with policy-makers and general public. GWR can also be useful in policy design and assessment. Different governmental policies and interventions for deforestation reduction may be appropriate in different areas; depending on local conditions. Therefore the design of locally sensitive policies is likely to be more efficient than nation-wide policies. Alternatively, GWR can be used to evaluate the success of a given policy already in place by determining areas where the intervention was more successful and eventually why.

Acknowledgements. This research has been funded by the *Consejo Nacional de Ciencia y Tecnología* (CONACyT) and the *Secretaría de Educación Pública* (grant CONACYT-SEP CB-2012-01-178816) and CONAFOR project: *Construcción de las bases para la propuesta de un nivel nacional de referencia de las emisiones forestales y análisis de políticas públicas*. The authors would like to thank the four reviewers for their careful review of our manuscript and providing us with their comments and suggestion to improve the quality of the manuscript.

References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B., Csaki, F. (eds.) 2nd Symposium on Information Theory, pp. 267–281. Akademiai Kiado, Budapest (1973)
2. Alix-García, J., de Janvry, A., Sadoulet, E.: A tale of two communities: explaining deforestation in Mexico. World Dev. **33**(2), 219–235 (2005)
3. Belsley, D., Kuh, E., Welsch, R.: Regression Diagnostics: Identifying Inuential Data and Sources of Collinearity. Wiley, New York (1980)
4. Bivand, R., Yu, D.: Package spgwr, Geographically weighted regression. http://cran.open-source-solution.org/web/packages/spgwr/spgwr.pdf
5. Bonilla-Moheno, M., Redo, D.J., Mitchell Aide, T., Clark, M.L., Grau, H.R.: Vegetation change and land tenure in Mexico: a country-wide analysis. Land Use Policy **30**(1), 355–364 (2013)
6. Bray, D.B., Duran, E., Ramos, V.H., Mas, J.F., Velázquez, A., McNab, R.B., Barry, D., Radachowsky, J.: Tropical deforestation, community forests, and protected areas in the Maya Forest. Ecol. Soc. **13**(2), 56 (2008)
7. Bezaury Creel, J.E., Torres, J.F., Ochoa-Ochoa, L., Castro Campos, M., Moreno Díaz, N.G.: Bases de datos georeferenciadas de áreas naturales protegidas y otros espacios dedicados y destinados a la conservación y uso sustentable de la biodiversisad en México. The Nature Conservancy (2011). Database on CD. Mexico
8. CONAPO: Índices de marginaci por localidad. http://www.conapo.gob.mx/es/CONAPO/Indice_de_Marginacion_por_Localidad_2010

9. FAO: Global resources assessment. Forestry paper, 140 (2001)
10. Figueroa, F., Sánchez-Cordero, V., Meave, J.A., Trejo, I.: Socioeconomic context of land use and land cover change in Mexican biosphere reserves. Environ. Conserv. **36**(3), 180–191 (2009)
11. Fotheringham, S.A., Brunsdon, C., Charlton, M.: Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, Chichester (2002)
12. García-Barrios, L., Galván-Miyoshi, Y.M., Valdivieso Pérez, I.A., Masera, O.R., Bocco, G., Vandermeer, J.: Neotropical forest conservation, agricultural intensification and rural outmigration: the Mexican experience. BioScience **59**(10), 863–873 (2009)
13. Gollini, I., Lu, B., Charlton, M., Brunsdon, C., Harris, P.: GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models. J. Stat. Softw. **63**(17), 1–50 (2015). http://www.jstatsoft.org/v63/i17/
14. Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G.: High-resolution global maps of 21st-century forest cover change. Science **342**, 850–853 (2013)
15. INEGI: Carta topográfica escala 1: 250000. INEGI, México (2004)
16. INEGI: Conteo de población y vivienda 2005. Indicadores del censo de Población y vivienda. INEGI, México (2005)
17. INEGI: Censo de población y vivienda 2010. INEGI, México (2010)
18. Klooster, D.: Beyond deforestation: the social context of forest change in two indigenous communities in highland Mexico. J. Lat. Am. Geogr. **26**, 47–59 (2000)
19. Lu, B., Harris, P., Charlton, M., Brunsdon, C.: The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models. Geospatial Inf. Sci. **17**(2), 85–101 (2014). http://www.tandfonline.com//abs/10.1080/10095020.2014.917453
20. Mas, J.F., Pérez, V.A., Andablo, R.A., Castillo Santiago, M.A., Flamenco, S.A.: Assessing modifiable areal unit problem in the analysis of deforestation drivers using remote sensing and census data. Int. Arch. Photogrammetry Remote Sens. Spat. Inf. Sci. (ISPRS Archives) **XL-3/W3**, 77–80 (2015)
21. Mas, J.F., Velázquez, A., Díaz-Gallegos, J.R., Mayorga-Saucedo, R., Alcántara, C., Bocco, G., Castro, R., Fernández, T., Pérez-Vega, A.: Assessing land/use cover changes: a nationwide multidate spatial database for Mexico. Int. J. Appl. Earth Obs. Geoinformatics **5**, 249–261 (2004)
22. Mennis, J.L.: Mapping the results of geographically weighted regression. Cartographic J. **43**(2), 171–179 (2006)
23. Openshaw, S.: Ecological fallacies and the analysis of areal census data. Environ. plann. A **16**, 17–31 (1984)
24. Pineda-Jaimes, N.B., Bosque Sendra, J., Gómez Delgado, M., Franco Plata, R.: Exploring the driving forces behind deforestation in the state of Mexico (Mexico) using geographically weighted regression. Appl. Geogr. **30**, 576–591 (2010)
25. QGIS Development Team: QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org
26. Core, R., Team, R.: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna (2014). http://www.R-project.org/
27. Rudel, T.A., Horowitz, B.: Tropical Deforestation: Small Farmers and Local Clearing in the Ecuadorian Amazon. Columbia University Press, New York (2013)
28. SAGARPA: Listas de beneficiarios de PROCAMPO y PROGAN (2008–2011)

XQuery-Based Query Processing in Open Street Map

Jesús M. Almendros-Jiménez(✉) and Antonio Becerra-Terón

Information Systems Group, University of Almería, 04120 Almería, Spain
{jalmen,abecerra}@ual.es

Abstract. Volunteered geographic information (VGI) makes available a very large resource of geographic data. The exploitation of data coming from such resources requires an additional effort in the form of tools and effective processing techniques. One of the most established VGI is Open Street Map (OSM) offering data of urban and rural maps from the earth. In this paper we present a library for querying OSM with XQuery. This library is based on the well-known spatial operators defined by Clementini and Egenhofer, providing a repertoire of XQuery functions which encapsulate the search on the XML document representing a layer of OSM, and make the definition of queries on top of OSM layers easy. In essence, the library is equipped with a set of OSM Operators for OSM elements which, in combination with Higher Order facilities of XQuery, facilitates the composition of queries and the definition of keyword based search geo-localized queries. OSM data are indexed by an R-tree structure, in which OSM elements are enclosed by Minimum Bounding Rectangles (MBRs), in order to get shorter answer time.

1 Introduction

Volunteered Geographic Information (VGI) is a term introduced by Goodchild [13,14] to describe geographic information systems based on crowdsourcing, in which users collaborate to a collection of spatial data of urban and rural areas from the earth. *Open Street Map* (OSM) [5,26] is one of the most relevant VGI systems, with almost two millions of registered users. OSM data can be visualized from the OSM Web site[1], and many applications have been built for the handling of OSM maps[2]. OSM data can be represented with many formats and several tools are able to export OSM to XML, KML and SVG, among others.

The main tasks OSM tools carry out are edition, export, rendering, conversion, analysis, routing and navigation, but little attention has been paid on querying. Querying urban maps can be seen from many points of view. One of the most popular querying mechanism is the so-called *routing* or *navigation*: the

This work was funded by the EU ERDF and the Spanish Ministry of Economy and Competitiveness (MINECO) under Project TIN2013-44742-C4-4-R, and by the Andalusian Regional Government (Spain) under Project P10-TIC-6114.

[1] http://www.openstreetmap.org.

[2] http://wiki.openstreetmap.org/wiki/Software.

C. Grueau and J.G. Rocha (Eds.): GISTAM 2015, CCIS 582, pp. 50–68, 2016.
DOI: 10.1007/978-3-319-29589-3_4

most suitable route to go from one point to another of the city. In this case, the inputs of the query are two points (or streets) and the output is the sequence of instructions needed to reach the destination.

Nevertheless, querying an urban map can also be interesting for city sightseeing. In this case, places of interests around a given geo-localized point are the major goal. The inputs of the query are a point and a city area, close to the given point, and the output is a set of points. The tourist would also like to query streets close to a given street when looking for a hotel, querying parking areas, restaurants, high ways to go out, etc. In such queries, the input is a given point (or street) and the output would be a number of streets, parking areas, restaurants, high ways, etc.

Most tools are able to query OSM with very simple commands: searching by tag and relation names. This is the case of *JOSM*[3] and *Xapiviewer*[4]. The *OSM Extended API or XAPI*[5] is an extended API that offers search queries in OSM with a XPath flavoring. The *Overpass API* (*or OSM3S*)[6] is an extension to select certain parts of the OSM layer. Both XAPI and OSM3S act as a database over the web: the client sends a query to the API and gets back the dataset that corresponds to the query. *OSM3S* has a proper query language which can be encoded by an XML template. *OSM3S* offers more sophisticated queries than *XAPI*, but it is equipped with a rather limited query language.

XQuery [3,27] is a programming language proposed by the W3C as standard for handling XML documents. It is a functional language in which *for-let-orderby-where-return* (*FLOWR*) expressions are able to traverse XML documents. It can express Boolean conditions, and provides a format to output documents. XQuery has a sublanguage, called *XPath* [6], whose role is to address nodes on the XML tree. XPath is properly a query language equipped with Boolean conditions and many path-based operators. XQuery adds expressivity to XPath by providing mechanisms to join several XML documents.

In this paper, we present a library for querying OSM with XQuery. This library is based on the well-known spatial operators defined by Clementini [8] and Egenhofer [9], providing a repertoire of XQuery functions which encapsulate the search on the XML document representing a layer of OSM, and making the definition of queries on top of OSM layers easy. Basically, the library is equipped with a set of *OSM Operators*, for OSM elements which, in combination with *Higher Order* facilities of XQuery, makes the *Composition of Queries* and the definition of *Keyword based* search *Geo-Localized* queries easy. OSM data are indexed by an R-tree structure [16], where OSM elements are enclosed by *Minimum Bounding Rectangles* (*MBRs*) in order to get shorter answer time.

Our work focuses on the retrieval of information and querying from urban maps. Although navigation is a interesting type of query, we are more interested in querying the elements of a urban map in a certain area or layer and taking as

[3] https://josm.openstreetmap.de/.
[4] http://osm.dumoulin63.net/xapiviewer/.
[5] http://wiki.openstreetmap.org/wiki/Xapi.
[6] http://overpass-api.de/.

input a given point, street, building, etc. The advantages of our approach are that our XQuery library makes the definition of queries on top of OSM layers easier. A repertoire of OSM spatial operators are implemented in terms of the spatial operators of Clementini and Egenhofer. Such repertoire of operators is specific for OSM maps, that is, it handles the particular nature of the XML representation of OSM. Our proposal also includes a batch of coordinate (i.e., latitude and longitude) based operators, allowing the definition of interesting geo-localized queries. Higher order functions in XQuery[7] permit definitions of nested queries and keyword based search queries. Queries are expressed in terms of filtering, folding, traversal, composition and set-based operators (union, intersection and difference).

For instance, a typical query in our approach is something like: "Retrieve the schools close to a street, wherein *"Calzada de Castro"* street ends" which combines proximity to a street, keywords (i.e., *school*), as well as the OSM spatial operator (i.e., *isEndingTo*). It can be expressed as follows:

```
let $waysAllEndingTo :=
fn:filter(
rt:getLayerByName(.,"Calle Calzada de Castro"),
osm:isEndingTo(osm:getElementbyName(., "Calle Calzada de Castro"),?))
return
fn:filter(
fn:for-each($waysAllEndingTo, rt:getLayerByName(.,?)),
osm:searchTags(?,"school"))
```

which uses Higher-order functions (i.e., *filter* and *for-each*) of XQuery, for filtering and traversal OSM elements, respectively.

A good performance of query processing is ensured by the OSM data indexing. An R-tree structure, implemented as an XML document, is used to index OSM nodes and ways, which are enclosed by MBRs. Using the R-tree structure, we are able to efficiently retrieve the OSM elements *close* to a given OSM element, and thus, to process in reasonable time, queries focused on the vicinity of a point, street, building, etc. even for large city maps. Thus, for geo-localized queries, we can get better answer times.

We have implemented our library with the *BaseX* XQuery processor [15]. The implementation is based on the transformation of geometric shapes of OSM into the corresponding GML data. Then GML data are handled by the *Java Topology Suite* (*JTS*) [28], an open source API that provides a spatial object model and a set of spatial operators. JTS is available for most of XQuery processors due to the *XQuery Java Binding* mechanism. This is the case of *Exist* [24] and *Saxon* [19] processors, as well as *BaseX*. Thus, the library is portable to other XQuery implementations. We have also tested our approach with the *JOSM* tool [17], working with the XML representation of OSM data, but customized with an style to highlight results of the queries. We have evaluated our library with datasets of several sizes. Finally, the developed library is available at http://indalog.ual.es/osm. The examples shown in this paper can also be downloaded here.

[7] http://www.w3.org/TR/xpath-functions-30/#higher-order-functions.

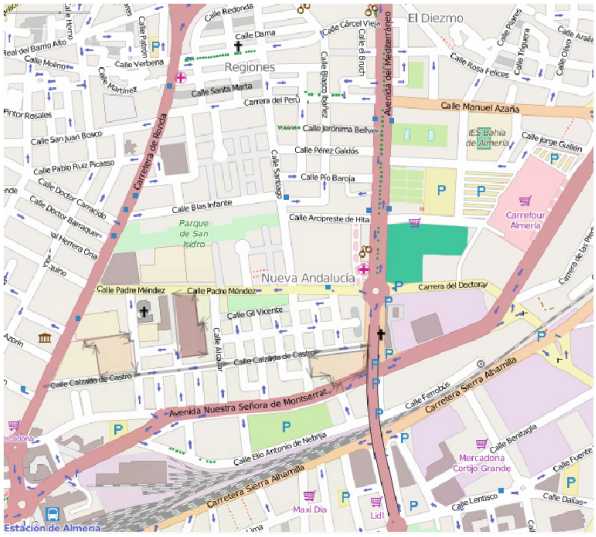


Fig. 1. (Spain) Almería city map.

The rest of this article is organized as follows. Section 2 will present the basic elements of Open Street Map. Section 3 will define the XQuery library. Section 4 will show examples of queries and give benchmarks for several datasets. Section 5 will compare with related work and finally, Sect. 6 will conclude and present future work.

2 Open Street Map

OpenStreetMap uses a topological data structure which includes the following core elements: (1) *Nodes* which are points with a geographic position, stored as coordinates (pairs of a latitude and a longitude) according to WGS84. They are used in ways, but also to describe map features without a size like points of interest or mountain peaks. (2) *Ways* are ordered lists of nodes, representing a poly-line, or possibly a polygon if they form a closed loop. They are used in streets and rivers as well as areas: forests, parks, parkings and lakes. (3) *Relations* are ordered lists of nodes, ways and relations. Relations are used for representing the relationship of existing nodes and ways. (4) *Tags* are key-value pairs (both arbitrary strings). They are used to store *metadata* about the map objects (such as their type, their name and their physical properties). Tags are attached to a node, a way, a relation, or to a member of a relation.

As an example of OSM, Fig. 1 shows the visualization with JOSM of a piece of the Almería (Spain) city map. In order to represent such a map, OSM uses XML labels: *node*, *way* and *relation*, and each label can have several attributes; for instance, *node* has *lat* and *lon*, among others, for representing latitude and longitude of the node. A node can have nested *tags* for adding information about

the node, using attribute pairs key (k) and value (v) with this end. For instance, the museum "Museo Arqueologico" of Almería city is represented as follows:

```
<node lat='36.8386557' lon='-2.4556049'>
    <tag k='name' v='Museo Arqueologico' />
    <tag k='tourism' v='museum' />
</node>
```

The main element of the OSM is the *way* that serves not only to represent streets but also buildings, parkings, etc. Ways are described by a sequence of node references, called *nd*, which link ways to nodes, and *tag*s as follows:

```
<way>
    <nd ref='-3625' />
    <nd ref='-3623' />
    <nd ref='-3621' />
    <tag k='highway' v='residential' />
    <tag k='name' v='Calle Calzada de Castro' />
</way>
```

When the way represents a building, parking, etc., specific nested tags are used, for instance:

```
<way id='27161540'>
    <nd ref='298004115' />
    <nd ref='298004116' />
    <nd ref='298004119' />
    <nd ref='298004128' />
    <nd ref='298004115' />
    <tag k='amenity' v='parking' />
</way>
```

In spite of the simplicity of the XML representation of OSM, many features[8] in a OSM layer can be described. Finally, relations are used to relate elements of the map, for instance, bus routes:

```
<relation id='147091'>
    <member type='way' ref='27197940' role='3,11,12' />
    <member type='way' ref='27197939' role='3,7,11,12' />
    <member type='way' ref='35031199' role='3,11,12' />
    <member type='way' ref='27197944' role='7' />
    <member type='way' ref='27197945' role='7' />
    <member type='way' ref='25586878' role='3,11,12' />
    <member type='way' ref='30953417' role='3,11,12' />
    <member type='way' ref='25585669' role='3,5,6,11,12' />
    <member type='way' ref='27161590' role='5,6,12' />
    <member type='way' ref='27210271' role='3' />
    <member type='way' ref='31484654' role='12' />
    <member type='way' ref='27210293' role='3,5,6,12' />
    <member type='way' ref='50004718' role='12' />
<tag k='route' v='bus' />
<tag k='type' v='route' />
</relation>
```

3 XQuery Library for OSM

Our main goal is to provide a repertoire of OSM Operators, implemented as a XQuery library which, in combination with higher order facilities of XQuery,

[8] http://wiki.openstreetmap.org/wiki/Map_Features.

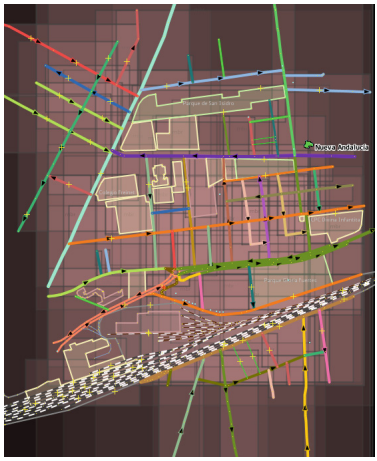


Fig. 2. R-tree based indexing of OSM Maps.

enables the definition of queries over OSM maps. Moreover, we have to ensure shorter answer time for large maps. An R-tree structure to index OSM maps has been implemented, and a set of XQuery functions to retrieve the layer of objects close to a given node/way has been developed. Next, we will show the elements of the XQuery library, which includes:

(1) *OSM Indexing* to generate an R-tree and retrieve elements from it,
(2) *Transformation Operators* to transform OSM geometries into GML ones,
(3) *OSM Spatial Operators* to check spatial relations over OSM geometries,
(4) *Higher Order functions* to facilitate the composition of queries and keyword based search queries.

3.1 OSM Indexing

In order to handle large city maps, in which the layer can include many objects, an R-tree structure to index objects is used. The R-tree structure is based, as usual, on MBRs to hierarchically organize the content of an OSM map. Moreover, MBRs are also used to enclose the nodes and ways of OSM in the leaves of such structure. Figure 2 shows a visual representation of the R-tree of a OSM layer for the Almería city map (streets are highlighted in different colors and MBRs are shown in black color).

The R-tree structure has been implemented as an XML document. That is, the tag based structure of XML is used for representing the R-tree. A tag called *node* represents the inner nodes (i.e., MBRs), while a tag called *leaf* stores the MBRs of OSM ways and nodes. The tag *mbr* is used to represent MBRs. For instance, the R-tree of the OSM map of Fig. 1 is represented in XML as follows:

```
<node x="-2.4574724" y="36.8305714"  z="-2.4473768" t="36.849285">
<node x="-2.4565026" y="36.8319462" z="-2.4476476" t="36.849285">
<node x="-2.4557511" y="36.8319462" z="-2.4491401" t="36.8414807">
<leaf x="-2.4557511" y="36.8347249" z="-2.4522051" t="36.8396123">
<mbr x="-2.4533564" y="36.8383646" z="-2.452359" t="36.8384662">
<way>....</way>
</mbr>
```

The root element of the XML document is the root node of the R-tree, and the children can be also nodes and, in particular, leaves. x, y, z and t attributes of *node*s are the left (x, y) and right corners (z, t) of the MBRs. MBRs are also represented by left and right corners.

We have implemented in XQuery a set of functions to handle R-trees. The function *load_file* generates a R-tree from an OSM layer. The function *getLayerbyName* obtains, given the name of a node or a way, the nodes or ways of the OSM layer whose MBRs *overlap* the MBRs of the given node or way. In case of points, overlapping means inclusion. In other words, *getLayerbyName* retrieves the elements that are *close* to the given node or way. Additionally, we have implemented the following two functions: *getElementByName* to retrieve an OSM element by name, and *getElementsByKeyword* to retrieve OSM elements by keyword. *getElementByName* and *getElementsByKeyword* does not use the spatial indexing, rather than the XML indexing of the document is used.

Our proposed query language uses *getLayerbyName* as basis, in the sense that, queries have to be focused on a certain area of interest, given by the name of a node (park, pharmacy, etc.,) or by the name of a way. In other words, our query language is useful and efficient for geo-localized queries. Once the layer of the area of interest is retrieved, the repertoire of OSM operators in combination with higher order functions can be applied to produce complex queries. The answer of a query is an OSM layer including nodes and ways of the area of interest. Nevertheless, *getElementsByKeyword* can be used to retrieve OSM elements by keyword, and thus enabling keyword based queries. Using keywords instead node and way names the XML indexing is used.

3.2 Transformation Operators

In order to handle OSM elements (i.e., nodes and ways), OSM geometries of these elements have to be transformed into GML data. Once transformed, GML data are handled by the JTS library based on Clementini's operators. Functions *_osm2GmlLine* and *_osm2GmlPoint* have been defined with this end, in order to transform OSM *ways* and *nodes* into GML multi-lines and points, respectively. *_osm2GmlLine* is defined as follows:

```
declare function osm_gml:_osm2GmlLine($way as node()){
if ($way//way) then
<gml:MultiLineString>{<gml:LineString><gml:coordinates>{
for $node in $way/node
return (concat(concat(data($node/@lat),','),data($node/@lon)))
} </gml:coordinates></gml:LineString>}</gml:MultiLineString>
```

```
else
<gml:Point>{<gml:coordinates>{
for $node in $way/node
return (concat(concat(data($node/@lat),',',),data($node/@lon)))
} </gml:coordinates>}</gml:Point>
};
```

3.3 OSM Spatial Operators

A repertoire of *OSM Operators* suitable for OSM city maps has been designed. The repertoire is specific for OSM maps, handling the XML representation of OSM. Clementini's operators are shown in Fig. 3[9]. and our proposal of (Boolean) *OSM Operators* is shown in Figs. 4 and 5. We consider two kinds of operators:

(a) *Coordinate based OSM Operators*, shown in Fig. 4;
(b) *Clementini based OSM Operators*, shown in Fig. 5.

They are designed to cover most of urban queries involving nodes and ways, and geo-positioning: streets at north, points at east, and so on; the street in which a given point is located; if two points are located in the same street; whether two streets are crossing in any point or not; whether a street ends to another one, and finally, whether a street is a continuation of another one.

Name	*Definition*
Equals(a,b)	Their interiors intersect and no part of the interior or boundary of one geometry intersects the exterior of the other
Disjoint(a,b)	They have no point in common
Touches(a,b)	They have at least one boundary point in common, but no interior points
Contains(a,b)	No points of b lie in the exterior of a, and at least one point of the interior of b lies in the interior of a
Covers(a,b)	Every point of b is a point of (the interior of) a
Crosses(a,b)	They have some but not all interior points in common (and the dimension of the intersection is less than at least one of them)
Overlaps(a,b)	They have some but not all points in common, they have the same dimension, and the intersection of the interiors of the two geometries has the same dimension as the geometries themselves

Fig. 3. Clementini Spatial Operators.

Next, we will show the implementation of our OSM Operators. For instance, the coordinate based operator *furtherNorthPoints*, which is *true* whenever the first point is further north than the second point, is defined as follows:

[9] Clementini has also defined the logic negation of some operators, that is, *Intersects* (for *Disjoint*), *Within* (for *Contains*) and CoveredBy (for *Covers*).

Name	*Definition*	*Spatial Operator*
furtherNorthPoints(p1,p2)	Returns true whenever the point p1 is further north than the point p2	Using latitudes of nodes in north and south hemispheres
furtherSouthPoints(p1,p2)	Returns true whenever p1 is further south than p2	furtherNorthPoints negation
furtherEastPoints(p1,p2)	Returns true whenever p1 is further east than p2	Using latitudes of nodes in west and east hemispheres
furtherWestPoints(p1,p2)	Returns true whenever p1 is further west than p2	furtherEastPoints negation
furtherNorthWays(s1,s2)	Returns true whenever all points of the street s1 are further north than all points of the street s2	Using furtherNorthPoints
furtherSouthWays(s1,s2)	Returns true whenever all points of s1 are further south than all points of s2	furtherNorthWays negation
furtherEastWays(s1,s2)	Returns true whenever all points of s1 are further east than all points of s2	Using furtherEastPoints
furtherWestWays(s1,s2)	Returns true whenever all points of s1 are further west than all points of s2	furtherEastWays negation

Fig. 4. Coordinate based OSM Operators.

```
declare function osm:furtherNorthPoints($node1 as node(), $node2 as node())
{
  let $lat1 := $node1/@lat, $lat2 := $node2/@lat
  return
  (: Case 1:  both nodes in positive Ecuador hemisphere :)
    if ($lat1 > 0 and $lat2 > 0) then
        if (($lat2 - $lat1) > 0) then true()
                                 else false()
    else
  (: Case 2: both nodes in negative Ecuador hemisphere :)
      if ($lat1 < 0  and $lat2 < 0) then
      if (((-$lat2) - (-$lat1)) < 0) then true()
                                     else false()
      else
  (: Case 3: First node in positive Ecuador hemisphere,
             Second node in negative Ecuador hemisphere:)
     if ($lat1 > 0 and $lat2 < 0) then false()
  (: Case 4: First node in negative Ecuador hemisphere,
            Second node in positive Ecuador hemisphere :)
                                  else true()
};
```

The Clementini based operator *inWay*, which checks whether a point is located in a street, is defined, using Clementini's operator *contains*, as follows:

```
declare function osm:inWay($point as node(), $way as node()) {
  let $point := osm_gml:_osm2GmlPoint($point),
```

Name	*Definition*	*Clementini's Operator*
inWay(p,s)	Returns true whenever p (point) is in s (street)	Contains
inSameWay(p1,p2)	Returns true whenever p1 (point) and p2 (point) are in the same way	Equals
isCrossing(s1,s2)	Returns true whenever s1 (street) crosses s2 (street)	Crosses
isNotCrossing(s1,s2)	Returns true whenever s1 does not cross s2	Disjoint
isEndingTo(s1,s2)	Returns true whenever s1 ends to s2	Touches (neither initial nor final point)
isContinuationOf(s1,s2)	Returns true whenever s2 is a continuation of s1	Touches (either initial or final point)

Fig. 5. Clementini based OSM Operators.

```
      $line := osm_gml:_osm2GmlLine($way)
  return geo:contains($line,$point)
};
```

The Clementini based operator *inSameWay*, which returns *true* whether two points are located in the same street, uses the auxiliary function *WaysOfaPoint* to retrieve the street (or streets) in which the points are located. *inSameWay* uses Clementini's operator *equals*, and is defined as follows:

```
declare function osm:inSameWay($node1 as node(), $node2 as node(), $document as node()*)
{
  let
      $way1 := osm:WaysOfaPoint($node1,$document),
      $way2 := osm:WaysOfaPoint($node2,$document)
  return
   some $x in $way1 satisfies
   (some $y in $way2 satisfies
    (let $line1 := osm_gml:_osm2GmlLine($x),
              $line2 := osm_gml:_osm2GmlLine($y)
                     return geo:equals($line1,$line2)))
};
```

Now, the Clementini based operator *isCrossing*, which checks if two streets are crossing, is defined, by using Clementini's operator *crosses*, as follows:

```
declare function osm:isCrossing($way1 as node(), $way2 as node()) {
  osm:booleanQuery($way1,$way2,"geo:crosses")
};
```

Here, a *Boolean query pattern* is used, called *booleanQuery*, which makes the definition of the Clementini based OSM operators easier, and is defined as follows:

```
declare function osm:booleanQuery($way1 as node(), $way2 as node(), $functionName as xs:string)
{
  let $mutliLineString1 := osm_gml:_osm2GmlLine($way1),
      $multiLineString2 := osm_gml:_osm2GmlLine($way2)
  let $spatialFunction :=
                 fn:function-lookup(xs:QName($functionName),2)
  return $spatialFunction($mutliLineString1,$multiLineString2)
};
```

Name	*Semantics*
fn:for-each(s,f)	Applies the function f to every element of the sequence s
fn:filter(s,p)	Selects the elements of the sequence s for which p is true
fn:for-each-pair(s1,s2,f)	Zips the elements of s1 and s2 with the function f
fn:fold-left(s,e,f)	Folds (left) the sequence s with f starting from e
fn:fold-right(s,e,f)	Folds (right) the sequence s with f starting from e

Fig. 6. Higher order functions of XQuery.

This pattern takes as arguments two *streets* and a *functionName*. *functionName* is a Clementini's operator from JTS. The patterns applies the *functionName* to the streets. The Boolean query pattern is also used for the implementation of *isNotCrossing*. The cases *isEndingTo* and *isContinuationOf* are special cases of OSM operators that are not direct instances of the Boolean query pattern. They can be derived from Clementini's operators. Both functions use the Clementini's operator *touches*, as well as start and end point of the street, in order to check the ending or continuation of the street. For instance, *isEndingTo* is defined as follows:

```
declare function osm:isEndingTo($way1 as node(),$way2 as node())
{
 if (osm:booleanQuery($way1,$way2,"geo:touches"))
 then
    let $mutliLineString1 := osm_gml:_osm2GmlLine($way1),
        $multiLineString2 := osm_gml:_osm2GmlLine($way2),
        $intersection_point := geo:intersection($mutliLineString1,$multiLineString2),
        $start_point := geo:start-point($mutliLineString1/*),
        $end_point := geo:end-point($mutliLineString1/*)
    return
            (geo:equals($intersection_point/*,$start_point/*) or
            geo:equals($intersection_point/*,$end_point/*))
    else false()
};
```

3.4 Higher Order XQuery Facilities

On the other hand, XQuery 3.0 is equipped with higher order facilities. It makes possible to define functions in which arguments can also be functions. Moreover, XQuery provides a library of built-in higher order functions (see Fig. 6). Making use of this capability, our library has been equipped with functions enabling: (a) Query composition by combining higher order functions and Coordinate and Clementini based operators; (b) Keyword based search queries by combining higher order functions and keyword based operators.

For instance, with respect to (a), the higher order function *filter* together with the OSM operator *isCrossing* can be used to retrieve all the streets crossing a given street (for instance, *"Calzada de Castro"* in the Almería city map) as follows. Note the simplicity and natural formulation of this query.

```
fn:filter(rt:getLayerByName(.,"Calle Calzada de Castro"),
osm:isCrossing(?, osm:getElementbyName(., "Calle Calzada de Castro")))
```

Here, *getLayerByName* returns all the streets close to *"Calle Calzada de Castro"* street[10] from the indexed OSM layer, and *getElementbyName* retrieves

[10] "Calle" means street in spanish.

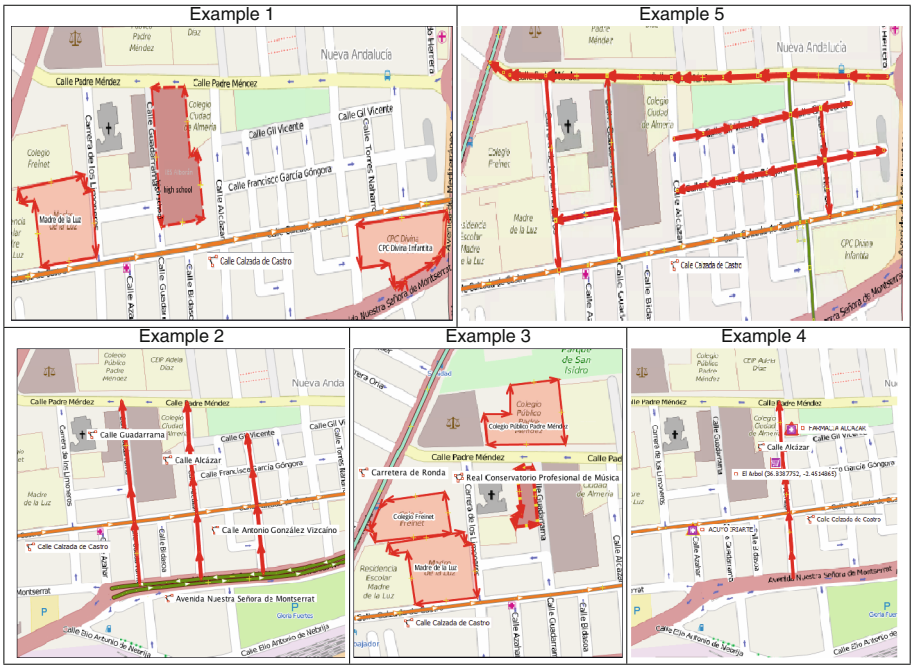


Fig. 7. Results for Examples.

"Calzada de Castro" street (i.e., the OSM way representing *"Calzada de Castro"*). The symbol "?" indicates which argument of *isCrossing* is filtered.

With respect to (b), a new repertoire of keyword based operators has been defined for adding (*addTag*), removing (*removeTag*), replacing (*replaceTag*) and retrieving (*searchOneTag* and *searchTags*) *keywords* on OSM elements. Keyword based operators work on OSM nodes and ways. For instance, the function *searchTags*, which searches a set of keywords in a OSM element, is defined as follows:

```
declare function osm:searchTags($node as node(),$SetofKeywords as xs:string*)
{
  some $value in
  (distinct-values(
  for $keyword in $SetofKeywords
    return osm:searchOneTag($node,$keyword)))
                    satisfies ($value = true()))
};
```

As example of use *searchTags* in combination with the higher order function *filter* can be used to retrieve all the schools close to *"Calzada de Castro"* street from the indexed OSM map. Keyword search is restricted in this case to a geolocalized street (*"Calzada de Castro"*):

```
fn:filter(rt:getLayerByName(.,"Calle Calzada de Castro"), osm:searchTags(?,"school"))
```

4 Examples

In this section, we show some examples of use of our library. In addition, we also provide benchmarks from datasets of several sizes. Assuming the map of Fig. 1 (i.e., Almería city), we can consider the following batch of queries whose results are depicted in Fig. 7.

Example 1: Retrieve the schools and high schools close to *"Calzada de Castro"* street:

```
fn:filter(rt:getLayerByName(.,"Calle Calzada de Castro"),
                        osm:searchTags(?,("high school", "school")))
```

In this query, the higher order function *filter* in combination with the function *searchTags* is used. It enables the retrieval of the schools and high schools from the layer; i.e., to search the keywords *"school"* and *"high school"* from the tags included in the layer of *"Calzada de Castro"*. It is assumed that the R-tree has been previously generated and loaded in memory of the XQuery interpreter. The function *getLayerByName* retrieves from the R-tree, the layer of nodes and ways close to *"Calzada de Castro"* street (i.e., those objects whose MBRs overlap with the MBR of *"Calzada de Castro"* street).

Example 2: Retrieve the streets crossing *"Calzada de Castro"* and ending to *"Avenida Montserrat"* street:

```
let $waysCrossing :=
fn:filter(rt:getLayerByName(.,"Calle Calzada de Castro"),
osm:isCrossing(?, osm:getElementbyName(., "Calle Calzada de Castro")))
return
fn:filter($waysCrossing,
osm:isEndingTo(?, osm:getElementbyName(., "Avenida Montserrat")))
```

Here, the function *filter* has been used in combination with the OSM operators *isCrossing* and *isEndingTo*. In this query, first of all the streets crossing *"Calzada de Castro"* street are filtered, and then, from these streets, the streets ending to *"Avenida de Montserrat"* street are retrieved.

Example 3: Retrieve the schools close to a street, wherein *"Calzada de Castro"* street ends.

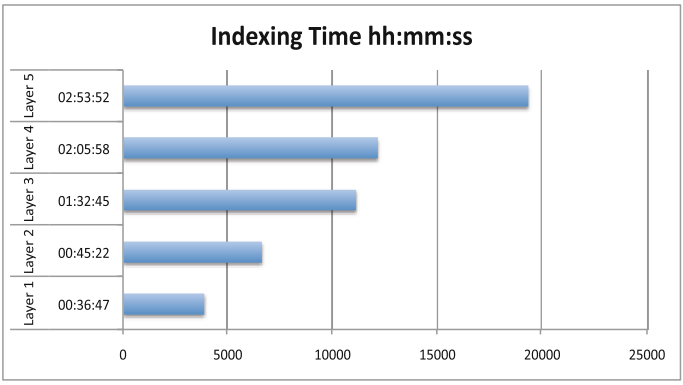
```
let $waysEndingTo :=
fn:filter(rt:getLayerByName(.,"Calle Calzada de Castro"),
osm:isEndingTo(osm:getElementbyName(., "Calle Calzada de Castro"),?))
return
fn:filter(fn:for-each($waysEndingTo, rt:getLayerByName(.,?)),
osm:searchTags(?,"school"))
```

Here, we can see how queries are nested: on the one hand, the OSM operator *isEndingTo* is used to retrieve the streets wherein *"Calzada de Castro"* street ends, and, on the other hand, the keyword *school* is searched from the OSM elements occurring in the layer of each street. *filter* is used twice.

Example 4: Retrieve the streets close to *"Calzada de Castro"* street, in which there is a supermarket *"El Arbol"* or a pharmacy (or chemist's).

```
osm:intersectionQuery(
 osm:unionQuery(
 rt:getLayerByName(.,"El Arbol"),
 rt:getLayerByName(.,"pharmacy")),
 rt:getLayerByName(.,"Calle Calzada de Castro"))
```

Here, we can see an additional feature of our library; i.e. *the handling of set-based operators*, such as *union*, *intersection* and *difference* of sequences. Functions *unionQuery*, *intersectionQuery* and *exceptQuery* of the library can be used to



(a)

	Layer 275 Objects	Layer 950 Objects	Layer 3.519 Objects	Layer 13.764 Objects
Example 5	184,29	237,06	339,62	414,86
Example 4	52,42	103,87	172,24	255,58
Example 3	172,59	236,63	306,64	422,67
Example 2	242,54	273,23	322,77	518,08
Example 1	75,92	108,45	137,89	209,45

(b)

Fig. 8. Benchmarks. Layer (1) 38666 Objects = 5495 Ways + 33171 Nodes (4101 KB) - Layer (2) 49287 Objects = 4244 Ways + 45583 Nodes (6043 KB) - Layer (3) 102620 Objects = 17783 Ways + 88837 Nodes (12 MB) - Layer (4) 168048 Objects = 7019 Ways + 161029 Nodes (16,3 MB) - Layer (5) 207357 Objects = 21258 Ways + 186099 Nodes (20,9 MB) (Color figure online).

produce more complex queries. In this case, the intersection of streets close to *"Calzada de Castro"* street and streets with a supermarket (named *"El Arbol"*) or a pharmacy is requested.

Example 5: Retrieve the streets to the north of *"Calzada de Castro"* street:

```
fn:filter(
rt:getLayerByName(.,"Calle Calzada de Castro"),
osm:furtherNorthWays(
osm:getElementbyName(., "Calle Calzada de Castro"),?))
```

Finally, we can see and example of geo-positioning queries. Streets close to *"Calzada de Castro"* are retrieved, and then, the further north streets are filtered.

4.1 Benchmarks

Now we would like to show the benchmarks with our library. We have used the *BaseX Query* processor in a HP Proliant (two processors and 16 MB RAM Memory) with Windows Server 2008 R2. The goal of the benchmarks is (1) to measure the time required to generate an index (i.e., the R-tree structure) for large datasets, and (2) to measure the time required to execute queries. The time (1) is not so crucial for the performance of the library, since indexes are only generated the first time a given dataset is retrieved. We have experimented with large datasets from maps of several cities; i.e. *Alexandria* (Layer 1), *Santa Barbara* (Layer 2), *Alburquerque* (Layer 3), *Cusco* (Layer 4) and, finally, *Cork* (Layer 5). Sizes range from 38666 to 207357 Objects. For all the cities the execution times for indexing are shown in Fig. 8(a). Moreover, query execution times for Examples 1 to 5, with sizes ranging from two hundred to fourteen thousand objects, corresponding to: from a zoom to *"Calzada de Castro"* street to the whole Almería city map (around 10 km^2 square kilometers), are shown in Fig. 8(b). From the benchmarks, we can conclude that increasing the map size, does not increase, in a remarkable way, the answer time.

Unfortunately, we cannot compare our benchmarks with existent implementations of similar tools due to the following reasons. Even when OSM has been used for providing benchmarks in a recent work [11], they use OSM as dataset for Description Logic based reasoners rather than to evaluate spatial queries. There are some proposals for defining spatial datasets for benchmarking Spatial RDF stores [12,20], mainly focused on Clementini's and Egenhofer's operators whereas our query language offers more sophisticated queries.

5 Related Work

GQuery [7] is an early proposal for adding spatial operators to XQuery. Manipulation of trees and sub-trees are carried out by XQuery, while spatial processing is performed using geometric functions and *JTS*. *GeoXQuery* approach [18] extends the *Saxon* XQuery processor [19] with functions that provide geo-spatial operations. It is also based on *JTS* and a GML to SVG transformation library for the

XQuery processor is defined in order to show query results. *GML Query* [23] is also a contribution in this research line that stores GML documents in a spatial RDBMS. This approach performs a simplification of the GML schema that is then mapped to its corresponding relational schema. The basic values of spatial objects are stored as values of the tables. Once the document is stored, spatial queries can be expressed using the XQuery language with spatial functions. The queries are translated to their equivalent in SQL which are executed by means of the spatial RDBMS. In our approach, we have followed the same direction as [7,18,23], adopting XQuery for querying. Our XQuery library is specifically designed for OSM, while the quoted approaches are focused on GML. Also the introduction of higher order functions in XQuery (which was adopted by the W3C in 2013), and their use for querying spatial data is a novelty of our approach with respect to the quoted works.

With regard to *OSM3S* (i.e., *Overpass API*), focused on OSM, it is specifically designed for search criteria like location, types of objects, tag values, proximity or combinations of them. Overpass API has the query languages *Overpass XML* and *Overpass QL*. Both languages are equivalent. They handle OSM objects ((a) standalone queries) and set of OSM objects ((b) query composition and filtering). With respect to (a), the query language permits the definition of queries to search a particular object, and is equipped with forward or backward recursion to retrieve links from an object (for instance, it permits to retrieve the nodes of a way). With respect to (b), the query language is able to express queries involving several search criteria. Among others, it can express: to find all data in a bounding box (i.e., positioning), to find all data near something else (i.e., proximity), to find all data by tag value (exact value, non-exact value and regular expressions), negation, union, difference, intersection, and filtering, with a rich set of selectors, by polygon, by area pivot, and so on. However, Overpass API facilities (i.e., query composition and filtering) cannot be combined with spatial operators such as Clementini's crossing or touching. In Overpass API, only one type of spatial intersection is considered (proximity zero with the across selector). For instance, the query (expressible in our approach) *"Retrieve the streets crossing Calzada de Castro street and ending to Avenida de Montserrat street"* is not expressible in Overpass API. On the other hand, Overpass API has a rich query language for keyword search based queries. We plan to extend our library to handle a richer set of keyword search based queries.

Linked Geospatial Data is an emerging line of research (see [21] for a survey) focused on the handling of the RDF based representation of geo-spatial information, adopting a Semantic Web point of view [10], and using SPARQL style query languages like SPARQL-ST [25], stSPARQL [22] and GeoSPARQL [4]. The *LinkedGeoData* dataset [29] is a work of the AKSW research group at the University of Leipzig that uses GeoSPARQL and well-known text (WKT) RDF vocabularies to represent OSM data. SPARQL based query languages offer a rich set of spatial operators. For instance, stSPARQL is equipped with Clementini's operators, as well as MBRs based operators. Also directional operators are considered and, functions for constructing new objects are included: buffer, boundary, envelope, convexHull, union, intersection and difference as well as distance-based

operators: distance and area. Finally, temporal operators are also considered. The RDF representation of OSM and the use of SPARQL style query languages, offer also the opportunity to describe more complex queries than OSM3S and XAPI. Although we are working on the XML representation of OSM data, and with XML-based query languages, most of ideas handled by our approach could be adopted by RDF and SPARQL based query languages. Unfortunately, SPARQL and its spatial dialects are not equipped with higher order (although there exists a recent proposal [2] for SPARQL) and thus, the queries we propose, are not so easy to express in them. Even more, spatial dialects of SPARQL have to work with the graph based structure of OSM RDF. It makes the definition of some queries more difficult, if not impossible [1]. The same happens when a spatial RDBMS is used. While spatial RDBMS can offer the same functionality than the proposed XQuery extension, higher order makes the work easier.

6 Conclusions and Future Work

We have presented an XQuery library for querying OSM. We have defined a set of OSM Operators suitable for querying OSM maps. We have shown how higher order facilities of XQuery enable the definition of complex queries over OSM involving composition and keyword searching. We have shown benchmarks for several datasets, revealing that the R-tree structure used to index OSM ensures short answer times. As future work we would like to extend our library to handle distance based queries, aggregation operators and top-k queries. Also we would like to develop a Web-based application to execute and show results of queries.

References

1. Alkhateeb, F., Baget, J.F., Euzenat, J.: Extending SPARQL with regular expression patterns (for querying RDF). Web Semant. Sci. Serv. Agents World Wide Web **7**(2), 57–73 (2011)
2. Atzori, M.: Toward the web of functions: interoperable higher-order functions in SPARQL. In: Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., Goble, C. (eds.) ISWC 2014, Part II. LNCS, vol. 8797, pp. 406–421. Springer, Heidelberg (2014)
3. Bamford, R., Borkar, V., Brantner, M., Fischer, P.M., Florescu, D., Graf, D., Kossmann, D., Kraska, T., Muresan, D., Nasoi, S., et al.: XQuery reloaded. Proc. VLDB Endowment **2**(2), 1342–1353 (2009)
4. Battle, R., Kolas, D.: Enabling the geospatial semantic web with Parliament and GeoSPARQL. Semant. Web **3**(4), 355–370 (2012)
5. Bennett, J.: OpenStreetMap - Be your own cartographer. Packt Publishing Ltd. (2010)
6. Berglund, A., Boag, S., Chamberlin, D., Fernandez, M., Kay, M., Robie, J., Siméon, J.: XML path language (XPath) 2.0. W3C (2010)
7. Boucelma, O., Colonna, F.: GQuery: a query language for GML. In: Proceedings of the 24th Urban Data Management Symposium, pp. 27–29 (2004)
8. Clementini, E., Di Felice, P.: Spatial operators. ACM SIGMOD Rec. **29**(3), 31–38 (2000)

9. Egenhofer, M.J.: Spatial SQL: a query and presentation language. IEEE Trans. Knowl. Data Eng. **6**(1), 86–95 (1994)
10. Egenhofer, M.J.: Toward the semantic geospatial web. In: Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems, pp. 1–4. ACM (2002)
11. Eiter, T., Schneider, P., Šimkus, M., Xiao, G.: Using OpenStreetMap data to create benchmarks for description logic reasoners. In: Proceedings of the 3rd International Workshop on OWL Reasoner Evaluation (ORE 2014). CEUR Workshop Proceedings, vol. 1207, pp. 51–57 (2014)
12. Garbis, G., Kyzirakos, K., Koubarakis, M.: Geographica: a benchmark for geospatial RDF stores (long version). In: Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J.X., Aroyo, L., Noy, N., Welty, C., Janowicz, K. (eds.) ISWC 2013, Part II. LNCS, vol. 8219, pp. 343–359. Springer, Heidelberg (2013)
13. Goodchild, M.F.: Citizens as sensors: the world of volunteered geography. GeoJournal **69**(4), 211–221 (2007)
14. Goodchild, M.F., Li, L.: Assuring the quality of volunteered geographic information. Spat. Stat. **1**, 110–120 (2012)
15. Grun, C.: BaseX. The XML Database (2015). http://basex.org
16. Hadjieleftheriou, M., Manolopoulos, Y., Theodoridis, Y., Tsotras, V.J.: R-Trees – A dynamic index structure for spatial searching. In: Shekhar, S., Xiong, H. (eds.) Encyclopedia of GIS, pp. 993–1002. Springer, Heidelberg (2008)
17. Haklay, M., Weber, P.: Openstreetmap: user-generated street maps. IEEE pervasive comput. **7**(4), 12–18 (2008)
18. Huang, C.H., Chuang, T.R., Deng, D.P., Lee, H.M.: Building GML-native web-based geographic information systems. Comput. Geosci. **35**(9), 1802–1816 (2009)
19. Kay, M.: Ten reasons why saxon xquery is fast. IEEE Data Eng. Bull. **31**(4), 65–74 (2008)
20. Kolas, D.: A benchmark for spatial semantic web systems. In: International Workshop on Scalable Semantic Web Knowledge Base Systems (2008)
21. Sioutis, M., Nikolaou, C., Karpathiotakis, M., Kyzirakos, K., Koubarakis, M.: Data models and query languages for linked geospatial data. In: Eiter, T., Krennwallner, T. (eds.) Reasoning Web 2012. LNCS, vol. 7487, pp. 290–328. Springer, Heidelberg (2012)
22. Koubarakis, M., Kyzirakos, K.: Modeling and querying metadata in the semantic sensor web: the model stRDF and the query language stSPARQL. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010, Part I. LNCS, vol. 6088, pp. 425–439. Springer, Heidelberg (2010)
23. Zhou, S., Li, Y., Li, J.: GML storage: a spatial database approach. In: Wang, S., Tanaka, K., Zhou, S., Ling, T.-W., Guan, J., Yang, D., Grandi, F., Mangina, E.E., Song, I.-Y., Mayr, H.C. (eds.) ER Workshops 2004. LNCS, vol. 3289, pp. 55–66. Springer, Heidelberg (2004)
24. Meier, W.: eXist: an open source native XML database. In: Chaudhri, A.B., Jeckle, M., Rahm, E., Unland, R. (eds.) NODe-WS 2002. LNCS, vol. 2593, pp. 169–183. Springer, Heidelberg (2003)
25. Perry, M., Jain, P., Sheth, A.P.: SPARQL-ST: extending SPARQL to support spatiotemporal queries. In: Ashish, N., Sheth, A.P. (eds.) Geospatial Semantics and the Semantic Web, pp. 61–86. Springer, New York (2011)
26. Ramm, F., Topf, J., Chilton, S.: OpenStreetMap: using and enhancing the free map of the world. UIT Cambridge, Cambridge (2011)

27. Robie, J., Chamberlin, D., Dyck, M., Snelson, J.: XQuery 3.0: An XML query language. W3C (2014)
28. Shekhar, S., Xiong, H.: Java topology suite (JTS). In: Shekhar, S., Xiong, H. (eds.) Encyclopedia of GIS, pp. 601–601. Springer, New York (2008)
29. Stadler, C., Lehmann, J., Höffner, K., Auer, S.: Linkedgeodata: a core for a web of spatial open data. Semant. Web **3**(4), 333–354 (2012)

The K Group Nearest-Neighbor Query on Non-indexed RAM-Resident Data

George Roumelis[1], Michael Vassilakopoulos[2], Antonio Corral[3(✉)], and Yannis Manolopoulos[1]

[1] Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
{groumeli,manolopo}@csd.auth.gr
[2] Department of Electrical and Computer Engineering, University of Thessaly, Volos, Greece
mvasilako@inf.uth.gr
[3] Department of Informatics, University of Almeria, Almería, Spain
acorral@ual.es

Abstract. Data sets that are used for answering a single query only once (or just a few times) before they are replaced by new data sets appear frequently in practical applications. The cost of buiding indexes to accelerate query processing would not be repaid for such data sets. We consider an extension of the popular (K) Nearest-Neighbor Query, called the (K) Group Nearest Neighbor Query (GNNQ). This query discovers the (K) nearest neighbor(s) to a group of query points (considering the sum of distances to all the members of the query group) and has been studied during recent years, considering data sets indexed by efficient spatial data structures. We study (K) GNNQs, considering non-indexed RAM-resident data sets and present an existing algorithm adapted to such data sets and two Plane-Sweep algorithms, that apply optimizations emerging from the geometric properties of the problem. By extensive experimentation, using real and synthetic data sets, we highlight the most efficient algorithm.

Keywords: Spatial query processing · Plane-sweep · Group nearest-neighbor query · Algorithms

1 Introduction

Spatial database is a database that offers spatial data types (for example, types for points, line segments, regions, etc.), a query language with spatial predicates, spatial indexing techniques and efficient processing of spatial queries [1]. It has

G. Roumelis, M. Vassilakopoulos, A. Corral and Y. Manolopoulos—Work funded by the GENCENG project (SYNERGASIA 2011 action, supported by the European Regional Development Fund and Greek National Funds); project number 11SYN_8_1213.

A. Corral—Supported by the MINECO research project [TIN2013-41576-R].

C. Grueau and J.G. Rocha (Eds.): GISTAM 2015, CCIS 582, pp. 69–89, 2016.
DOI: 10.1007/978-3-319-29589-3_5

grown in importance in several fields of application such as urban planning, resource management, transportation planning, etc. Together with them come various types of complex queries that need to be answered efficiently.

One of the most representative and studied queries in Spatial Databases is the (K) Nearest-Neighbor Query (NNQ), that discovers the (K) nearest neighbor(s) to a query point. An extension that is important for practical applications is the (K) Group Nearest Neighbor Query (GNNQ), that discovers the (K) nearest neighbor(s) to a group of query points (considering the sum of distances to all the members of the query group). This query has been studied during recent years, considering data sets indexed by efficient spatial data structures. An example of its utility could be when we have a set of meeting points (data set) and a set of user locations (query set), and we want to find the set of one (K) meeting point(s) that minimizes the sum of distances for all user locations, since each user will travel from his/her location to each of the K meeting points. More specifically, user locations may represent residence locations and meeting points may represent points of interest (cultural landmarks). Each of the K points is visited by each user for whole day inspection and the user returns to his/her residence overnight, before visiting the next landmark on the following day. We may interested to solve such a problem for a specific pair of data and query sets only once, but we may face several such problems for different pairs of sets. Building indexes for the data sets would be needed only if several queries would be answered for the these sets, which might evolve gradually in the course of time and not be completely replaced by new ones.

One of the most important techniques in the computational geometry field is the Plane-Sweep (PS) algorithm, which is a type of algorithm that uses a conceptual sweep line to solve various problems in the Euclidean plane, E^2, [2]. The name of PS is derived from the idea of sweeping the plane from left to right with a vertical line (front) stopping at every transaction point of a geometric configuration to update the front. All processing is done with respect to this moving front, without any backtracking, with a look-ahead on only one point each time [3]. For instance, the PS technique has been successfully applied in spatial query processing, mainly for intersection joins [4].

In [5], the problem of processing K Closest Pair Query between RAM-based point sets was studied, using PS algorithms. Two improvements that can be applied to a PS algorithm and a new algorithm that minimizes the number of distance computations, in comparison to the classic PS algorithm, were proposed. By extensive experimentation, using real and synthetic data sets, the most efficient improvement was highlighted and it was shown that the new PS algorithm outperforms the classic one.

In this paper, we study (K) GNNQs, considering non-indexed data sets (a frequent case in practical applications, see the example given previously), unlike previous research presented in Sect. 2 that consider that one or both data sets are indexed by structures of the R-tree family. Our target is to design efficient non-index based algorithms for (K) GNNQs and highlight the most efficient among

them. Thus, we present three (RAM-based) algorithms[1], an existing one adapted to non-indexed data sets and two novel PS ones, that apply optimizations emerging from the geometric properties of the problem. Several experiments have been performed, using real and synthetic data sets, to show the most efficient algorithm.

The paper is organized as follows. In Sect. 2, we review the related literature and motivate the research reported here. In Sect. 3, three new PS algorithms for GNNQs are presented. In Sect. 4, a comparative performance study is reported. Finally, in Sect. 5, conclusions on the contribution of this paper and future work are summarized.

2 Related Work and Motivations

GNN queries are introduced in [7] and it consist in given two sets of points P and Q, a GNN query retrieves the point(s) of P with the smallest sum of distances to all points in Q. GNN queries are also known as aggregate nearest neighbor (ANN) queries [8]. In [7], the authors have developed three different methods, MQM (multiple query method), SPM (single point method) and MBM (minimum bounding method), to evaluate a GNN query that minimizes the total distance from a set of query points to a data point. In [8] these methods have been extended to minimize the minimum and maximum distance in addition to the total distance with respect to a set of query points. All these methods assume that the data points are indexed using an R-tree and can be implemented using both depth-first search and best-first search algorithms.

In general terms, MQM performs an incremental search for the nearest data point of each query point in the set and compute the aggregate distance from all query points for each retrieved data point. The search ends when it is ensured that the aggregate distance of any non-retrieved data point in the database is greater than the current K-th minimum aggregate distance, that is the K GNNs are found. It means MQM is a threshold algorithm, since it computes the nearest neighbor for each query point incrementally, updating different thresholds according to the target of the (K) GNN. The main disadvantage of MQM is that it traverses the R-tree multiple times and it can access the same data point more than once.

The other methods, SPM and MBM, find the K GNNs in a single traversal of the R-tree. SPM approximates the centroid of the query distribution area and continues the searching with respect to the centroid until the current (K) GNNs are determined. During the search, some heuristics based on triangular inequality are used to prune intermediate nodes and determine the real nearest neighbors to Q. MBM regards Q as a whole and uses its MBR M to prune the search space in a single query, in either a depth-first or best-first manner. Moreover, two pruning heuristics involving the distance from an intermediate node to M or query points are proposed and they can be used in either traversal

[1] This paper is a post proceedings enhanced version of [6], where the last two algorithms of the current paper are presented and compared.

policy. Experimental results showed that the performance of MBM is better than SPM and MQM for memory and disk resident query points, since it traverses the R-tree once and takes the query distribution area into account. Moreover, according to the comparison conducted in [7], MBM is better than SPM in terms of node access and CPU cost while MQM is the worst.

In [9], the authors propose two pruning strategies for (K) GNN queries which take into account the distribution of query points. Such methods employ an ellipse to approximate the extent of multiple query points, and then derive a distance or minimum bounding rectangle using that ellipse to prune intermediate nodes in a depth-first search via an R*-tree. These methods are also applicable to the best-first traversal. The experimental results show that the proposed pruning strategies are more efficient than the methods presented in [7].

A new method to evaluate a (K) GNN query for non-indexed data points using projection-based pruning strategies was presented in [10]. Two points projecting-based ANNQ algorithms were proposed, which can efficiently prune the data points without indexing. This new method projects the query points into a special line, on which their distribution is analysed, for pruning the search space.

In [11], a new property in vector space was proposed and, based on it some efficient bound estimations were developed for two most popular types of ANN queries (sum and maximum). Taking into account these bounds, indexed and non-index ANN algorithms were designed. The proposed algorithms showed interesting results, especially for high dimensional queries.

Other related contributions in this research line have been proposed in the literature. In [12] an efficient algorithm for (K) GNN query considering privacy preserving was proposed, and the existing (K) GNN algorithms [8] for point locations were extended to regions in order to preserve user privacy. In [13], the (K) GNN query in road networks based on network voronoi diagram was solved. In [14], the reverse top-K group nearest neighbor search is presented. In [15], the KNN and (K) GNN queries are extended to get a new type of query, so-called K Nearest Group (KNG) query. It retrieves closest elements from multiple data sources, and finds K groups of elements that are closest to a given query point, with each group containing one object from each data source. And recently, for uncertain databases, probabilistic (K) GNN query was studied by [16,17].

Therefore, the (K) GNN is an active research line nowadays and most of the contributions have used indexes (of the R-tree family) for their solutions. The main motivation of this paper is to exmanine the use of the Plane-Sweep technique to solve the problem proposed in [7], when neither of the inputs are indexed. Due to not using indexes, the algorithms proposed in this paper are completely different to previous solutions. To the best of our knowledge, there are not any existing solutions for the (K) GNNQ without indexes. The unnecessity of indexes is not infrequent in practical applications, when the data sets change at a very rapid rate, or the data sets are not reusable for subsequent queries (see the example in Sect. 1).

3 RAM-Based Algorithms for GNNQ

In this section we introduce three RAM-based algorithms for processing GNNQ. The input of this query consists of a set $P = \{p_0, p_1, \cdots, p_{N-1}\}$ of static data points in the Euclidean plane, E^2, and a group of query points $Q = \{q_0, q_1, \cdots, q_{M-1}\}$. The output contains the K (≥ 1) data point(s) with the smallest sum of distances to all points in Q.

The distance between a data point $p \in P$ and Q is defined as $sumdist(p, Q) = \sum_{i=0}^{M-1} dist(p, q_i)$, where $dist(p, q_i)$ is the Euclidean distance between $p \in P$ and a query point $q_i \in Q$. In the following, $dx\_dist(p, q)$ represents the dx-distance ($\Delta x(p, q)$) between two points p and q over the X-axis and $dy\_dist(p, q)$ represents the dy-distance ($\Delta y(p, q)$) over the Y-axis. The sum of distances (dx-distances) between one given point $p \in P$ and all query points of Q ($q_i \in Q$) is defined as $sumdist(p, Q) = \sum_{i=0}^{M-1} dist(p, q_i)$ ($sumdx(p, Q) = \sum_{i=0}^{M-1} dx\_dist(p, q_i)$).

The **first algorithm** that we present is called Single Point Method over Non-Indexed Data ($SPMNI$) and is a non-indexed data extension/reformation of the SPM algorithm proposed in [7] (this is the most efficient algorithm of [7] that can be adapted to non-indexed data, since MBM is based on the MBR concept used in tree and other indexes). Instead of the sorted list used in the SPM algorithm, we used a max binary heap (keyed by *sumdist* and called $MaxKHeap$) to keep the K data points with the smallest sum of distances to the query points found so far (the *sumdist* of the root of the $MaxKHeap$ is denoted by δ). In order to sort the points of the P data set acccording to their distance to the centroid (c) of query points, the $SPMNI$ algorithm uses an array of $|P|$ length named Centroid Nearest Neighbor List ($cNNl$). Every element of $cNNl$ is a pair of type $< i, dist(p, c) >$ where i is the index of the point p of P set and $dist(p, c)$ is the distance between the point p and the centroid c.

The algorithm works as follows. First, the algorithm, calculates the coordinates of the centroid, computes the sum of distances of the centroid to the query points ($sumdistCQ$), and after creates and sorts the $cNNl$ (preparation stage). For these, $SPMNI$ calls the functions $Calculate\_Centroid\_Coord(Q)$ (line 1), $Create\_CentroidNN\_List(P, c)$ (line 5), and $Sort\_CentroidNN\_List$ (line 6).

The search process of the $KGNN$ starts and until $MaxKHeap$ is full the steps bellow are repeated. The index i of the point p of the P set which is the next NN to the centroid c and the distance to the centroid $dist(p, c)$ is retrieved from the $cNNl$ list using the value of the index of the $cNNl$ list, j. Next, the $sumdist(p, Q)$ is calculated and the $< p, sumdist(p, Q) >$ pair is inserted into the heap. The index j is incremented in order to point to the next NN to the centroid c, and the iteration is repeated, unless the $MaxKHeap$ has become full.

When the $MaxKHeap$ is full, the same steps are repeated with a few differences. In [7] it was proved that for every data point p with $|Q| \cdot dist(p, c) \geq \delta + sumdist(c, Q)$, p can be ignored, without calculating any distance to the query points. Since the left part of the previous inequality grows for every sub-

Algorithm 1. SPMNI.

Input: Two X-sorted arrays of points $p[0, 1, \cdots, N-1]$, $q[0, 1, \cdots, M-1]$, and $MaxKHeap$.

Output: $MaxKHeap$ storing the K NNs having smallest sums of distances to all query points.

1: $c(x, y) = Calculate\_Centroid\_Coord(Q)$ ▷ calculate the coordinates of the Centroid
2: $sumdistCQ = 0.0$
3: **for** $k = 0; k < M; k++$ **do** ▷ for each query point q
4: $sumdistCQ+ = dist(c, q[k])$
5: $cNNl$=Create_CentroidNN_List(P,c) ▷ create the list of NN to the centroid
6: Sort_CentroidNN_List($cNNl$) ▷ sort entries of the list $cNNl$ according to their $dist$
7: $j = 0$
8: **while** $MaxKHeap$ *is not* $full$ **do**
9: $p = P[cNNl[j].i]$ ▷ retrieve the point p as current NN of the Centroid c
10: **for** $k = 0; k < M; k++$ **do** $sumdist+ = dist(p, q[k])$ ▷ $\forall$ q, add dist to current point
11: $MaxKHeap.insert(p, sumdist)$
12: $j = j + 1$ ▷ increment index j to the next NN
13: **while** $j < N$ **do**
14: $p = P[cNNl[j].i]$ ▷ retrieve the point p as current NN of the Centroid c
15: $dpc = cNNl[j].dist$ ▷ retrieve the distance dist(p,c) from the list $cNNl$
16: **if** $M \cdot dpc \geq MaxKHeap.root.dist + sumdistCQ$ **then** ▷ termination condition
17: **break** ▷ exit j, all other NNs have larger sum of distances
18: **for** $k = 0; k < M; k++$ **do** $sumdist+ = dist(p, q[k])$ ▷ $\forall$ q, add dist to current point
19: **if** $sumdist < MaxKHeap.root.dist$ **then**
20: $MaxKHeap.insertFull(p, sumdist)$
21: $j = j + 1$ ▷ increment index j to the next NN

sequent NN to the centroid that is retrieved, all the NNs after the one that makes this condition true can be ignored. Thus, this condition can be not only one pruning condition, but termination condition of the process of $KGNNQ$. This termination condition is checked in the beginning of every iteration in this second part of the algorithm. Moreover, for a data point p retrieved from the $cNNl$ list and after the $sumdist(p, Q)$ has been calculated, this sum of distances will be compared with the δ value of the $MaxKHeap.root.dist$. There are 2 cases:

1. *Case 1*: If $sumdist(p, Q)$ is larger than or equal to δ, then p will be not inserted in the heap.
2. *Case 2*: If the $sumdist(p, Q)$ is smaller than δ, then p will be inserted in the heap, after $MaxKHeap.root$ has been deleted (**rule 1**).

The next two algorithms that we developped are novel Plane-Sweep algorithms that make use of the *median* point of the query set Q. A simple application of Plane-Sweep, assuming that both data sets are sorted in ascending order of their X-values, would compute the sum of distances of each data point to all the query points, by examining the data points from left to right, along the sweeping axis (e.g. X-axis). Let p with $sumdx(p, Q) \geq \delta$, then, for every p' with $p'.x \geq p.x$, $sumdx(p', Q) \geq sumdx(p, Q)$. Moreover, $sumdist(p', Q) \geq sumdx(p', Q)$. Thus, $sumdist(p', Q) \geq \delta$ and we do not need to calculate any distance for p'. Note that, while the sweep line approaches (moves away from) the median point(s), $sumdx$ will be decreasing (increasing). This is proved in the Appendix. In the next two algorithms, we find a data point $p_i \in P$ that is X-closest to the *median* point of the query set Q (in case that the query set contains an even number of points, we choose the right of the two median points). This data point is found by binary search. The sweep line is located on p_{i-1} and *moves to left* until a data point p with $sumdx(p, Q) \geq \delta$ is found (**termination condition 1**). Then, the sweep line is located on p_i and *moves to right* until a data point p with $sumdx(p, Q) \geq \delta$ (**termination condition 2**). At this stage, $MaxKHeap$ will contain the K data points of P with the smallest sum of distances to the query points.

As we mensioned above, in [7] it was proved that for every data point p with $|Q| \cdot dist(p, c) \geq \delta + sumdist(c, Q)$, p can be ignored, without calculating any distance. In the third algorithm that we have developed, the centroid c of the query points is also used and the above condition is a pruning condition for points that saves a significant number of calculations. Moreover, in the third algorithm, when the sweep line is outside of the area of query points, then for the current data point p, $sumdx(p, Q) = |Q| \cdot |p.x - c.x|$. Using this condition, we save numerous calculations.

In the Appendix, we prove that the sum of dx-distances between one given point $p(x, y) \in P$ and all points of the query set Q ($sumdx(p, Q)$):

A Is minimized at the median point $q[m]$ (where $q[m]$ is the array notation of q_m),
B For all $p.x \geq q[m].x$, $sumdx$ is constant or increasing with the increment of x, and
C For all $p.x < q[m].x$, $sumdx$ is increasing while x decreases.

The **second algorithm** (that is only based on *median*) is called $GNNPS$ and it uses the helper algorithm *calc_sum_dist* and the function $find\_closest\_point$. Firstly, it calculates the initial position of the sweeping line (preparation state). For this, the algorithm must find the first point $p[i] \in P$ which is on the right of the median of query set $q[m]$ ($p[i].x > q[m].x$), by calling the function $find\_closest\_point$ (line 1). After this, the algorithm sets the sweeping line at the point $p[i-1]$ (line 3) and continues scanning the points of P set decreasing the index i until the *termination condition 1* will be true or the points of P set will have finished (lines 3–5). Lastly, the algorithm sets the sweeping line at the point $p[i]$ and continues scanning the points of P set

Algorithm 2. GNNPS.

Input: Two X-sorted arrays of points $p[0, 1, \cdots, N-1]$, $q[0, 1, \cdots, M-1]$, and $MaxKHeap$.

Output: $MaxKHeap$ storing the K NNs having smallest sums of distances to all query points.

1: $i = find\_closest\_point(0, P, q[m])$ ▷ STEP 1 : Preperation. $q[m]$ is the median query set Q
2: $j = i - 1$
3: **while** $j > -1$ **do** ▷ STEP 2 : Search in the range $p[j].x \leq q[m].x$, descending j (move to left)
4: **if** $calc\_sum\_dist(p[j--], Q, MaxKHeap) == err\_code\_dx$ **then** ▷ Termination cond. 1
5: **break**
6: **while** $i < N$ **do** ▷ STEP 3 : Search in the range $p[i].x > q[m].x$, ascending i (move to right)
7: **if** $calc\_sum\_dist(p[i++], Q, MaxKHeap) == err\_code\_dx$ **then** ▷ Termination cond. 2
8: **break**

increasing the index i until the *termination condition 2* will be true or the points of the P set will have finished (lines 6–8).

The **third algorithm** (that is based on *median* and *centroid*) is called $GNNPSC$ and it uses the helper algorithms *calc_sum_dist_in* and *calc_sum_dist_out* and the function $find\_closest\_point$. Firstly, the algorithm calculates the initial position of the sweeping line and the coordinates of the centroid (preparation state). For these, the algorithm calls the functions $find\_closest\_point$ (line 1) and $Calculate\_Centroid\_Coord(Q)$ (line 3). In the next step, it continues scanning the points of P set decreasing the index j until the *termination condition 1* will be true or the X-coordinate of the current point of P set is smaller than or equal to the X-coordinate of the first query point $q[0]$ ($p[j].x \leq q[0]$). In this state, $GNNPSC$ calls the function *calc_sum_dist_in* to calculate the sum of distances. After exiting the previous loop and if the *termination condition 1* has not arisen (line 12), the algorithm continues decreasing j until the *termination condition 1* will be true or the points of P set will have finished (lines 13–15). Lastly, the algorithm sets the sweeping line at the point $p[i]$ and continues scanning the points of P set increasing the index i just like in the previous step (lines 17–20 inside query set Q and lines 21–24 outside query set Q). Note that the *calc_sum_dist_in* function is the same as *calc_sum_dist*, adding two new parameters (the centroid of Q (c) and its sum of distances to all query points ($sumdistCQ$)) and the following statements just after the line 9.

9 : $distpc = dist(p, c)$
10 : **if** $M \cdot distpc \geq MaxKHeap.root.dist + sumdistCQ$ **then**
11 : **return** $err\_code\_dist\_centroid$

And the remaining statements of *calc_sum_dist_in* from line 12 (12–22) are the same as *calc_sum_dist*.

Algorithm 3. *calc_sum_dist.*

Input: One point p, the sorted array of query points $q[0, 1, \cdots, M-1]$, and $MaxKHeap$.

Output: Value $successful\_insertion$ or err_code_dx or err_code_dist and $MaxKHeap$ updated with p if rule 2 was true.

```
1: function calc_sum_dist(p, Q, MaxKHeap)
2:     sumdist = 0.0, sumdx = 0.0
3:     if MaxKHeap is not full then
4:         for k = 0; k < M; k++ do                          ▷ for each query point q
5:             sumdist+ = dist(p, q[k])      ▷ dist(): the Euclidean distance between p and q[k]
6:         MaxKHeap.insert(p, sumdist)
7:         return sucessful_insertion
8:     else
9:         for k = 0; k < M; k++ do                          ▷ for each query point q
10:            sumdx+ = dx_dist(p, q[k])   ▷ dx_dist(): the dx-distance between p and q[k]
11:        if sumdx ≥ MaxKHeap.root.dist then                              ▷ Rule 1
12:            return err_code_dx          ▷ exit k, all other points have longer distance
13:        for k = 0; k < M; k++ do                          ▷ for each query point q
14:            sumdist+ = dist(p, q[k])    ▷ add the distance (dist) from the current point
15:        if sumdist < MaxKHeap.root.dist then                            ▷ Rule 2
16:            MaxKHeap.insertFull(p, sumdist)
17:            return sucessful_insertion
18:        else
19:            return err_code_dist        ▷ not inserted because of sum of distances (sumdist)
```

The following examples illustrate the execution of the algorithms. The point data P set is defined as $P = \{p_0(1,7);\ p_1(2,4);\ p_2(3,1);\ p_3(3,13);\ p_4(8,2);\ p_5(8,18);\ p_6(9,10);\ p_7(10,19);\ p_8(12,12);\ p_9(13,4);\ p_{10}(14,12);\ p_{11}(16,6);\ p_{12}(19,8);\ p_{13}(19,17);\ p_{14}(20,3);\ p_{15}(22,7)\}$, and the point query set Q is defined as $Q = \{q_0(9,7);\ q_1(10,11);\ q_2(12,4);\ q_3(17,7);\ q_4(19,11)\}$. In Fig. 1, P and Q (they are sorted in ascending order of their X-values), the centroid and the median of the query points and the initial position of the sweep line are drawn.

$SPMNI$ starts the preparation stage by calculating the coordinates of centroid point $c(x, y) = (13.4, 8)$ and then calculates the sum of distances between the centroid and the query points $sumdist(c, Q) = 23.374$. Next, $SPMNI$ creates the $cNNl$ list of pairs of type $< i, dist(p, c) >$ for all points of P set. This list is sorted in ascending order for each point in the P set with respect to $dist(p, c)$. Thus, the final form of the sorted list is: $\{< 11, 3.280 >, < 9, 4.020 >, < 10, 4.045 >, < 8, 4.238 >, < 6, 4.833 >, < 12, 5.600 >, < 4, 8.072 >, < 14, 8.280 >, < 15, 8.658 >, < 13, 10.600 >, < 5, 11.365 >, < 7, 11.513 >, < 3, 11.539 >, < 1, 12.081 >, < 0, 12.440 >, < 2, 12.536 >\}$. In other words, we have one complete list of Nearest Neighbors to the Centroid beginning from the closest one.

Algorithm 4. GNNPSC .

Input: Two X-sorted arrays of points $p[0, 1, \cdots, N-1]$, $q[0, 1, \cdots, M-1]$, and $MaxKHeap$.

Output: $MaxKHeap$ storing the K NNs having smallest sums of distances to all query points.

1: $i = find\_closest\_point(0, P, q[m])$ ▷ STEP 1 : Preperation. $q[m]$ is the median of query set Q
2: $j = i - 1$
3: $c(x, y) = Calculate\_Centroid\_Coord(Q)$ ▷ calculate the coordinates of the Centroid
4: $sumdistCQ = 0.0$
5: **for** $k = 0; k < M; k++$ **do** ▷ for each query point q
6: $sumdistCQ+ = dist(c, q[k])$
▷ STEP 2 : Search in the range $p[j].x \leq q[m].x$, descending j (move to left)
7: $cont\_search = true$ ▷ initialize the flag
8: **while** $j > -1$ **and** $p[j].x > q[0].x$ **do** ▷ $\forall$ $p[j]$ inside the query MBR in sweeping axis
9: **if** $calc\_sum\_dist\_in(p[j--], Q, c, sumdistCQ, MaxKHeap) == err\_code\_dx$ **then**
10: ▷ Termination condition 1
11: $cont\_search = false$
12: **break**
13: **if** $cont\_search = true$ **then**
14: **while** $j > -1$ **do** ▷ for each point $p[j]$ on the left of the query MBR in sweeping axis
15: **if** $calc\_sum\_dist\_out(p[j--], Q, c, sumdistCQ, MaxKHeap) == err\_code\_dx$ **then**
16: ▷ Termination condition 1
17: **break**
▷ STEP 3 : Search in the range $p[i].x > q[m].x$, ascending i (move to right)
18: $cont\_search = true$
19: **while** $i < N$ **and** $p[i].x < q[M-1].x$ **do** ▷ $\forall$ $p[i]$ inside the query MBR in sweeping axis
20: **if** $calc\_sum\_dist\_in(p[i++], Q, c, sumdistCQ, MaxKHeap) == err\_code\_dx$ **then**
21: ▷ Termination condition 2
22: $cont\_search = false$
23: **break**
24: **if** $cont\_search = true$ **then**
25: **while** $i < N$ **do** ▷ for each point $p[i]$ on the left of the query MBR in sweeping axis
26: **if** $calc\_sum\_dist\_out(p[i++], Q, c, sumdistCQ, MaxKHeap) == err\_code\_dx$ **then**
27: ▷ Termination condition 2
28: **break**

Index j is initially set to 0, that is, to the first element of the $cNNl$ list and $SPMNI$ continues with main stage. The search process of the $KGNNQ$ starts

Algorithm 5. *calc_sum_dist_in.*

Input: One point p, set of query points Q, centroid c, its sum of distances to all query points $sumdistCQ$ and $MaxKHeap$.

Output: Value $successful\_insertion$ or err_code_dx or err_code_dist and $MaxKHeap$ updated with p if rule 2 was true.

1: **function** *calc_sum_dist_in*(p, Q, c, $sumdistCQ$, $MaxKHeap$)
2: $sumdist = 0.0$, $sumdx = 0.0$
3: **if** $MaxKHeap$ *is not full* **then**
4: **for** $k = 0; k < M; k++$ **do** ▷ for each query point q
5: $sumdist += dist(p, q[k])$ ▷ $dist()$: the dx-distance between p and $q[k]$
6: $MaxKHeap.insert(p, sumdist)$
7: **return** $sucessful\_insertion$
8: **else**
9: $dpc = dist(p, c)$ ▷ $dist()$: the distance between p and c
10: **if** $M \cdot dpc \geq MaxKHeap.root.dist + sumdistCQ$ **then** ▷ prune p w/o computing dists
11: **return** $err\_code\_dist\_centroid$; ▷ not inserted because of sum of distances
12: **for** $k = 0; k < M; k++$ **do** ▷ for each query point q
13: $sumdx += dx\_dist(p, q[k])$ ▷ $dx\_dist()$: the dx-distance between p and $q[k]$
14: **if** $sumdx \geq MaxKHeap.root.dist$ **then** ▷ Rule 1
15: **return** err_code_dx ▷ exit k, all other points have longer distance
16: **for** $k = 0; k < M; k++$ **do** ▷ for each query point q
17: $sumdist += dist(p, q[k])$ ▷ add the distance ($dist$) from the current point
18: **if** $sumdist < MaxKHeap.root.dist$ **then** ▷ Rule 2
19: $MaxKHeap.insertFull(p, sumdist)$
20: **return** $sucessful\_insertion$
21: **else**
22: **return** err_code_dist ▷ not inserted because of sum of distances ($sumdist$)

with empty $MaxKHeap$. The first NN is p_{11}. $sumdist(p_{11}, Q) = 26.599$ is calculated and the pair $< p_{11}, 26.599 >$ is inserted in the $MaxKHeap$ as the first one (lines 8–11). Index j is incremented and the second and third NN are retreived sequentially from the $cNNl$ list, while $MaxKHeap$ is not full. The pairs $< p_9, 27.835 >$ and $< p_{10}, 30.370 >$ are inserted in the $MaxKHeap$. At the end of the rhird iteration the $MaxKHeap$ is full and $MaxKHeap.root.dist$ has value 30.370. The second part of the main stage is started executing lines 13–21. The forth NN is p_8 with $dist(p_8, c) = 4.238$. The terminal condition $|Q| \cdot dist(p, c) \geq \delta + sumdist(c, Q)$ is tested (line 16). The condition $5 \cdot 4.238 \geq 30.370 + 23.374$ is false. Therefore, $SPMNI$ continues calculating $sumdist(p_8, Q)$ (line 18). The variable $sumdist$ is set to the value 30.209. The condition $sumdist(p_8, Q) < MaxKHeap.root.dist$ is true and the previous $MaxKHea.root$ is deleted, because the pair $< p_8, 30.209 >$ must be inserted in

Algorithm 6. *calc_sum_dist_out.*

Input: One point p, set of query points Q, centroid c, its sum of distances to all query points $sumdistCQ$ and $MaxKHeap$.

Output: Value $successful\_insertion$ or err_code_dx or err_code_dist and $MaxKHeap$ updated with p if rule 2 was true.

```
1: function calc_sum_dist_out(p, Q, c, sumdistCQ, MaxKHeap)
2:     sumdist = 0.0, sumdx = 0.0
3:     if MaxKHeap is not full then
4:         for k = 0; k < M; k++ do                          ▷ for each query point q
5:             sumdist+ = dist(p, q[k])    ▷ dist(): the dx-distance between p and q[k]
6:         MaxKHeap.insert(p, sumdist)
7:         return sucessful_insertion
8:     else
9:         dx = dx_dist(p, c)   ▷ dx_dist(): the dx-distance between p and c (Δx(p, c))
10:        if M · dx ≥ MaxKHeap.root.dist then                          ▷ Rule 1
11:            return err_code_dx;       ▷ exit k, all other points have longer distance
12:        dy = dy_dist(p, c)              ▷ dy_dist(): the dy-distance between p and c
13:        distpc = √(dx² + dy²)
14:        if M · distpc ≥ MaxKHeap.root.dist + sumdistCQ then
15:            return err_code_dist_centroid;
16:        for k = 0; k < M; k++ do                          ▷ for each query point q
17:            sumdist+ = dist(p, q[k])
18:        if sumdist < MaxKHeap.root.dist then                         ▷ Rule 2
19:            MaxKHeap.insertFull(p, sumdist)
20:            return sucessful_insertion
21:        else
22:            return err_code_dist          ▷ not inserted because of sum of distances
   (sumdist)
```

the full $MaxKHeap$ as new root (line 20). So, $MaxKHeap.root.dist = 30.209$ and the index j is incremented (line 21). The fifth NN is p_6 with $dist(p_6, c) = 4.833$. The terminal condition is tested as above (line 16). The condition $5 \cdot 4.833 \geq 30.209 + 23.374$ is false. Therefore, $SPMNI$ continues calculating the $sumdist(p_6, Q)$ (line 18). The variable $sumdist$ is set to the value 29.716. The condition $sumdist(p_6, Q) < MaxKHeap.root.dist$ is true and the previous $MaxKHea.root$ is deleted, because the pair $< p_6, 29.716 >$ must be inserted in the full $MaxKHeap$. The pair $< p_9, 27.835 >$ becomes new root (line 20). So, $MaxKHeap.root.dist = 27.835$. From the sixth to the tenth NNs (p_{12} with $dist(p_{12}, c) = 5.6$, p_4 with $dist(p_4, c) = 8.072$. p_{14} with $dist(p_{14}, c) = 8.280$, p_{15} with $dist(p_{15}, c) = 8.658$ and p_{13} with $dist(p_{13}, c) = 10.6$), the terminal condition is false (line 16). Therefore, $SPMNI$ continues calculating the variable $sumdist$ for each point (line 18). The condition $sumdist(p, Q) < MaxKHeap.root.dist$ is false and the pairs $\{< p_{12}, 32.835 >, < p_4, 43.299 >, < p_{14}, 45.635 >, < p_{15}, 46.089 >, < p_{13}, 55.922 >\}$ must not be inserted in the full $MaxKHeap$. The eleventh and final NN is p_5 with $dist(p_5, c) = 11.365$. The terminal condition is tested as above (line 16) and $5 \cdot 11.365 \geq 29.716 + 23.374$ is true. Therefore,

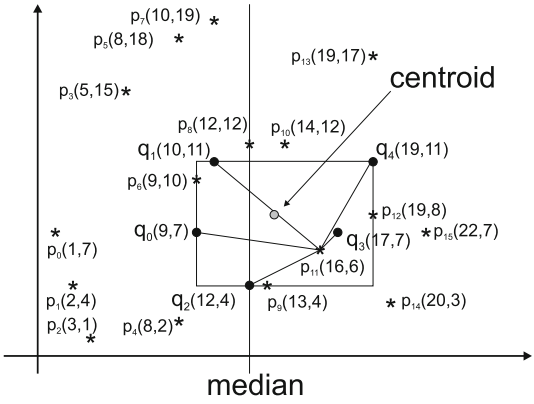


Fig. 1. The points of P and Q, the centroid, the median of the query points and the initial position of the sweep line.

$SPMNI$ is terminated (line 17) by breaking the while ... do loop. While executing this algorithm, we made 71 complete point-point distance calculations, 142 point-point dx-distance calculations, 5 points with their sum of distances were inserted in the $MaxKHeap$ and 10 of 16 points of P set were fully examined and their sumdist distances to the query points were calculated. One point of 16 (the last one) has been partially examined. $dist(p, c)$ has been calculated for all (16) points and all the P set members have been sorted.

In $GNNPS$, firstly (in Step 1) the algorithm searches for the point of the P set which is on the right of the median q_2(12,4) query point (line 1). That is p_9(13,4) point. In Step 2 (lines 3–5) it starts calculating the sum of distances between point p_8(12,12) and all query points. The result is $sumdist(p_8, Q) =$ 30.209 and the point p_8 is inserted in the $MaxKHeap$ (*calc_sum_dist*:lines 2–7). In the next iteration the point p_7(10,19) is examined. The $MaxKHeap$ is full and the second part of the *calc_sum_dist* function (lines 9–19) is executed. The sum of distances is $sumdist(p_7, Q) = 61.108$ larger than the $MaxKHeap.root.dist =$ 30.209 (condition in the *calc_sum_dist*:line 15 is false), so the point is rejected (*calc_sum_dist*:line 19). In the third iteration the point p_6(9,10) is examined and the sum of distances is $sumdist(p_6, Q) = 29.716$ which is smaller (condition of *calc_sum_dist*:line 15 is true) than the $MaxKHeap.root.dist$ therefore the point p_6 is inserted in the $MaxKHeap$ (*calc_sum_dist*:lines 16,17) by replacing the previous root (p_8). In the fourth and fifth iterations for the points p_5 and p_4 the sum of distances are $sumdist(p_5, Q) = 60.317$ and $sumdist(p_4, Q) = 43.299$, respectively; both larger than the $MaxKHeap.root.dist$ and the points are rejected. In the sixth iteration, the point p_3 has $sumdx(p_3.x, Q) = 52$ (condition in *calc_sum_dist*:line 11) which is larger than the $MaxKHeap.root.dist$ and the process (scanning the P set on the left) ends (*calc_sum_dist*:line 12) because it is impossible to find other points of P set on the left of p_3 having sum of distances smaller than 52. The algorithm continues scanning the

points of P set to the right of the median q_2, starting from the p_9 point. Its $sumdist(p_9, Q) = 27.835$ is smaller than the $MaxKHeap.root.dist = 29.716$ so it replaces the existing point in the root of $MaxKHeap$. The next point p_{10} has $sumdist(p_{10}, Q) = 30.370$ and it is rejected. The next iteration will try the point p_{11} which has $sumdist(p_{11}, Q) = 26.599$ the smallest sum of distances and this point (p_{11}) is inserted in the $MaxKHeap$ replacing the previous root p_9. In the last iteration the algorithm examines the point p_{12} which has $sumdx(p_{12}, Q) = 28$ larger than the $MaxKHeap.root.dist = 26.599$ and the process is finally finished. While executing this algorithm we made 46 complete point-point distance calculations, 84 point-point dx-distance calculations, 4 points with their sum of distances were inserted in the $MaxKHeap$ and 10 of the 16 points of P set were examined.

$GNNPSC$ starts (Step 1) by finding the first point of P set which is on the right of the median point of query set Q. That is the point p_9. Afterwards it calculates the coordinates of centroid point $c(x, y) = (13, 8)$ and then calculates the sum of distances between the centroid and the query points $sumdist(c, Q) = 23.374$. $GNNPSC$ continues with Step 2. In that step, the points of P set are scanned on the left of the p_9 in two particular steps. First from p_8 up to p_7 which have X-coordinate larger than $q_0.x = 9$ by calling the *calc_sum_dist_in* function. There is $sumdist(p_8, Q) = 30.209$ and this point is inserted in the $MaxKHeap$ as the first point while the $MaxKHeap$ is empty (*calc_sum_dist_in*:lines 3–7). The point p_7 is examined next and it is rejected without a need to calculate $sumdist(p_7, Q)$ because the condition of the function *calc_sum_dist_in*:line 10 is true. Step 2 continues scanning the points of P set which are on the left (outside) of the q_0 query point by calling the function *calc_sum_dist_out*. The point p_6 with $sumdist(p_6, Q) = 29.716$ is inserted (*calc_sum_dist_in*:lines 9–20), while points p_5 and p_4 are rejected with $sumdist(p_5, Q) = 60.137$ and $sumdist(p_4, Q) = 43.299$ respectively, both larger than the $MaxKHeap.root.dist = 29.716$ with the point p_6. The next point p_3 is the last point to be examined because it has $sumdx(p_3, Q) = 52$ larger than the current $MaxKHeap.root.dist$. The algorithm continues by executing Step 3, scanning the points of P set on the right of the median query point q_2. The algorithm continues scanning the points of P set to the right starting from the p_9 point. Its $sumdist(p_9, Q) = 27.835$ is smaller than the $MaxKHeap.root.dist = 29.716$ so it replaces the existing point in the root of $MaxKHeap$. The next point p_{10} has $sumdist(p_{10}, Q) = 30.370$ and it is rejected. The next iteration will try the point p_{11} which has $sumdist(p_{11}, Q) = 26.599$ the smallest sum of distances and this point is inserted in the $MaxKHeap$ replacing the previous root p_9. In the last iteration we examine the point p_{12} which has $sumdx(p_{12}, Q) = 28$ larger than the $MaxKHeap.root.dist = 26.599$ and the process is finally finished. While executing this algorithm we made 42 complete point-point distance calculations, 38 point-point dx-distance calculations, 4 points with their sum of distances were inserted in the $MaxKHeap$ and 10 of 16 points of P set were examined.

4 Experimentation

In order to evaluate the behaviour of the proposed algorithms, we have used 6 real spatial data sets of North America, representing cultural landmarks (CL with 9203 points) and populated places (PP with 24493 points), roads (RD with 569120 line-segments) and railroads (RR with 191637 line-segments). To create sets of points, we have transformed the MBRs of line-segments from RD and RR into points by taking the center of each MBR (i.e., $|RD| = 569120$ points, $|RR| = 191637$ points). Moreover, in order to get the double amount of points from RR and RD, we chose the two points with min and max coordinates of the MBR of each line-segment (i.e. $|RDD| = 1138240$ points and $|RRD| = 383274$ points). The data of these 6 files were normalized in the range $[0, 1]^2$. The real data sets we used are geographical. In order to test the performance of our algorithms with data appearing in Science, we have created synthetic clustered data sets of 125000 ($125K$), 250000 ($250K$), 500000 ($500K$) and 1000000 ($1000K$) points, with 125 clusters in each data set (uniformly distributed in the range $[0, 1]^2$), where for a set having N points, $N/125$ points were gathered around the center of each cluster, according to Gaussian distribution (this distibution is common for natural properties of systems within Science). The first real data set (CL) was used to make the query set (Q) by selecting the appropriate number of points randomly. Then the coordinates of these points were appropriately scaled in order to get the MBR of the query points to get a pre-defined size in comparison to the MBR of the data set (P). The other 9 data sets were used as data sets (P) within which we were looking for the NNs.

All experiments were performed on a Laptop PC with Intel Core i5-3210M (2.5 GHz) CPU with 4 GB of RAM and several GBs of secondary storage, with Ubuntu Linux v. 14.04 64 bit, using the GNU C/C++ compiler (gcc). The performance measurements were: (1) the response time (total query execution time) of processing the (K) GNNQ, not counting reading from disk files to main memory and sorting of the data sets, (2) the number of points involved in calculations, (3) the number of X-axis distance computations (dx-distance) and (4) the number of distance computations.

In every experiment the query set was moved on X-axis in 8 equal size steps from the top left corner of the area of the data set (P) up to the right corner and after this, one step down on the Y-axis and so on. The total execution time, and the other experimentation metrics, for each one experiment, were computed as an average of all (the 64) queries.

In Fig. 2 (left), we depict the effect of the number of query points, M (the cardinality of Q), on execution time of all algorithms for the RD data set (the number of group nearest-neighbors, K, was equal to 8 and the size of query-set MBR was 8 % of the data set space). In analogous diagrams created for dx-distance and $dist$ calculations, in most cases, $GNNPS$ had the worse performance, $SPMNI$ was next and $GNNPSC$ had the best performance. It is obvious that the increase of M leads to an increase of the execution time, but with a smaller rate of increase. $GNNPSC$ needs less time than $GNNPS$, because of the use of centroid (the computation of the distance between the centroid and the

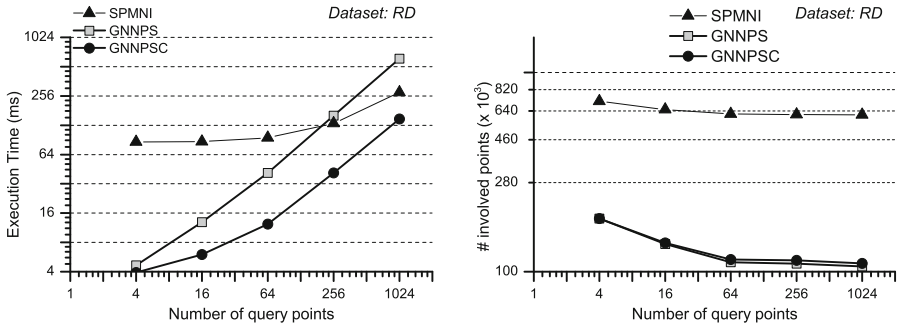


Fig. 2. (Left) Execution time of the algorithms as a function of M (RD data set). # (Right) Points involved in *sumdist* calculations of the algorithms as a function of M (RD data set).

reference point of P set needs one calculation of distance while the computation of the sum of distances between the reference point and all query points needs M distance calculations). Moreover, $GNNPSC$ needs less time than $SPMNI$, although both algorithms make use of centroid, because $SPMNI$ intially calculates the distance of the whole P set from the centroid and sorts the whole P set, while $GNNPSC$ does not access a big part of the P set, due to the termination condition.

For the same parameter settings and data set, in Fig. 2 (right), we depict the effect of M on the number of data set points involved in calculations. We observe that this number of points is reduced as M increases. In $SPMNI$, all points of P set are involved in calculations, since this algorithms intially calculates the distance of the whole P set from the centroid and sorts the whole P set. Regarding the other two algorithms, note that the sums of distances of the points of data P set near the median are enlarged to a smaller extent, compared to the *sumdist* of the points outside the query-set MBR. This enables the termination conditions and makes it possible to get nearest to the median query point. Moreover, we can observe in Fig. 2 that $GNNPSC$ needs more involved points and it is the fastest. This behaviour could be due to that in function *calc_sum_dist_in* we firstly apply the pruning condition of centroid and next the termination condition 1 or 2 is checked. So it is possible that some points may be pruned in $GNNPSC$ rather than being the cause of termination of the scanning.

In Fig. 3 (left), we depict the effect of the size of the query-set MBR, on dx-distance calculations of all algorithms for the $1000K$ data set (the number of group nearest neighbors, K, was equal to 8 and the number of query points was equal to 128).

Analogous diagrams created for distance calculations had similar appearance. In most of these diagrams, $SPMNI$ had the worse execution time, while $GNNPSC$ was always the best. It is obvious that the increase of the size of the query-set MBR leads to an increase of the execution time, but with a smaller

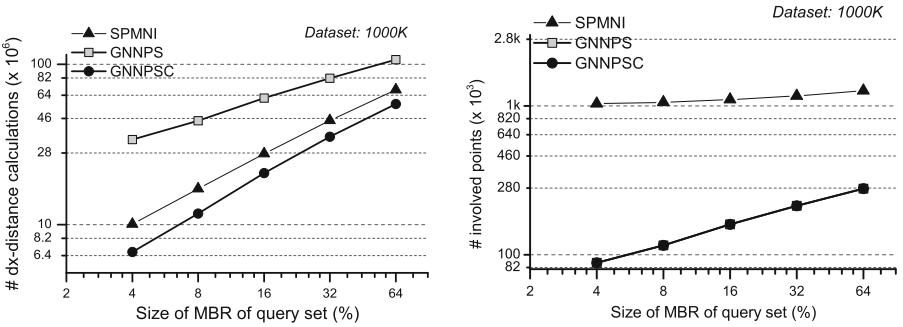


Fig. 3. (Left) # dx-distance calculations of the algorithms as a function of the size of MBR $\mathcal{M}$ ($1000K$ data set). (Right) # Points involved in calculations of the algorithms as a function of the size of MBR $\mathcal{M}$ ($1000K$ data set).

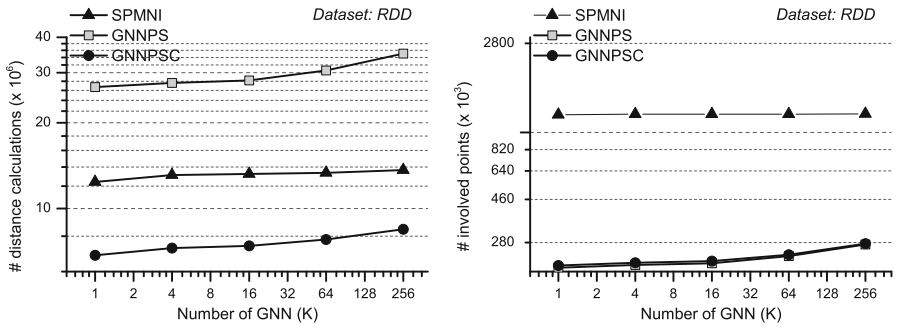


Fig. 4. (Left) # distance calculations of the algorithms as a function of K (RDD data set). (Right) # Points involved in calculations of the algorithms as a function of K (RDD data set).

rate of increase. The size of MBR was increased with a ratio of 4. The execution time, dx-distance and complete distance ($dist$) calculations was increased with ratio in the range 1.2 up to 2 for all data sets of real and synthetic data. For the same parameter settings and data set, in Fig. 3 (right), we depict the effect of the size of the query-set MBR on the number of points involved in calculations. We observe that this numbers of points are increased as the query-set MBR increases with a ratio smaller than 1.4 for $GNNPS$ and $GNNPSC$. We observe in this figure that the number of points involved almost identical and the lines are for $GNNPS$ and $GNNPSC$ overlapped. In $SPMNI$, the number of points involved in calculations is much higher, as explained in the interpretation of Fig. 2 (right).

In Fig. 4 (left), we depict the effect of the number of group nearest-neighbors, K, on distance calculations of all algorithms for the RDD data set (the number of query points, M, was equal to 128 and the size of query-set MBR was 8 % of the data set space). Analogous diagrams created for distance calculations had

similar appearance. Regarding execution time, the comments of Fig. 3 (left) hold for this figure, too. For the same parameter settings and data set, in Fig. 4 (right), we depict the effect of K on the number of points involved in calculations. We observe that this number of points is increased so slowly that it is going to be seen for values of K larger than 64.

From the above experiments, we conclude that:

- The number of data-set points involved in the calculations of $GNNPS$ and $GNNPSC$ algorithms is almost equal. However, the execution time for $GNNPSC$ remains always lower than the execution time of $GNNPS$, due to the pruning condition and the lower dx-distance calculations cost. This number is always signifficantly larger for $SPMNI$, since this algorithms intially calculates the distance of the whole P set from the centroid and sorts the whole P set.
- The main advantages of the Plane-Sweep method are the absence of recalculation, as each point is used in calculations once at most, and the absence of backtracking.
- The number of points involved in calculations is decreased when the number of query points is increased, provided that K and the query-set MBR size remain constant.

5 Conclusions and Future Work

Processing of GNNQs has been based on index structures, so far. In this paper, for the first time, we present new algorithms that can be efficiently applied on RAM-based data for processing the GNNQ. Extending [6], we present a comparison of new PS algorithms that we developed with respect to the best algorithm presented in [7] that can be transformed to work on non-indexed data sets, and we observe the PS algorithms achieve signifficantly better performance. As the experimentation that we performed shows, using synthetic and real data sets, the use of median and centroid in $GNNPSC$, prunes the number of points involved in processing and the number of calculations, in relation to $SPMNI$ and $GNNPS$.

In the future, we plan to compare the best of our algorithms to existing index based solutions. Moreover, the algorithms we present could be transformed/extended to work on high volume, disk resident data that are transferred in RAM in blocks. Additionally, the application of Plane-Sweep to other spatial queries (like Reverse NNQ) could lead to interesting techniques.

Appendix

Lemma: The sum of dx-distances between one given point $p(x, y) \in P$ and all points of the query set Q ($sumdx(p, Q)$):

A Is minimized at the median point $q[m]$ (where $q[m]$ is the array notation of q_m),

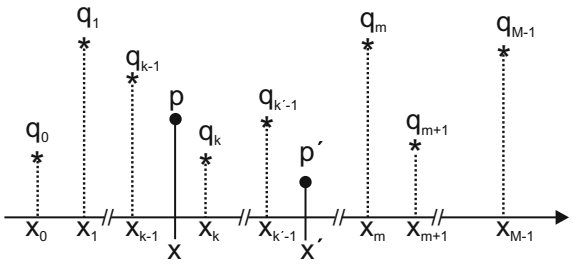


Fig. 5. The point p has K query points on the left and the point p' ($p'.x > p.x$) has K' query points on the left.

B For all $p.x \geq q[m].x$, *sumdx* is constant or increasing with the increment of x, and
C For all $p.x < q[m].x$, *sumdx* is increasing while x decreases.

Proof: Property **A** has been proved in [18]. To prove property **B**, for every point $p \in P$ and $q \in Q$, we use $\Delta x(p,q) = \begin{cases} p.x - q.x \text{if } p.x \geq q.x \\ q.x - p.x \text{if } p.x < q.x \end{cases}$
If the point p has K query points on the left ($p.x < q[K-1].x$) and $M-K$ query points on the right (Fig. 5), then: $sumdx(p,Q) = \sum_{i=0}^{K-1}(p.x - q[i].x) + \sum_{i=K}^{M-1}(q[i].x - p.x) = Kp.x - \sum_{i=0}^{K-1} q[i].x + \sum_{i=K}^{M-1} q[i] - (M-K)p.x = (2K-M)p.x - \sum_{i=0}^{K-1} q[i].x + \sum_{i=K}^{M-1} q[i].x$
For another point $p' \in P$ with $p'.x > p.x$ which has K' query points on the left (Fig. 5) and $M - K'$ query points on the right, it is: $sumdx(p',Q) = (2K' - M)p'.x - \sum_{i=0}^{K'-1} q[i].x + \sum_{i=K'}^{M-1} q[i].x$
The difference between dx-distances of the points p' and p is: $\Delta sumdx = sumdx(p',Q) - sumdx(p,Q) = (2K - M)(p'.x - p.x) + 2\left[(K'-K)p'.x - \sum_{i=K}^{K'-1} q[i].x\right]$. If the set of the query points Q has cardinality M and this is an even number then there are two medians $q[m1]$ and $q[m2]$, while if M is odd then there is only one median point $q[m]$.
B.1 M is even and $q[m1].x \leq p.x < p'.x$ then $M \leq 2K \leq 2K'$ so $(2K-M) \geq 0$, $(p'.x - p.x) \geq 0$ and $(K'-K)p'.x - \sum_{i=K}^{K'-1} q[i].x \geq 0$ because $p'.x \geq q[i].x$, whereas $K \leq i \leq K'$

B.2 All of the above apply to M if it is odd and it is only one median point $q[m].x \leq p.x < p'.x$. It is proven that for all points p on the right of the median query point the sum of dx-distances is increasing.
C For both types of cardinality of the query set Q and for the case $p.x < p'.x < q[m].x$ it is: $\Delta sumdx = (2K - M)(p'.x - p.x) + 2(K' - K)p'.x - 2\sum_{i=K}^{K'-1} q[i].x \leq (2K - M)(p'.x - p.x) + 2(K' - K)p'.x - 2(K' - K)p.x = 2(K - M)(p'.x - p.x) + 2(K' - K)(p'.x - p.x) = (2K - M + 2K' - 2K)(p'.x - p.x) = (2K' - M)(p'.x - p.x) < 0$. It is proven that for all points p on the left of the median query point the sum of dx-distances is strictly decreasing. □

References

1. Rigaux, P., Scholl, M., Voisard, A.: Spatial Databases - with Applications to GIS. Elsevier, San Francisco (2002)
2. Preparata, F.P., Shamos, M.I.: Computational Geometry - An Introduction. Springer, New York (1985)
3. Hinrichs, K., Nievergelt, J., Schorn, P.: Plane-sweep solves the closest pair problem elegantly. Inf. Process. Lett. **26**, 255–261 (1988)
4. Jacox, E.H., Samet, H.: Spatial join techniques. ACM Trans. Database Syst. **32**, 7 (2007)
5. Roumelis, G., Vassilakopoulos, M., Corral, A., Manolopoulos, Y.: A new plane-sweep algorithm for the K-closest-pairs query. In: Geffert, V., Preneel, B., Rovan, B., Štuller, J., Tjoa, A.M. (eds.) SOFSEM 2014. LNCS, vol. 8327, pp. 478–490. Springer, Heidelberg (2014)
6. Roumelis, G., Vassilakopoulos, M., Corral, A., Manolopoulos, Y.: Plane-sweep algorithms for the k group nearest-neighbor query. In: GISTAM Conference, pp. 83–93. Scitepress (2015)
7. Papadias, D., Shen, Q., Tao, Y., Mouratidis, K.: Group nearest neighbor queries. In: ICDE Conference, pp. 301–312. IEEE (2004)
8. Papadias, D., Tao, Y., Mouratidis, K., Hui, C.K.: Aggregate nearest neighbor queries in spatial databases. ACM Trans. Database Syst. **30**, 529–576 (2005)
9. Li, H., Lu, H., Huang, B., Huang, Z.: Two ellipse-based pruning methods for group nearest neighbor queries. In: ACM-GIS Conference, pp. 192–199. ACM (2005)
10. Luo, Y., Chen, H., Furuse, K., Ohbo, N.: Efficient methods in finding aggregate nearest neighbor by projection-based filtering. In: Gervasi, O., Gavrilova, M.L. (eds.) ICCSA 2007, Part III. LNCS, vol. 4707, pp. 821–833. Springer, Heidelberg (2007)
11. Namnandorj, S., Chen, H., Furuse, K., Ohbo, N.: Efficient bounds in finding aggregate nearest neighbors. In: Bhowmick, S.S., Küng, J., Wagner, R. (eds.) DEXA 2008. LNCS, vol. 5181, pp. 693–700. Springer, Heidelberg (2008)
12. Hashem, T., Kulik, L., Zhang, R.: Privacy preserving group nearest neighbor queries. In: EDBT Conference, pp. 489–500. ACM (2010)
13. Zhu, L., Jing, Y., Sun, W., Mao, D., Liu, P.: Voronoi-based aggregate nearest neighbor query processing in road networks. In: ACM-GIS Conference, pp. 518–521. ACM (2010)

14. Jiang, T., Gao, Y., Zhang, B., Liu, Q., Chen, L.: Reverse top-k group nearest neighbor search. In: Wang, J., Xiong, H., Ishikawa, Y., Xu, J., Zhou, J. (eds.) WAIM 2013. LNCS, vol. 7923, pp. 429–439. Springer, Heidelberg (2013)
15. Zhang, D., Chan, C., Tan, K.: Nearest group queries. In: SSDBM Conference, p. 7. ACM (2013)
16. Lian, X., Chen, L.: Probabilistic group nearest neighbor queries in uncertain databases. IEEE Trans. Knowl. Data Eng. **20**, 809–824 (2008)
17. Li, J., Wang, B., Wang, G., Bi, X.: Efficient processing of probabilistic group nearest neighbor query on uncertain data. In: Bhowmick, S.S., Dyreson, C.E., Jensen, C.S., Lee, M.L., Muliantara, A., Thalheim, B. (eds.) DASFAA 2014, Part I. LNCS, vol. 8421, pp. 436–450. Springer, Heidelberg (2014)
18. Ahn, H.-K., Bae, S.W., Son, W.: Group nearest neighbor queries in the L_1 plane. In: Chan, T.-H.H., Lau, L.C., Trevisan, L. (eds.) TAMC 2013. LNCS, vol. 7876, pp. 52–61. Springer, Heidelberg (2013)

Validation and Integration of Wheat Seed Emergence Prediction Model with GIS and Numerical Weather Prediction Models

R. Al-Habsi[1], Y.A. Al-Mulla[1(✉)], Y. Charabi[2], H. Al-Busaidi[1], and M. Al-Belushi[1]

[1] Department of Soils, Water and Agricultural Engineering, College of Agricultural and Marine Sciences, Sultan Qaboos University, P.O. Box 34 Al-Khod 123, Muscat, Oman
{rahma.alhabsi,yalmula}@squ.edu.om
[2] Department of Geography, College of Arts and Social Sciences, Sultan Qaboos University, P.O. Box 42 Al-Khod 123, Muscat, Oman

Abstract. The main factors affecting wheat emergence are climatic condition, soil properties and planting depth. Time and percentage of the wheat emergence depend on the interaction between above factors which can be predicted by using wheat simulation model (WSM). WSM is based on three main factors which are soil water potential, soil temperature and planting depth. The general objective of this study was to delineate the best location for wheat production in arid regions such as Oman through linking Wheat Simulation Model (WSM) with Numeric Weather Prediction Model (NWPM) in the platform of the Geographical Information Systems (GIS). Soil temperature and water potential raster layers which were obtained from NWPM were analyzed using spatial analysis tools in ESRI ArcGIS software. Four field trials, over two seasons, have validated positively the linkage of the developed WSM with GIS. The developed model can be promoted as a tool for decision makers to delineate the best location for wheat production in arid regions.

Keywords: Numeric weather prediction model · Simulation model · Wheat · Emergence · Geographic information system · Arid regions

1 Introduction

Food security is an important issue especially after the recent food crisis that hit many parts of the world. Food crisis was not limited only to the substantial rise in the prices of food imports, but extended to the lack of universal availability, and then the scarcity and the difficulty of obtaining these goods. One of the most important crops for food in the world is wheat, where more than two thirds of food is provided by cereals [1] and one third production of cereals is wheat [2]. The total production and wheat planting area in Oman for years from 2009 to 2013 are presented in Table 1 [3].

However, Oman wheat production covers only 0.5–1 % of its need. The food gap between production and consumption increases due to increase in the population. Where the gap increased from about 108 thousand tons in 2003 to about 169 thousand

C. Grueau and J.G. Rocha (Eds.): GISTAM 2015, CCIS 582, pp. 90–103, 2016.
DOI: 10.1007/978-3-319-29589-3_6

tons in 2008. The price of local wheat in the market is about 0.600 R.O/kg compared to imported wheat which is about 0.200 R.O/Kg. This difference between the two prices is due to low yield and high cost of production of local wheat which is 59.5 R.O/ha [4].

Table 1. Total wheat production and planting area in Oman.

Year	Area (ha)	Production (ton)
2009	583	1874
2010	535	2286
2011	620	2126
2012	413	1421
2013	287	1045

Knowing when to plant wheat crop is one of the most important factors for better emergence timing which leads to better wheat yield especially in arid regions. The emergence time, however, is affected by climatic conditions, soil property and planting depth [5]. Moreover, despite the fact that an arid country like Oman has no problem with soil temperature for wheat emergence, Omani wheat growers usually sow almost double number of seeds into a land in order to assure they will get higher wheat production efficiency of that land due to non-emergence of portion of the sowed seeds. That might due to seeds viability and to soil properties especially the optimum soil temperature and soil water potential in addition to planting depth which all play an important role in determining the time of emergence for wheat seedlings. Therefore, in order to increase the wheat production efficiency, there is a need to determine the optimum soil temperature, water potential and planting depth of wheat and so the best time and percentage of wheat emergence can be predicated. However, given the above constraining factors and due to varying values of these factors from time to time and from place to place, estimation of emergence time for wheat will be much easier by using a computer simulation model.

Simulation is a computerized model that describes the behavior of a complex system based on a set of data and dynamic variables and interactive components. The Computer simulation models become popular in many natural systems in physics, chemistry and biology, human systems in economics and social science. Biosystems field is considered as one that requires the use of simulation and computer modeling including the predicting the time and percentage of wheat emergence in field applications.

A wheat simulation model (WSM) was developed by [5] to predict the time and percentage of winter wheat emergence based on the three above mentioned factors: planting depth, soil temperature, and soil water potential. The developed wheat emergence model is based on the hydrothermal time concept which was proposed and further developed, but for germination only, by [6–9]. The WSM governing equation has the following form:

$$t(f)_E = \frac{\theta_{HT}}{(T - T_b)[\psi - \overline{\psi}_b - (probit(f)\sigma_{\psi b})]} \tag{1}$$

where t(f)E is time from sowing seeds to emergence (days), ψ is soil water potential (MPa), and σψb is standard deviation of the base water potential respectively (MPa), T is soil temperature (°C), Tb is the base soil temperature (°C), and probit(f) indicates the number of standard deviations away from the mean that any fraction of the seed population lies. It linearizes a cumulative normal distribution which facilitates modeling efforts [10].

The parameter θ_{HT} is the hydrothermal time to emergence (MPa degree-days). It can be determined using the following formula:

$$\theta_{HT} = (\psi - \psi_b)(T - T_b)t_E \tag{2}$$

where all parameters explained above.

The parameter probit(f) can be determined as follows:

$$probit(E) = \frac{\psi - \frac{\theta_{HT}}{(T-T_b)t_E} - \psi_b(50)}{\sigma_{\psi_b}} \tag{3}$$

Where $\psi_b(50)$ is base water potential for 50 % population of planted seeds.

The developed WSM also includes the calculation of the maximum percent of emergence (Emax) by using the following equation which works as a threshold for the developed model:

$$E_{\max} = c_1 D^2 + c_2 D + c_3 \tag{4}$$

where D is sowing depth (cm), and c1, c2, and c3 are constants related to ψ and T.

The soil temperature and water potential information required for the emergence model are not easy to find for any lands. That's due to the need of using special sensors to measure these factors at time of emergence of the wheat crop. In this study a novel approach is explored in order to utilize the WSM model without the need of using any sensors by finding a better way for extracting the needed model's input data from Numerical Weather Prediction Model (NWPM) which is based on numerical models that deal with set of motion equations. These equations govern the fluid flow partial differential equations [11]. Hence, by linking WSM with NWPM in Geographic Information System (GIS), the best location of wheat production in arid regions as the Sultanate of Oman can be delineated and that was the main aim of this study. The specific objectives of the projects were: (1) to validate the developed simulation model for predicting the time and percentage of emergence using local field data, (2) to examine the performance of the model using different wheat cultivars, (3) to extract needed data from Numerical Weather Prediction Model (NWPM), (4) to re-design the simulation model by linking it with Geographic Information system (GIS) and NWPM, and (5) to create maps delineating the best location of wheat production in the Sultanate of Oman.

2 Materials and Methods

2.1 Field Experiment

Field experiment was conducted at Agriculture Extension Station in Sultan Qaboos University in the Sultanate of Oman (Fig. 1) between first of December 2013 until End of May 2014. The field work was divided into two times. In time one, planting started in 1st December 2013 until the end of harvesting stage March 2014. Planting in second time started on 22 January 2014 until 21 May 2014. The field was divided to four lines, the first two lines for time one and the last two lines for time two. Each line was divided to seven 2 m × 2 m plots. Each plot was divided to ten cultivated lines irrigated by drip irrigation system. The space between one line to another was 20 cm and the wheat seeds were sowed adjacent to the lines with 5 cm spacing. Two wheat varieties were used. One of them is local variety which is Coli and the second one was imported from Kuwait which is KW1. The seeds were sowed at two different depths: 2.5 cm and 5 cm. Hence, the experiment treatment factors are: 2 times × 2 varieties × 2 planting depths (Fig. 2). Three replicates for each treatments planting with one control plot of each varieties were conducted. The planting method in the control plots was by broad-casting. Fertilization was applied as recommend by AES staff experience. By which fertilizer was applied two times a week through fertigation system where combination of urea, potassium nitrate and phosphoric acid was applied trough the irrigation system.

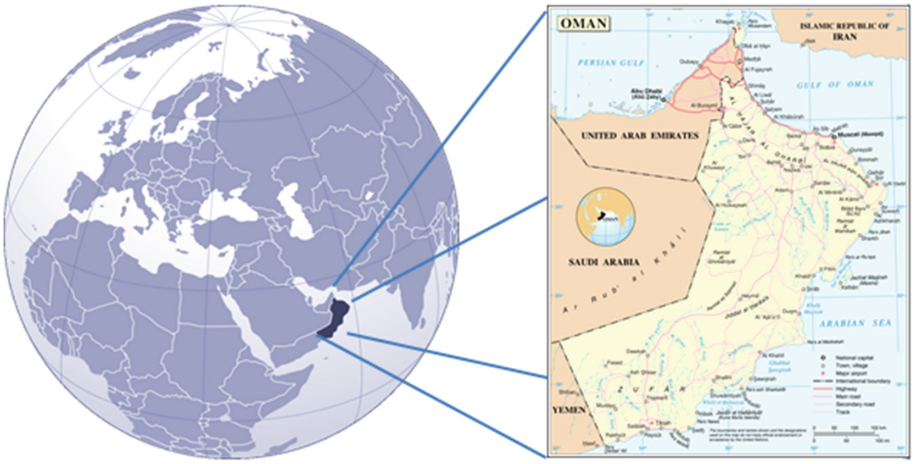


Fig. 1. Sultanate of Oman [12].

2.2 Germination Test

Ten petri dishes were prepared for each variety and in each petri dish a filter paper and ten seeds from each variety were putted and irrigated with water. So, 100 seeds in total for each variety were examined. Germination of seeds was counted and recorded on daily basis.

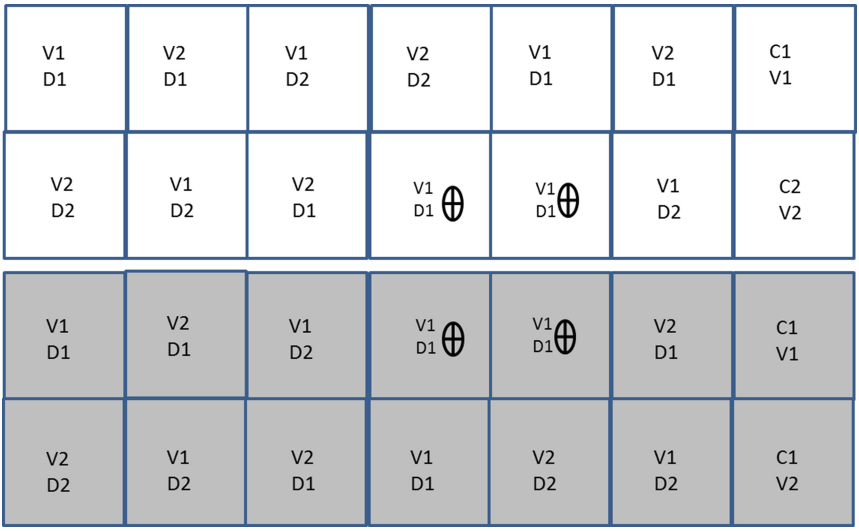


Fig. 2. Field experimental layout. White area is for time one while grey area is for time tow. V1 and V2: variety 1 and 2, D1: 2.5 cm, D2: 5 cm, C1: Control 1, C2: Control 2, and ⊕: locations of 5TE and MPS2 sensors.

2.3 Physical and Chemical Soil Analysis

Soil Sampling. One composite soil sample was taken by collecting different soil samples from different plots and mixed together. Then soil was air dried and sieved through 2 mm.

Moisture Content and Oven Dry Weight. After the soil sample was air_dried and sieved by 2 mm sieve, unspecified amount of soil was taken and its weight was determined before and after drying in oven at 104 °C for 24 h. Then percentage of moisture content and Oven dry weight were calculated and result used in soil texture calculation.

Soil pH, EC and SAR Measurements. Paste saturation sample was prepared and soil solution was extracted by air vacuum for Electrical Conductivity (EC) and pH measurements using electrical electrode. For Sodium Absorption Ratio (SAR), concentration of Sodium (Na), Calcium (Ca), and Magnesium (Mg) were measured using ICP. To find SAR value, the following equation was used:

$$SAR = \frac{[Na]}{\sqrt{[Ca + Mg]}} \tag{5}$$

Soil Texture. Soil sample was air-dried and sieved by 2 mm sieve. From sieved soil, 50 g of soil was assigned into a baffled stirring cup with 10 ml of 0.5 N sodium hexametaphosphate and distilled water was added until half fill of the cup. The mixture

was stirred for five minutes and transferred to 1000 ml graduated cylinder which was filled with distilled water until 1000 ml mark. The suspension was mixed and at the end of 20 s from mixing, the hydrometer was inserted. The first reading of hydrometer was recorded after 40 s and also temperature of suspension was recorded by the thermometer. Then, the hydrometer was removed and re-shacked again. After two hours, the reading of hydrometer and temperature were recorded. The percentage of each particle size was calculated according to these equations:

$$\mathrm{HRc} = \mathrm{HR} + (0.2 \text{ for every } 1^{\circ}\mathrm{C} \text{ above } 20^{\circ}\mathrm{C}) \tag{6}$$

$$\%silt = \frac{HRc\ at\ 40\ \mathrm{sec} - HRc\ at\ 2hr}{ODW} * 100 \tag{7}$$

$$\%Clay = \frac{HRc\ at\ 2hr}{ODW} * 100 \tag{8}$$

$$\%Sand = 100 - (\%Silt + \%Clay) \tag{9}$$

where HRc is corrected hydrometer reading.

2.4 Irrigation

Drip irrigation was used in the study. Ten drip lines were crossing the plots with 20 cm spacing between the emitters. Water required was calculated based on crop evapotranspiration for each stage of wheat development. The historical data like daily maximum and minimum soil temperature and humidity, wind speed and the meteorological data for Seeb weather station were used. Visual basic in Excel sheet was used to create sheet for reference water requirement (ET^{o}). The sheet is based on Penman – Monteith method. To determine the time of irrigation, the discharge of water from meter was recorded. Hence, the time of irrigation was calculated by dividing volume of water required by the discharge.

2.5 Sensors

The inputs parameters including water potential and soil temperature, for wheat simulation model can be obtained from the sensors installed in the field. A 229 heat dissipation sensor from Campbell Scientific Company, USA, is used to measure water potential indirectly using principle of heat dissipation. The principle of heat dissipation is whenever there is water potential gradient between sensors and the surrounding soil, water movement between sensor and soil take certain time to reach equilibrium. Hydraulic Equilibration time depend on magnitude of water potential gradient and hydraulic conductivity. The changes in water content of sensor ceramic matric lead to change in the thermal conductivity of sensor/soil complex. There is exponential relationship between water content and thermal conductivity. As water content in the ceramic sensor increase, the thermal conductivity increase. A 229 heat dissipation sensor is porous ceramic

cylindrical shaped with thermocouple and heating element at the middle of the cylinder. It has the ability to measure a wide range of matric potential from −10 to −2500 kPa and it is compatible with most Campbell Scientific data logger and multiplexer. Also, it is known by long lasting without need for maintenance. The 229 should be installed horizontally at the desired depth and good contact between the ceramic cylinder and soil must be exist (Instruction manual of model 229, Campbell Scientific Inc.). Soil water potential can also be measured by using MPS2 from Decagon Devices Inc., USA. MPS2 is ideal sensor for a range of water potential measurement between −0.01 and −0.5 MPa and soil temperature between −40 and 60C (Decagon Devices Inc.). A 5TE sensor from Decagon devices, USA, is used to measure volumetric water content, soil temperature and electrical conductivity. It measures the three parameters independently. The volumetric water content is obtained by measuring the dielectric constant of the media using electromagnetic field supplied from the sensor with 70 MHz while the soil temperature is obtained from thermistor which is installed at surface of the sensor with a range of readings of −40 to 50° C. Electrical conductivity is obtained by using stainless steel electrode array and the reading is taking within the range of 0 to 23 ds/m. The sensor is easily installed in the field through pushing it directly to undisturbed. EC is measured by applying alternating current to two electrodes and measuring the resistance between them soil (Decagon Devices Inc.).

2.6 Linking WSM with GIS

To link the wheat simulation model in GIS, soil temperature and volumetric water content raster layers were required as input for the model and they were extracted from the European Centre for Medium-Range Weather Forecasts (ECMWF) website which is an independent intergovernmental organization established in 1975. The Centre provides a catalogue of forecast data worldwide that can be purchased by businesses and other commercial customers for national community. Different spatial analysis tools in GIS were used to create the final emergence map mainly rater calculator and reclassifying tools (Fig. 3).

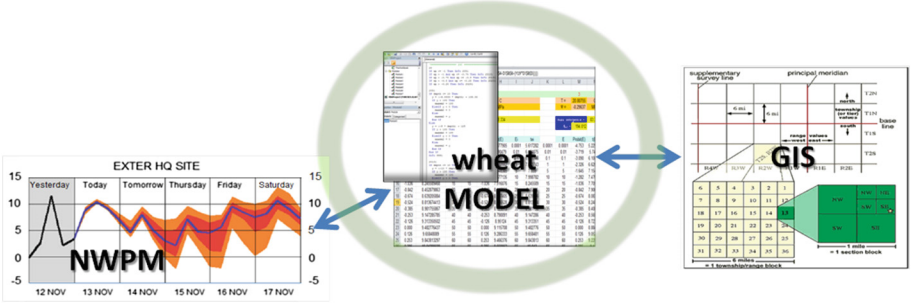


Fig. 3. Schematic chart of creating wheat prediction model in GIS.

2.7 Statistical Analysis of Physical Characteristics of Wheat

Physical characteristics of wheat like length of plant, number of spikes per plant, number of seeds per spike and weight of 100 seeds of each variety were examined at the late season of each planting time. Plant height was measured by taking average length of random selection of plants using meter tab from the soil surface to the begging of the spike. The average plant length in each plot was recorded. Random plants were also selected to find out number of spikes in it and average numbers of seed in each spike were counted. After the seeds removed from its spike, 100 seeds from each plot was weighed using digital scale. For statistical analysis, ANOVA for single factor was used to examine the differences among each factor.

3 Results and Discussion

3.1 Germination Test

Germination was started after second day of planting. For local variety (Coli) germination was 99 % in second day and 100 % in third day. For KW1, germination was 46 % in second day and after second day till day six of planting, germination was stopped at 60 %.

3.2 Physical and Chemical Soil Analysis

Soil Texture Class for planting field is silt loam which is medium textured soil. For wheat production, the best soil type is well drained fertile loamy soil to sandy loam soil [13]. Electrical conductivity for paste soil extract is 478 μS\cm, Sodium absorption ratio (SAR) is 2.35 and pH is 8. The best range of soil pH for wheat production is between 6 and 7.5 [13]. The pH value in this study exceeded the optimum range a little bit. Alkalinity affects wheat production negatively in two ways by decreasing water infiltration and decreasing availability of micronutrients. Since that the solubility of Iron (Fe), manganese (Mn), copper (Cu) and zinc (Zn) decrease with increasing soil pH. However, deficiency of the last three ions due to alkalinity has not severe effect in wheat production area [14]. Based on the EC and SAR values, soil was classified as non-saline and non-sodic soil. It was normal soil and suitable for wheat production.

3.3 Irrigation

Crop evapotranspiration and Irrigation Scheduling for each plant stage are presented in Tables 2 and 3 for planting time one and two. As expected, for first time of planting, crop coefficient (Kc) started with small value but it increased as plant grown and developed and then decreased at late season as plant senesce. The crop water requirement (ET_C) followed the same pattern of the Kc. Because the plant need more water at growth and developing stages, the duration of irrigation increased until the late season where it decreased and stopped for one week before harvesting. The ET_o and

Table 2. Evapotranspiration and Irrigation Scheduling calculation for time one.

Stage	Duration days	k_c	Average ET_o (mm/day)	Average etc (mm/day)	Volume (m^3)	Irrigation time (min)
Initial stage	20	0.4	3.3	1.32	0.07	1.3
Growth and branching	30	1	3.1	3.1	0.17	3.1
Completion of growth and flowering	30	1.2	3.5	4.2	4.2	4.3
Composition and grain filling	30	1	4.7	4.7	4.7	4.8
Late season	15	0.4	5.3	2.12	0.12	2.1

Table 3. Evapotranspiration and Irrigation Scheduling calculation for time tow.

Stage	Duration days	k_c	Average ET_o (mm/day)	Average etc (mm/day)	Volume (m^3)	Irrigation time (min)
Initial stage	20	0.4	3.6	1.4	0.08	1.4
Growth and branching	30	1	4.5	4.5	0.25	4.6
Completion of growth and flowering	30	1.2	5.4	6.5	0.36	6.7
Composition and grain filling	30	1	6.7	6.8	0.38	7
Late season	15	0.4	8.8	3.5	0.20	3.6

ETc values increased during time two of planting than time one due to the increase of air temperature. That resulted in increasing the duration of irrigation supply in time tow more than in time one.

3.4 Linking WSM to GIS

Since WSM requires soil water potential as one of the input data and since both sensors the 5TE and MPS-2 measure volumetric water content for, there was a need to convert these sensors data into water potential. A relationship between soil's volumetric water content and water potential were obtained. By which a regression equation of calculating water potential from volumetric water content was found as follows:

$$\Psi = -20.208wc^2 + 11.747x - 1.7176 \quad (10)$$

Where Ψ is soil water potential (MPa) and wc is soil water content (%).

The time of emergence equation of the WSM has a probit parameter. It is a value that is used to indicate the number of standard deviation away from the mean that any fraction of seed population lies. Also, it is used to linearize a cumulative normal distribution which makes the model easy to work [5, 10]. Since it is not possible to calculate the probit value in GIS directly, relationship was found between emergence percentage and the probit value which can be expressed in the following equation:

$$Probit = 0.3383 \ln(E) - 1.996 \tag{11}$$

where Probit is probit value and E is emergence percentage (%).

The soil temperature and volumetric water content data at level 1 were extracted from ECMWF by selecting a month long (October 2013) of the required data with time steps of 24 h. Equation 10 was used in the raster calculator to find soil water potential from the volumetric water content layer extracted from ECMWF. The data format was in Network Common Data Form (NetCDF), so they needed to be converted to raster data format for GIS. That was done by selecting "Make NetCDF Raster Layer" which is a sub menu under Multidimintion menu of ArcToolbox.

After raster layers of soil temperature and water potential were created from NetCDF file, coding the equations of hydrothermal time and maximum emergence was done using raster calculator. Equations of maximum emergences are based on certain range of temperature and water potential with a total of 23 conditions. Hence, 23 raster layers of pixels that match with emergence conditions have certain values of maximum emergences and therefore they made to appear in colored pixel. However, the other pixels that do not match the emergence conditions of emergence have no data. To merge all emergence raster layers in one layer to be used in the final equation of time of emergence, the no data pixels were converted to zero value using Reclassify tools from spatial analysis. All reclassified emergence raster layers were combined in one raster layer using addition tool in raster calculator producing a single layer map. Figure 4 shows the outcome of the GIS-Linked WSM modified model for the Arabian Peninsula region after extracting the NWMP data from the ECMWF.

3.5 Statistical Analysis of Physical Characteristics of Wheat

The statistical results shown that the two planting depth (2.5 and 5 cm) for wheat planting at the end of the January has no significant effect on the wheat physical parameters of both varieties.

This result of effect of planting depth agrees with previous studies done by [15, 16] who found that wheat is less sensitive to the planting depth ≤4 cm and there is no significant reduction in total emergence percentage.

To find out effect of planting time of same planting depth of same variety, also ANOVA of single factor was used. It is shown that there is effect of planting time on Coli variety which was planted at 2.5 cm depth ($P < 0.05$) that is the average height of wheat which planted at the first of December 2013 (120.3 cm) taller than wheat which planted at end of January 2014 (84.6 cm). The same result was got for Coli variety which planted at 5 cm depth. Plant height of wheat at time one (119.6 cm) is taller than

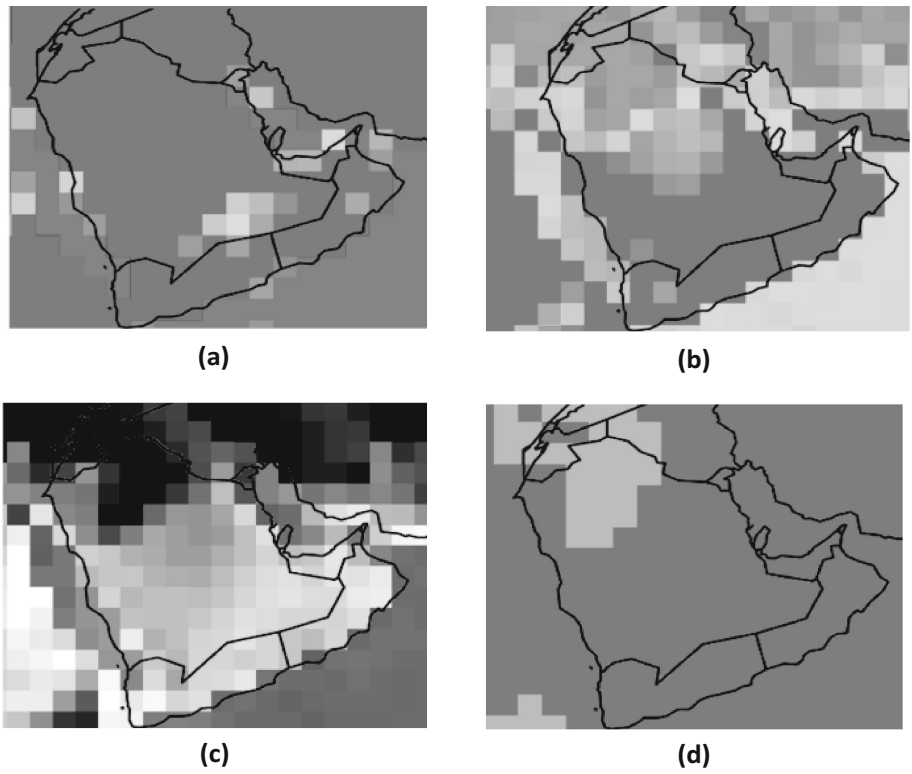


Fig. 4. GIS linked WSM outcome: (a) soil Water Potential, (b) Soil temperature, (c) Hydrothermal time factor and (d) Time to emergence.

plants at time two (61.6 cm). On the other hand, statistical results shown that planting time has no effect on KW variety at two depths 2.5 cm and 5 cm ($P > 0.05$).

For the Coli variety at two planting depths and for KW variety at depth 2.5 cm, results shown there is no effect of planting time on number of spike per tree. But for KW variety which planted at 5 cm, planting date has effect on number of spike per plant ($P = 0.0351$) where the average number of spikes per plant was more at time two (9 spikes/plant) compared with time one (6 spikes/plant).

Both planting depths of Coli variety were affected by the planting time ($P \leq 0.05$) where the average number of seeds per spike at time one planted at depth of 2.5 cm was 25 seeds/spike and at time two was 15 seeds/spike. The same result was found for planting depth 5 cm. The number of seeds per spike of KW variety was less affected by changing time for both depths.

The weight of 100 seeds of Coli variety at two planting depths was not affected by time of planting whereas KW variety was affected by the time of planting. The average weight of 100 seeds at time one of both planting depth was almost 4 g whereas for time two was about 2 g.

Most of literatures shown delay in sowing wheat from the proper planting time lead to poor stand establishment and less number of productive tillers and less yield [17, 18]. Because of that in Oman it is usually planted at the second half of November [19].

3.6 Comparison Between Field Experiment, Wheat Prediction Model in Excel and GIS

Figure 5b shows emergence results from time one. Emergence started after day 4 from planting for all treatments and reached maximum emergence at day 12 after planting. The maximum percentage was achieved for Coli variety at 2.5 cm planting depth which was 75 % at day 12 after planting while at planting 5 cm depth, the maximum emergence was 62 %. For KW variety, at planting depth 2.5 cm, the maximum emergence was achieved after day 12 with 53 % and at planting depth of 5 cm, the maximum emergence was 36 %. Emergence results of time two are presented in Fig. 6b. The emergence of both varieties started after day 5 from planting and reached the maximum at day 12 from planting. Coli variety had higher percentage of emergence than KW variety where at 2.5 cm planting depth it reached the highest percentage among other treatments with 77 %. The KW variety reached its higher emergence percentage at shallow planting depth (50 %) in comparison to deep planting depth (44 %).

The WSM predictions for the wheat to start emerging were 5 days after planting (DAP) in time one and 5 DAP in time two, while in the field it took 4 DAP in time one and 5 DAP in time two (Fig. 5a). The DAP to reach 50 % emergence was predicted by WSM as 6 DAP and 6 DAP in both times one and two, whereas, in the field, the wheat reached 50 % emergence after 7 DAP and 6 DAP in time one and time two (Fig. 6a). The maximum percentage achieved in the field, for time one, was 74 % whereas WS prediction was 67 % while for time two, the maximum percentage was achieved in the field was 77 % whereas WSM prediction was 80 %. The DAPs to reach maximum emergence was predicted by the WSM as 7 DAP while in the field it was achieved after 11 DAP for time one. For time two, the SWM prediction was 6 DAP in the field it was achieved after two more days which was 8 DAP.

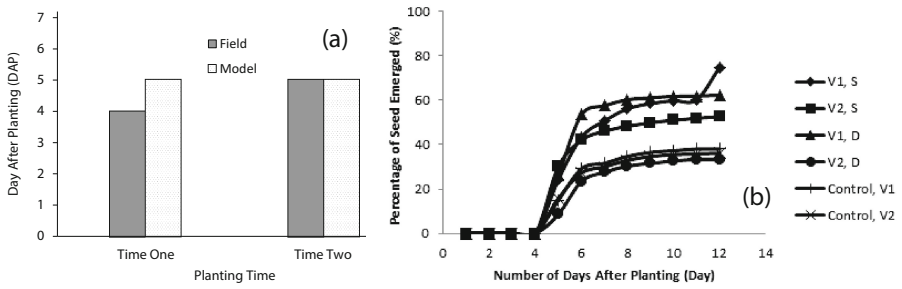


Fig. 5. Wheat seed Emergence as DAP (a) and percentage (b) for Time one, where V1 Coli variety, V2 KW variety, S: shallow planting depth (2.5 cm), D: deep planting depth (5 cm).

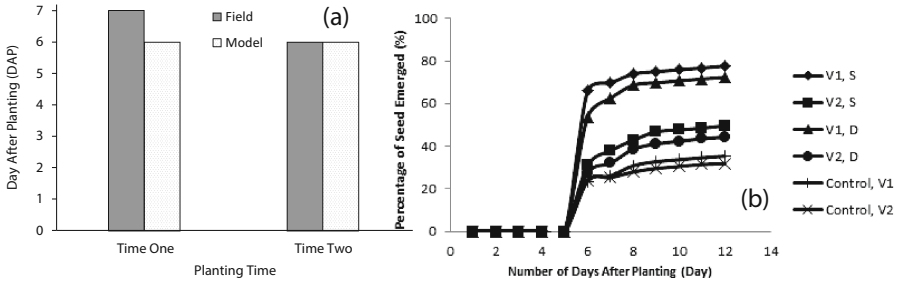


Fig. 6. Wheat seed Emergence as DAP (a) and percentage (b) for Time Two, where V1 Coli variety, V2 KW variety, S: shallow planting depth (2.5 cm), D: deep planting depth (5 cm).

4 Conclusion

Wheat is one of the most important crops for food in the world including arid regions as the Sultanate of Oman. For arid regions, It would be especially useful to develop a method that can increase wheat production location but with minimum possible water consumption. By integrating GIS technology, a Wheat Simulation Model (WSM) was further developed for optimal wheat production location in arid regions. WSM was tested with field data and the effectiveness of the model was proven. The study illustrates a good example of GIS application for solving spatial and temporal crop production optimization problem. By which, it becomes possible to delineate (map) the WSM outcome that can cover any part of world including arid regions by eliminating the need of using any sensors to run the simulation.

Acknowledgement. The corresponding author, as a PI, appreciates the financial support for this research provided by Sultan Qaboos University (IG/AGR/SWAE/12/02).

References

1. Bushuk, W., Rasper, V.F.: Wheat production, properties and quality. Blakie Academic and professional an imprint Chapman and Hall, pp. 273–310 (1994)
2. Carver, B.F.: Wheat Science and Trade. World Agriculture Series, 1st edn. Wiley-Blackwell, Ames (2009)
3. MAF: Ministry of Agriculture and Fishery, Annual Agricultural and Fishery Census Book. Developmental Media Department (2013)
4. MAF (Ministry of Agriculture and Fisheries): Wheat Production in the Sultanate of Oman: Reality, Challenges, and Future Vision. Wheat Production Project National Committee Report. SQU, Oman (2009)
5. Al-Mulla, Y.A., Huggins, D.R., Stöckle, C.O.: Modeling the emergence of winter wheat in response to soil temperature, water potential and planting depth. Trans. ASABE **57**(3), 761–775 (2014)
6. Gummerson, R.J.: The effect of constant temperatures and osmotic potentials on the germination of sugar beet. J. Exp. Bot. **37**, 729–741 (1986)

7. Bradford, K.J.: A water relations analysis of seed germination rates. Plant Phys. **94**, 840–849 (1990)
8. Bradford, K.J.: Water relations in seed germination. Seed Development and Germination, pp. 351–396. Marcel Dekker Inc., New York (1995)
9. Cheng, Z., Bradford, K.J.: Hydrothermal time analysis of tomato seed germination responses to priming treatments. J. Exp. Bot. **50**, 89–99 (1999)
10. Bauer, M.C., Meyer, S.E., Allen, P.S.: A simulation model to predict seed dormancy loss in the field for Bromus tectorum L. J. Exp. Bot. **49**, 1235–1244 (1998)
11. Stensrud, D.: Parameterization Schemes Keys to Understand Numerical Weather Prediction Model. Cambridge University Press, Cambridge (2007)
12. Mapsof: Where is Oman Located? (2014). http://mapsof.net/map/
13. DAFF (Department of Agriculture, Forestry and Fisheries): Wheat Production Guideline, Public of South Africa (2010)
14. Heyne, E.G.: Wheat and Wheat Improvement, 2nd edn. Madison, Wisconsin (1987)
15. Jim, H., John, J., Dottie, C.: Wheat Planting Depth Study, Department of Agronomy, Princeton (2000)
16. Keshtkar, E., Farnazkordbacheh, M.B.M., Mashhad, H.R., Alizadeh, H.: Effects of the sowing depth and temperature on the seedling emergence and early growth of wild barley (Hordeum spontaneum) and wheat. Weed Sci. Soc. Jpn. Weed Biol. Manage. **9**, 10–19 (2009)
17. Akther, M., Ahmed, N., Nasrullah, M., Ali, B., Zahid, A.R., Shahid, I.: Effect of late planting on emergence, tillering and yield of various varieties of wheat. Pakistan J. Anim. Plant Sci. **22**(4), 1136–1166 (2012). ISSN: 1018-7081
18. El-Gizawy, N.: Effect of planting date and fertilizer application on yield of wheat under no till system. World J. Agric. Sci. **5**(6), 777–783 (2009). ISSN: 1817-3047
19. Qabashi, A.W.: Most Important Cereal Crop in Sultan of Oman, Minstry of Agriculture (2003)

Towards Geospatial Tangible User Interfaces: An Observational User Study Exploring Geospatial Interactions of the Novice

Catherine Emma Jones[1(✉)] and Valérie Maquil[2]

[1] CVCE, Luxembourg, Luxembourg
catherine.jones@cvce.eu
[2] Luxembourg Institute of Science and Technology (LIST), Luxembourg, Luxembourg
valerie.maquil@list.lu

Abstract. Tangible user interfaces (TUI) such as tangible tabletops have potential as novel and innovative learning environments for mapping applications across a wide range of geospatial learning activities. This is because they offer a more natural and intuitive class of interface to users and they are fun to use. For realising their potential as a new type of geo-technology, they must be easy and straightforward to learn and remember how to use. Furthermore, the different types of tangible object interactions should align to the mental models and cultural perceptions of different types of users. This paper reports on the results of an initial observational of a small set of novice users. Users were recorded completing six tasks whilst thinking aloud. The resulting analysis revealed how easy it was for the novice to discover the different types of geospatial tangible interactions (e.g., zoom, pan, adding layers, working with layers). Formed around the categories of (1) everyday cartographic elements and their everyday metaphors, (2) object manipulations, and (3) offline interactions we propose a set usability guidelines for geospatial tangible tables. The aim is to provide an evidence base on which to improve future iterations of the improving their usability, usefulness and increasing their potential as a learning interface.

Keywords: Tangible user interface · Cartography · Geospatial information · Usability · Qualitative study · User study · Tangible interfaces · TUI · Mapping · User centred design · Interactive tabletops

1 Introduction

Almost everything that happens, happens somewhere and knowing where something happens is critically important [20]. Encoding location within a multitude of everyday interactions, objects and events, subsequently means we can unlock solutions to a wide

This article is a substantially extended version of an already published conference paper, published in the conference proceedings of the GISTAM 2015 Conference [41]. Paper presented at the 1st International Conference on Geographical Information Systems Theory, Applications and Management (GISTAM 2015), Barcelona, Spain]. Substantial changes have been made for this extended version, but some parts have remained unchanged.

C. Grueau and J.G. Rocha (Eds.): GISTAM 2015, CCIS 582, pp. 104–123, 2016.
DOI: 10.1007/978-3-319-29589-3_7

variety of complex problems. The age old adage "a picture tells a story of 1000 words", rings true. Maps as learning, exploration and analysis tools have great power, aiding understanding of complex phenomena and to instigate and engage discussion. Knowing something about where and being able to unravel complexity has given rise to age of the map. Furthermore, if we are able to integrate maps within technology and interfaces that facilitates learning, discussion and collaboration there will be even more potential for knowledge and capacity building.

In recent decades, advances in geotechnologies have transformed the methods for acquiring, processing and sharing Geographic information (GI) [9]. the now ubiquitous web-mapping applications based on the online "slippy map" API's (Bing Maps, Google Maps, OpenStreetMap) [32] have largely contributed to the public availability of geospatial information on web sites, services, and apps making geographic data accessible across a wide array of devices. They enable large volumes of geographic data to be served as either raster or vector tiles (as in the case of for example Mapbox) allowing easy and fast consumption of interactive maps. Indeed such interfaces developed around 'slippy maps' are a primary GI tool for the lay person and have been adopted by National Mapping Agencies, such as the UK Ordnance Survey or the Geoportail.lu produced by the National Cadastral and Mapping Agency for Luxembourg. Such changes together with the Open data and Open Software movement drove a transformation in how we consume and produce maps. These simple and intuitive interfaces were a breakthrough for increasing participation in mapping activities and use of resulting maps. They prove that easy to use interfaces adhering to usability principals are effectively useful and usable to a broad range of users [18].

With this revolution in GI come new opportunities for user research and novel interface design. Turning from the over reliance on sophisticated and often complex interactions within desktop solutions which often isolates users and hinders knowledge construction and moving towards new technologies and interfaces with users at the core, further adding value by being fun and engaging [8, 11, 16, 21, 37, 38]. We are observing a shift towards a focus in user centric design for geospatial interfaces and mapping representations that are more in tune with natural intuitive user interactions that align with users' mental models. Turning away from the over reliance on sophisticated and often complex interactions which alienate users and impairs knowledge construction and moving towards the adoption of new geo interfaces that place users at the core.

Tangible user interfaces (TUI) as an emerging class of interfaces [14], have a clear potential for improving collaboration, learning and knowledge for mapping. TUI's offer large representations that encourage collaborative working amalgamated with intuitive and tactile user interactions [22]. More broadly, there are a number of benefits. Namely, the inherent knowledge of the physical objects helps to provide users with a feeling of intuitive directness [1]. TUIs are natural supports for collaboration; they enhance group productivity, by bringing users around a shared discussion space and supporting them in coordinating their actions using the physical objects [12]. Moreover, applications available via tangible devices have an inherent spatiality, both literally and metaphorically [28]. These benefits have led to the implementation of a variety of geospatial TUI research scenarios: landscape modeling, maritime operations,

urban planning, topographic change, tourism, emergency response, disaster management, and collaborative decision-making [22, 24, 28, 36].

Tangible tables have potential as novel and innovative learning environments for mapping applications but to realise this potential they must be easy and straightforward to learn and use and interaction must match user' mental models. Their functionality must be memorable and peripheral so that users do not focus on interactions but on knowledge construction. For tangible interfaces to be useful learning interfaces they must also provide users with a sense of satisfaction. They must not cause frustration or fear. Such aspects are typically studied in usability and user experience.

In the context of GUI online mapping sites, various usability studies have been conducted exploring issues of user interface design, and user experience (for example, [34]). Nivala et al. [31] evaluated the usability of four different mapping sites and identified 403 usability problems. They defined a series of guidelines related to the user interface, the map, and the search operations. Similar insights are provided by Haklay and Tobón [11] and Jones and Weber [18] who identify the need for mapping interfaces to be designed with the notion of less is more.

Comparable usability studies related to mapping on tabletop interfaces, however, have not yet been conducted. Whilst there is research interest in developing geospatial applications on interactive tabletops, it has focused on technical implementations at the expense of detailed user research and understanding. The few studies have only marginally incorporated user centric approaches and testing. Based on usability questionnaires and design research, [30] found that the fluidity of interactions creates an engaging user experience, and that users handle well the switching of views. Scott et al. [36] evaluated the usability of a pen-based tabletop using a task-based approach, identifying benefits related to the reposition and reorientation of windows, multiple access points, and the physical manipulations with the pen creating a high level of awareness in the group. Problems have been identified regarding the inconsistency in the orientation of information windows, clutter of information at certain zoom levels, as well as the lack of linking between related information items.

Given this information, it is fair to say that there is a lack of research centered on users of TUIs, their geo-understanding and the explicit interaction with GI via tangible objects. This is due to an absence of joint work between geographical experts and designers of these emerging technologies. There is a research opportunity associated with the disciplinary gap. Traditional desktop mapping and GIS systems are developed from the notion that spatial is special [20], but currently geospatial research with TUIs fail to recognize the spatial and unique nature inherent to geographic data, visualisation and interaction. So whilst a TUI is a different, highly innovative interface it is still subject to the challenges of working with spatial data and their unique interactions, requiring detailed practical investigations enriched by individual experiences. To ensure these classes of technologies reach their full geo-potential, it is necessary to provide greater understanding of the unique geo-interaction patterns and collaboration potential that tangibles afford.

While previous works [30, 36] have already provided first insights on how users interact with maps on interactive tables, the study presented in this paper is novel in two aspects. First, it focusses on cartographic interactions on such tables, i.e. the basic interactions required to explore and analyse a digital map. Second, it investigates the

use of tangible objects for that purpose: our cartographic interactions are implemented by means of manipulations with physical objects. In this paper, we explore the relationship between the user and the spatial interactions in order to determine what is the most intuitive and effective use of tangibles for geographic interactions such as zoom, pan and working with layers and their legends. The aim is to describe the results of an initial qualitative usability study carried out on a geospatial tangible table and to provide insights on how novice users interact with geospatial data through a tangible table.

An existing interface was the starting point for the study [25]. The tangible interface was developed to incorporate geospatial elements using Rapid Application Development Techniques integrating existing projects' geospatial data and cartography. The purpose being to develop proof of concept that tangible geospatial interfaces can be useful for particular scenarios, in this instance the focus was on a Logistics scenario.

2 About the Geospatial Tangible Table

The geospatial tangible table allows users to explore and analyse digital maps projected onto a tabletop. Interactions with the map are carried out using physical objects that are placed, shifted, and twisted on the tabletop [25].

The rounded tabletop is sized 150 × 105 cm, with an interactive surface of 120 × 75 cm. Using the tracking framework "reacTIVision", the tangible objects are tagged with optical markers that are detected by a camera mounted on the bottom of the table. The user interface and digital maps together with other types of feedback are then projected onto the tabletop (see Fig. 1).

Fig. 1. Digital maps on the tangible table.

The system was developed in 2013–2014 in an iterative approach with user input. The digital maps have been created in the context of sustainable freight transport in

North Western Europe, related to the EU Interreg IVb Weastflows[1] project. The project addresses the societal challenge of the drastically increasing volumes of freight moved in Europe, with the vast majority being transported by road. To achieve more sustainable freight transport, congestion issues need to be addressed while reducing the environmental impact of freight movements. This requires the joint work of multiple stakeholders, who are not experts in manipulating geographical information, to understand current and planned freight transport infrastructure and to identify future opportunities for sustainable and more efficient supply chains, an inherently geographical problem. To support such discussion of new, sustainable solutions for urban freight transport, data on the existing logistical infrastructure (roads, railways, freight airports, ferry crossings,..) was gathered and cartographically represented by the Interreg IVb Weastflows project and made available via web mapping services (WMS). Further, to support face-to-face collaboration of multiple stakeholders, the need of a tabletop interface has been identified. In order to be able to study the role of physical objects in this context, we decided to implement the geospatial application as tangible tabletop. In multiple iterations we designed a series of basic geospatial interactions, that we progressively extended by more advanced interactions using the methodology of rapid prototyping. This allowed us to create a first prototype of a geospatial TUI, and to test, improve and refine it based on input from different types of users.

While the system with all the interactions, as well as the used software architecture has been reported elsewhere [25] we here describe the basic spatial interactions that were investigated in the usability study (see Fig. 2):

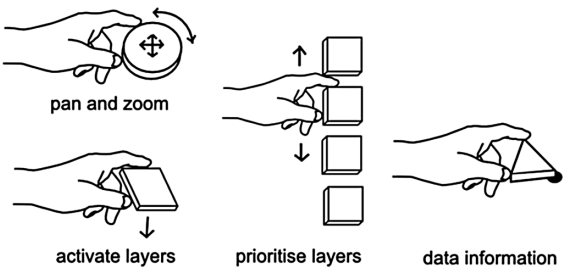


Fig. 2. Basic cartographic interactions implemented with tangible objects.

- **Panning:** A circular object can be placed on the map. Through dragging it across the table, the map view is moved in the same direction (panning). When lifted and dropped at another location, no panning is performed.
- **Zooming:** The same circular object can be rotated to the right to zoom in, and rotated to the left to zoom out.
- **Activate Layers:** A set of square objects is provided, each object representing a different geographical data layer. To activate the layer, the object is placed

[1] http://www.weastflows.eu.

anywhere on the table. The legend is then visualized in a box displayed on the right of the placed object. When the object is removed from the table, the layer is deactivated.

- **Prioritise Layers:** The vertical position of each layer object determines the order in which information layers are drawn. Layer objects laying nearest the bottom of the table are drawn first, while layers lying at the top are drawn last – hence may occlude any layers underneath.
- **Data Information:** A triangular object shows a black dot at one of its corners. This dot can be placed on any geographical object within the map. When the object remains in a same position for 500 ms, a window with the description is opened next to the graphical element. When removing the object, the window closes again.

3 Experiment Methodology

A qualitative approach was adopted to collect user interaction data describing an enriched view of the participants' perspective of the tangible geospatial interface. The study combined participant observation (video and observer) together with the Think Aloud protocol and a short post-test interview. For the experiments, we were interested in understanding the ease of use from the perspective of a new user. The Think Aloud protocol (ISO/TR 16982:2002) has proven an effective tool for investigating and understanding users' personal experiences with mapping technology (e.g. Jones and Weber [18, 38]). It requires participants to describe what they are doing, thinking and feeling as they complete the experiment worksheet.

The experiment had 6 tasks, which took participants 30 min to complete. The tasks were designed around the basic cartographic interaction patterns a new user would be expected to learn (zoom, pan, adding layers, rearranging layers, working with legends and interpreting thematic maps). Complexity of the tasks increased as users completed the work sheet and some interactions were repeated to enable analysis of memorability of functionality. A task sheet provided participants with the detailed activities they had to complete as well as a set of questions they should answer as they worked through each task. The tasks were designed around the European logistics scenario whereby participants could explore a variety of associated geographical data for Europe. Tasks were based on the following themes:

- Locating Luxembourg and its greater region (zoom, pan and adding data layer).
- Adding more logistics information (working with multiple data layers)
- Working with different information and prioritising it (zoom, pan, switching data on and off, rearranging layers to create visual hierarchy)
- Interpreting meaning from the map (using legends, working with the info tool)
- Working with thematic maps and different layers (zoom, pan, working with layers and legends).

A pilot experiment evaluated the test protocol for consistency, errors and timeliness. Prior to commencing the experiments participants were (1) provided with an information sheet outlining what they would be doing and why, (2) given the

opportunity to ask questions, (3) informed what data were being collected and how it would be used, (4) asked to sign a consent form and (5) completed a brief general IT questionnaire to gauge computer literacy and experience with GIS. An experiment room was set up, comprising of the tangible table and the objects, a camera (GoPro Hero 3) and seating for observers. The video camera was placed on the wall opposite the participant looking down at the table (see Fig. 3 (left)) in order to record the interactions with the objects and how they used the space. At the beginning of each experiment the objects were set in the same place and order (see Fig. 3 (right)). Participants were not provided with any other instructions other than the task sheet to provide an indication of how learnable such interfaces are.

Fig. 3. Setup of the experiment room (left) and initial position of the objects (right).

Eight participants (N = 8) were recruited by identifying colleagues in other research departments who were interested in the technology. There were two pre-requisites (a) participants must never have used a tangible table interface before and (b) participants must be comfortable Thinking Aloud in English. An equal mix of genders participated (4 females and 4 males) with an age range between 20 and 45. All were familiar with online mapping websites such as Google Maps (3 frequent and 5 occasional users). No testers routinely used desktop GIS although 2 participants have used it: 1 described himself as a novice with less than 1 years' experience, the other as an intermediate user with 1 to 3 years' experience). All participants were IT literate with 4 participants stating they have experience in application development.

There are many debates on the number of users required for usability testing. The number of recommended users range from 5 [19], 10−20 [3] to 10±2 [13], justifying our sample size, there are diminishing returns for discovering additional issues.

4 Analysis of Results

A qualitative analysis was jointly conducted by a mapping expert and a TUI designer to enable detailed analysis of the fundamental functions required in interactive digital mapping applications. It focused on user interactions associated with the basic cartographic functions: zooming, panning and working with layers and their associated legend. Our purpose was to understand issues associated with ease of learning and ease of use. In mid-term perspective, the results of this analysis should provide input for an

iterative research revealing insight into how learnable tangible mapping interfaces are and the extent to which they are intuitive to the novice user(s).

In the first instance the functions under investigation are (1) creating a basic map with one vector layer (2) navigating the map by zooming and panning and (3) working with more than one data layer.

4.1 Getting Started: Making a Basic Map with One Data Layer

Participants were asked to create a map, made up of one geographic boundary layer. They were asked to add data for European country borders onto the table and thus creating their first, albeit simple, vector map of Europe. To complete the task, which all participants managed to do (100 %), they had to select the correct object and place it on the table. From observing the participants it was clear they were uncertain how to start using the table, none of them had used such an interface before. At first they showed both confusion, bemusement and wonder, "*How do I start this?"* (P2). Four participants were observed shrugging their shoulders and/or waved hands or arms before either exploring the objects at the front of the table or touching the table (P1, P4, P6, P7). After their initial perplexity, users then explored the table based on their prior experience with technology. Wondering how to start, P7 touches the table, shrugs and then says "*... am I supposed to click on something?"*. P2 first attempted a vertical stroke down the table with the exclamation "*OK, nothing happens..."* where as P4 waggles the fingers on the table to see if anything would happen and says, "*it seems to me I should switch something*".

All users were surprised by and hesitant to use the objects to create their map, believing that the identification of the correct object was, "*a lucky shot, I guess!"* (P1). Users unfamiliar with such interaction objects, initially explored the objects intently: scanning the objects, then selecting one, examining it by turning them around in their hands, turning them upside down. One participant even steps away from the table to look at them in situ from a distance prior to selecting the correct object and placing it on the table (P6).

4.2 Zoom and Pan

The next stage in the task sheet was to zoom to a specific Country. In this case, we asked participants to zoom to Luxembourg. The task sheet showed them an example map view that they should try and create with the tangible table interface. To complete this task, participants had to use the zoom and pan object to move the European border layer to locate Luxembourg in the centre of the map. Experience with prior technology, once again heavily influenced how participants investigated this interaction. On their first attempt at interaction, all participants used the familiar touch interactions common to mobile and tablet technology. We observed vertical swipes of the table from top to bottom (P1) pinching thumb and forefinger together (P2), using middle finger and forefinger pinch to try and zoom (P3), touching the table by moving two hands towards each other (see Fig. 4).

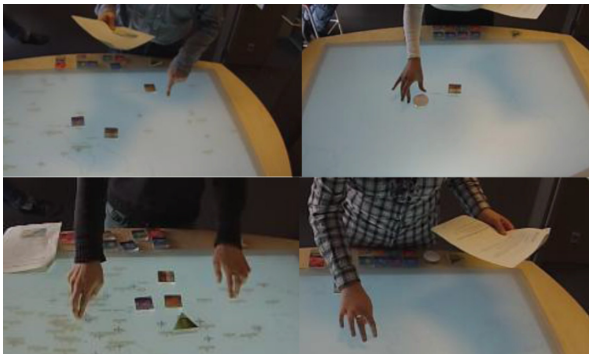


Fig. 4. First zoom and pan attempts with hand gestures: (a) vertical swipe (b) pinching (c) using two hands (d) middle and forefinger.

Panning, more than zoom, appeared to be the most intuitive and easy to learn. Once the object was correctly identified, to differentiate it from the layer objects it was a different shape and colour, all participants picked it up and dragged it across the table. Five participants took less than 10 s to work out the functionality. In a number of instances the map object was moved too quickly and the loading of the data was too slow. To stop panning, participants instinctively moved the object off the table. The type of interaction is closely reminiscent of the type of mouse interaction one would use to pan with an online slippy map.

The zoom functionality was less obvious, it was difficult to locate on the interface, as it was hidden within the pan object and was less intuitive. We observed seven participants zooming when using the object to pan. This was unintentional and unexpected. As they panned with the object it twisted causing the map to accidently change scale. This functionality led to confusion and frustration, "*it zooms but I don't know why*?" (P2) or "*I have no idea how to zoom, sometimes it works a bit*" (P3). Twisting the object was not the most instinctive action. Participants explored the object in many different ways: turning it over, doubling tapping the object on the table and even tapping the object until they finally considered twisting it. Furthermore, there was a lack of system visibility and feedback to the users: when the scale of the map was changed, users did not see the extent of the zooming in our out as there was no point of reference (i.e., scale indicator on the map).

4.3 Working with More Than One Data Layer

After creating a map with one data layer and exploring how to zoom and pan with the interface, the next logical task was to create a map with five different data layers, switching on the reference map (called full map on the object) and then to develop a visual hierarchy of the layers by prioritising the order the layers are displayed. Adding more layers had a 100 % completion rate. All participants added new layers with ease and confidence. P2, for instance, describes it as being "*straightforward*". They adopted

the same procedure, consisting of (1) reading the labels of the objects lying on the border of the table, (2) grasping the required object and (3) putting it onto the interactive surface. In case they could not spot the right object at the border, participants started to look at the ones placed on the map, then touching or pointing on the object in order to signal the step was complete. This approach could, for instance, be observed for P1: whilst search for the correct object P1 looks first at the objects lying on the border, then looks at the ones lying on the interactive surface. He finds the right layer, and moves it slightly, says "*Shipping… ..ok*!". It was easy to both learn and then remember how to do it a second time.

There were two distinct approaches to completing this task. Four of the participants repeat this procedure in a quick fashion until all layer objects are lying on the table and their information displayed. The remaining four carefully analyse the displayed information after having placed an object on the surface. An example is provided by P2, who immediately reads and analyses the legend after placing an object on the table: "*Ok, so now that's useful, now I understand why it is important to show it here*". Then she tries to identify the information on the map, and finally uses the zoom in order to view more detail on the map: "*at this zoom level it is not very legible…. but if I zoom in I suppose I will be able to see more… ok, yes*".

All participants hesitated on to where to place layer objects on the table. One assumption was that layer objects need to be placed into an area of interest. This was also observed during the first basic map creation. Participants were observed placing the objects at arms length, just below the optical centre of the map. Additional confusion was created due to the fact that some layers required several seconds to load, and there was no feedback to tell the users that data was refreshing. P4, for example, misinterprets the valid positions of the layer object he expects that the road layer object needs to be placed "*on the ground*". P4 takes the road object from the border, "*Let's create a map with roads*", and first makes a movement to place it inside UK, but then places it west of UK, inside the sea, waits a short moment, while looking at the map. Nothing is displayed, so he lifts it again. "*Maybe I should place it on the ground somewhere",* and places it inside France. Roads are displayed immediately… *"yes!*".

Switching on the reference map, labelled full map on the object, by twisting the country border object turned out to be straightforward for some participants whereas others needed additional time or explanations to explore different manipulations. This was surprising as all had previously used a twist of an object to change the map state when zooming. They were not expecting layer objects to have more than one layer.

Prioritising layers was easily completed by three users (see Fig. 5c), whilst five had trouble figuring out how to proceed. Participants had to place the objects in priority order, the object at the top of the map was the most visually important and the one nearest the bottom less so. Several participants intuitively stacked the objects to prioritise the layers. P4, for instance, stacks them with the first layer of the list lying at the bottom (see Fig. 5a), then, being unhappy with the results, creates a second stack with the last layer of the list lying at the bottom.

An alternative approach was provided by P5 who notices small arrows on the labels, and interprets the need of putting the layers side by side (see Fig. 5b),

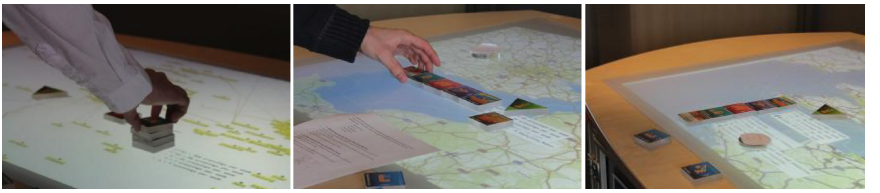


Fig. 5. Two approaches for prioritizing layers: a) stacking objects, b) putting them side by side (c) vertical array of objects

"*I'm trying now to have an idea how to set the order with the different objects. Now I realize that maybe the little symbol on the objects is like the order of the priority of the different objects in the map*"(P5). This demonstrates the labeling and the symbology of the objects can be misinterpreted as the arrow symbol is to show which layer is activated. Furthermore, participants were unsure whether they solved the task of prioritizing layers correctly as they were expecting additional feedback on the priorities. *"I'm not sure if we need to see that I gave a hierarchy on that, because we can't see that on the map. So I guess something is missing, but I don't know what."*

4.4 Interpreting Layers and Legends

After creating a visual hierarchy participants were asked to interpret meaning from the map. They were asked to identify, between France and UK, the shipping route with the most, and the least amount of traffic. At this stage, most of the participants were already familiar with the legend as they were noticing it in previous tasks. When a layer object is placed on the table, the legend appears to the right of the object. Participants read the legends and tried to interpret the information. We observed differences in how participants organized their workspace for solving this task. P3, for instance, works on the very left side of the table, preventing her to see much of the surroundings. However, she is not panning the map, nor mentioning a limitation in her field of view. In contrast, other participants make use of known features to create a map allowing them to best see the required information. For instance, P2 moves objects to the side and switches between full map and country borders to have a switch the background to more much detailed reference map view: "*Between… France and the UK…Let's move this here…*" She lifts and drops the line of objects one by one to the right. "S*o that I can see something… let's turn so that I see country borders*" She turns the object, looks at the map "*ula*!" Then she turns it back to full map. She repeatedly looks back and forth onto the map from different perspectives, and onto the legend of the info object.

A common procedure was also to remove unnecessary information to obtain a better view. For instance, P6, after a first look on the map, decides to remove some of objects. "*I will bring out everything which I don't use, to have something a bit more clear.*"(P6). Nevertheless, all participants were unable to locate the shipping route with the least amount of traffic. Participants were mentioning difficulties in seeing the difference between the lower types of ferries. This illustrates the need to improve the

cartographic representation of the data layers and the information and visualisation of the legends.

4.5 Requesting Additional Information

Finally, participants were asked to find out the names of the ports using the info tool. This task turned out to be particularly challenging for the participants. Only three participants identified the correct information (38 % completion rate, very low), however, even those who completed the task took a long time (average time was 03:53).

P8, for instance, first places it pointing down onto the centre of the shipping route, waits a short moment, and then turns it into the other direction, pointing upwards onto the centre of the route. Then she points onto one end of the route, and the other end. Then she taps the object. She is prompted to explain her action. While answering the question she replaces the object and a window opens.

This issue can be explained by the fact that the table sends the request for information after a short moment when there is no action on the object, and provides no hint when exactly this is done. Users were expecting immediate feedback, and since this was not provided immediately after an action, users have concluded that is was not the correct manipulation, thus there was a lack of feedback to the user.

5 Discussion

The research marks an initial study to explore the usability of geospatial tangible interfaces, the observations from which we are able to reveal how 8 participants discover a new interface for the very first time and to unravel the interaction patterns associated with ease of use and learnability. To provide insight into the observed usability issues we have classified the issues into three themes: (a) understanding cartographic elements on tangible tables; (b) Object manipulations (c) Use of non-responsive “offline spaces”

5.1 Understanding Cartographic Elements

The spatial cognition of maps is related to the way in which participants understand geographical data and information with a view to developing meaning. Therefore, one aim of the user study is to work towards improving spatial cognition of geo-tangible applications with improved use and design of geographic data and interfaces, which depend on human interpretation [29]. To aid the general spatial cognition of maps, there are a number of well-defined map elements that should be integrated into the digital mapping interface and then these interactions should be aligned to users’ mental models. The use of these simple mapping conventions and adopting metaphors for encoding interactions can improve process of engaging with the data and the subsequent interpretability and understanding of the information displayed.

Improving understanding can also be enhanced through making the interactions of the cartographic elements within the interface more efficient by ensuring an alignment

between the users' perception of how the different functions should work and how they have been designed to work. This type of efficiency is based on the notion of interface congruence [40] and can really aid the human-computer interaction experience. Therefore, it is important to harness the power of users' conceptual models to translate into interactions that build upon easy to understand everyday cultural metaphors.

The current implementation of the zoom interaction we observed is akin to a traditional volume dial where a clockwise twist increases the sound but in this case the map zoom level increases (zooms in) – the same interaction type is found in some in car Sat Nav/GPS systems. The interaction of the object in this form may be on conventional standards, using the dial to represent more or less, it does not align naturally with the users' mental models for navigating maps. The results suggest there are more effective methods of encoding zoom functionality for such geospatial tangible interface. Let us consider the zoom interaction on a desktop mapping package which can be accessed in two ways. For example, in QGIS the user is required to either (1) select the tool icon zoom in or zoom out (which is often represented by a magnifying glass) and then click in the desired area of the map or (2) use the mouse scroll wheel to zoom in (mouse scroll wheel forwards) or out (mouse scroll wheel backwards). When you want to see more information you roll the scroll wheel forwards – it represents moving towards the area or away from the area when you move the scroll wheel backwards. This interaction is not the most visible but it nicely encompasses the users' mental model of near (moving towards) and far (moving away from). The use of the scroll wheel is not as transparent. By comparison the interaction in webmap such as Googlemaps the user interacts with 2 buttons which are placed on top of each other: a plus (+) and a minus (−). The plus provides the user with a more detailed map by zooming in and the minus the opposite, with the more information on the top and less on the bottom – working on the metaphor more is up and less is down. For future implementations it would make sense to build on these cultural metaphors of distance (near and far) and orientation (top -bottom and up-down) to enhance the congruency of the user's mental models and how there are implemented in a tangible interface.

- More closely match user's cultural and mental models to the object interactions for zoom navigation controls.
- Consider mixed modal interfaces that combine touch interactions for zooming and panning with object interactions for adding layers etc.

Currently, the zoom object is unrestricted. Participants were able to zoom indiscriminately in or out of the map. On a number of occasions we observed users zooming so much that they ended up at the bottom of the ocean, with no clear understanding associated with what or why they were just seeing no data. This used to be an issue with web maps implemented prior to the slider scale interaction that is now commonplace. A lack of issues associated with the scale of the map and the zooming and panning interactions can be ameliorated with the following simple guidelines:

- Restrict the zoom capacity according to the scenario's context and common sense.
- Display feedback reflecting a change in the zoom made by a user by indicating map scale.

- Provide an inset map so users can see how and where they are navigating to on the map. Perhaps this could be in the form of a new object or could encoded within existing zooms or pan objects and activated automatically.
- Provide a tangible to reset the map zoom.
- Provide tighter control between the two functionalities of zoom and pan.

The essential map elements include scale and direction, title, inset maps and use of legends. We observed a number of issues associated with either an absence of or interaction difficulties with these cartographic elements. At present only the use of the legend has been implemented on the geo-tangible interface. This absence of fundamental map elements led to confusion and reduced the ease of use of the interface and reduced understanding and ability to interpret the geographic information. These types of elements are not new and have been clearly defined from research with paper maps but what is required now is in depth understanding on how they can be integrated into the TUI to ensure they enhance users' geographic understanding.

The legend is one of the most important map elements. Without a useful and meaningful legend users are less able to develop spatial cognition. It is one of the key mapping components to aid knowledge construction. As with many digital map legends, the legend details can be derived by default from the layer information, but default labels often reduce the ability to interpret data classifications. The legends used in this scenario were developed by default from the external projects database. Therefore, the legends fell foul to system naming defaults and labelling that made sense only to the technical developers. Furthermore, the automatic display of the legend to the right of the object once it was placed on the table lead to some misunderstanding and misinterpretation. For example, when only one piece of information was displayed, some users did not realise that this information was part of the legend and interpreted it as a further location on the map. We can explain this by the fact that there was no information to identify the symbol and label as the legend. The following suggestions would improve spatial cognition of legends:

- Explore impact of projecting the legend in the offline space or integrate it within a separate device (further work would be required to investigate impact this has on cognitive load).
- Differentiate the legend from the map using neat lines and titles and visual cues or use object (shake, or turn for example) to switch layer on or off.
- Design legends for the user: labels and text should reflect their mental models. Legend data classes should be rounded to whole numbers and arranged vertically with lowest numbers at the bottom. The textual descriptions for the data classes should add meaning for the user.
- Enable the ability to switch the legend on or off.

Conceptually, the original idea of map layers symbolise of a filing cabinet comprised of different drawers each containing files and folders for different themes. With this mental model in mind it seems natural that users thought that they could stack the tangible objects on top of each other. Thus when working with layers it is necessary to consider:

- The objects and how they represent everyday metaphors to better replicate the real world understanding of users – consider new forms of objects that can be stacked on top of each other.
- Enable users to filter information within different layers by switching on or off different classifications groups.
- Automatically change the cartographic styles of the layers based on where they are positioned in the visually hierarchy. For dynamic maps, the map title should say something about the general topic and indicate the purpose of the map. If not included in the object layer's legend(s) it should say something about the temporal nature of the data being mapped.

5.2 Object Manipulations

Particular to interactions on TUIs are the physical manipulations with tangible objects. In previous work, the mapping of physical objects to digital information has been seen as central [7, 33, 39], and aspects related to, for instance, embodiment, metaphor, location, or correspondence have been discussed. More recently, the physical objects and the associated movements have been considered in a broader scope. Fernaeus et al. [5] claim that the role of tangible artefacts goes beyond manipulating digital information and has impacts on a wider range of activities related to perception and sensory experience, physical manipulation, and referential, social and contextually oriented action.

In our analysis, we found that the ways in which the participants interacted with the object varied considerably. We observed participants frequently shifting, dropping and lifting the objects. Some used stacking, twisting, tapping to try and instigate a change in the map, initiate, or cancel an action. This suggests that in future versions of the interface, more could be made of these natural interactions by encoding functionality, and thus better conform to user's natural expectations.

Also observed were very different ways of manipulating the objects. Some work with two hands, other with one. Some prefer to lift and drop, others shift the object slowly around the table. Some make a lot of quick and short movements, other read and reflect a lot, then make only few considered movements. A larger study would most likely reveal more insight into the different working preferences of different types of people, but in the context of this study the tangible table needs to be able to support these different working styles and preferences towards manipulating the map with objects.

The current implementation of the interface was impacted by the different working practices which led to unexpected changes in the map state for the user: data not loading when quick movements were made, the map moved unexpectedly when users tried to move the zoom and pan object out of the area of interest on the map. The following guidelines could reduce the impact of unexpected results occurring from different working styles with the objects.

- Provide hints and tips to get started –a brief help video could be shown if the users touch the table when no objects are on it.

- Provide user feedback if an object is moved too quickly, like: “I think you are trying to move the map, try again but slower”.
- Enable users to go to their previous view by providing an object or turn/action with an object
- During panning, restrict the object to deactivate the zoom action.
- When the result of a geospatial interaction cannot be provided immediately during the manipulation, provide a visual feedback ensuring the user that the manipulation was correct and that something is happening. For instance, a data refreshing symbol.
- Provide clear labelling of interactions on the objects - arrow to show the need and direction of rotation for zoom.
- Ensure objects with different functions are uniquely differentiated, enabling them to be easily recognized. Make use of shape, colour, sizes and heights and group objects of similar functional types accordingly.

Where it makes sense consider the use of objects that represent everyday metaphors (e.g., toy cars for a road layer or trains for the railway network).

Further work is required to provide logical connections between the tangibles, the interface and the consequential changes to the interface that occur due to an interaction

5.3 Non-responsive (“Offline”) Spaces

One of the inherent properties of TUIs is the intense combination of the physical and the digital, forming a hybrid interaction space. This enables a rich range of interactions including a large number of actions being done offline, on non-responsive spaces. As already previously observed, TUIs enable an ‘extra layer of interaction’ on spaces that are not recorded into the system [42]. In collaborative settings, these spaces were noticed to be used to make suggestions, demonstrate next steps, or set a common focus [43]. [2] claim the support for offline activities being one of the major benefits of TUI. They describe how the hands and edges are used as offline repositories in single and paired settings. They were used to support cognitive actions and reduce mental load, hence complement task completion.

We have made similar observations with participants of the geo-tangible interface. In order to fulfill their tasks, participants were making use of offline interactions to reduce their mental load. In particular, observed offline interactions were related to organizing the workspace in order to have a better view on the map, as for instance, P2 who was shifting layer objects outside her field of view. Another type of offline interactions were used to aid cognition in the stepwise following of the tasks, i.e. P1, who was touching the object layers as soon he has found them. Finally, we saw that participants used non-responsive spaces to adopt another perspective. P2 was bending herself multiple times between two positions, as well as P6 who was leaning himself onto the border of the table while he felt stuck. Thus, the offline space is an important feature of the geo-tangible interface, therefore it is important that the table has the following features:

- Support a change of perspective: enable users to make a few steps, bend and stretch themselves, or lean against the table. The tabletop should provide a good view from different positions and support actions not only at the middle of the table, but also on the sides.
- Allow users to customize their views: provide a non-reactive area where objects can be placed when removed, consider providing dedicated repositories where users can place objects of different types – to aid relocating of the objects for future use.
- Also enable objects to be placed on different positions on the interactive surface to support users in customising their view. Non-reactive touching: allow for touch interactions that have no effect in the system, hence allowing the users to use them for externalising their cognition.

6 Conclusion, Limitations and Future Work

This paper has presented the first usability study of our geo-tangible mapping interfaces. It was conducted using established methodological practices designed around the completion of predefined tasks and the think-aloud protocol with collected data analysed using video analysis. The result is a descriptive review of the spatial interactions unique to geographical data and the basic controls required for interacting with such special data.

The qualitative study has provided a first set of insights into how the novice user explores tangible interfaces to carry out specific tasks. On first use with the objects, experiment participants were cautious and object movements were hesitant as they were uncertain of the interface. However, we observed all participants quickly becoming confident with using the objects to manipulate the map, with various different working styles emerging. Indeed, in the authors' experience, it would not be possible to learn so quickly to use a conventional desktop mapping application. A comparison with which would be a suitable topic for a further study.

Based on this observation, we can conclude that the geo-tangible user interface is particularly useful in situations involving lay users. Typically such situations appear in participatory approaches, such as participatory urban planning. A geo-tangible table could improve communication between heterogeneous stakeholders by, on one hand, allowing experts to explain geospatial phenomena to novices, and, on the other hand, supporting novices in sharing an own perspective with the expert. We also see its potential for the development as a teaching and learning platform for younger audiences. As interaction is simplified in TUI scenarios, complex GIS manipulations will be limited. So this approach is less useful for situations purely implicating geospatial expert users.

The results of our analysis highlight the necessity to consider three different dimensions in the design of geospatial tangible tables: cartographic elements, object manipulations, and non-responsive spaces. Observed issues dealt with the lack of cartographic elements and cultural metaphors which would enhance geographic meaning. We also observed a series of issues related to feedback in general. Although learned object movements could be easily repeated, i.e. they are memorable, they appeared

hidden to the users at the beginning. Better and timely feedback, informing the user of what is happening within the system interface, would allow him/her to appropriate the interactions more effectively.

We have formulated initial guidelines for the design of geospatial tangible tables to ensure their ease and straightforward to learn and use. In future work, we hope to investigate the most intuitive and effective use of tangibles for geographic interactions and understand how different types of objects and their interactions can be optimized for geospatial TUIs.

This study shows the real usefulness of user studies to establish guidelines for the development of novel interfaces such as the interactive tangible table. Our multidisciplinary approach combing the expertise of mapping science and TUI design considers the unique nature of both GI and TUIs. To successfully interact with such a system, special interactions are required, that, on one hand, build upon fundamental principles and, on the other hand, make use of new possibilities of emerging technologies. This requires dedicated user studies to ensure associated interfaces are both useable and useful. Improving the ways in which users interact with such systems can only aid the development of new learning technologies towards a new generation of natural interfaces.

Acknowledgments. This work has been partially funded by the Interreg IVb project Weastflows and the Luxembourg Institute of Science and Technology (LIST). We would like to extend our thanks to all the experiment participants who provided excellent data on which we could base this study.

References

1. Djajadiningrat, T., Wensveen, S., Frens, J., Overbeeke, K.: Tangible products: re-dressing the balance between appearance and action. Pers. Ubiquit. Comput. **8**(5), 294–309 (2004)
2. Esteves, A., Michelle S., Ian O.: Supporting offline activities on interactive surfaces. In: Proceedings of TEI 2013 (2013)
3. Faulkner, L.: Beyond the five-user assumption: benefits of increased sample sizes in usability testing. Behav. Res. Methods Instr. Comput. **35**(3), 379–383 (2003)
4. Fernaeus, Y., Jakob, T.: Looking at the computer but doing it on land: children' s interactions in a tangible programming space. In: McEwan, T., Gulliksen, J., Benyon, D. (eds.) People and Computers XIX — The Bigger Picture, pp. 3–18. Springer, London (2006)
5. Fernaeus, Y., Tholander, J., Jonsson, M.: Towards a new set of ideals: consequences of the practice turn in tangible interaction. In: Proceedings of TEI 2008, pp. 223–230. ACM (2008)
6. Fischer, G.: External and sharable artefacts as sources for social creativity in communities of interest. In: Proceedings of the 5th International Roundtable Conference: Computational and Cognitive Models of Creative Design, Australia, 9–13 December 2002
7. Fishkin, K.: A taxonomy for and analysis of tangible interfaces. Pers. Ubiquit. Comput. **8**(5), 347–358 (2004). doi:10.1007/s00779-004-0297-4
8. Fuhrmann, S., et al.: Making useful and useable geovisualization: design and evaluation issues. Exploring Geovisualization, 553–566 (2005)
9. Goodchild, M.F.: Twenty years of progress: GIScience in 2010. J. Spat. Inf. Sci. **1**, 3–20 (2010)

10. Goodchild, M.F.: Spatial thinking and the GIS user interface. Proc. Soc. Behav. Sci. **21**, 3–9 (2011)
11. Haklay, M., Tobón, C.: Usability evaluation and PPGIS: towards a user-centred design approach. Int. J. of Geograph. Inf. Sci. **17**(6), 577–592 (2003)
12. Hornecker, E., Buur, J.: Getting a grip on tangible interaction: a framework on physical space and social interaction. In: Proceedings of the CHI 2006 (2006)
13. Hwang, W., Salvendy, G.: Number of people required for usability evaluation: the 10±2 rule. Commun. ACM **53**(5), 130–133 (2010)
14. Ishii, H., Ullmer, B.: Tangible bits: towards seamless interfaces between people, bits and atoms. In: Proceedings of CHI 1997, 22–27 March 1997
15. ISO/TR 16982:2002: ISO/TR 16982:2002: "Ergonomics of human-system interaction – usability methods supporting human-centred design," International Standards for Business, Government and Society
16. Jones, C.E., Haklay, M., Griffiths, S., Vaughan, L.: A less-is-more approach to geovisualisation – enhancing knowledge construction across multidisciplinary teams. Int. J. Geograph. Inf. Syst. **23**(8), 1077–1093 (2009)
17. Jones, C.E.: Practical cartography. In: Haklay, M. (ed.) Interacting with Geospatial Technologies. Wiley, New York (2010)
18. Jones, C.E., Weber, P.: Towards usability engineering for online editors of volunteered geographic information: a perspective on learnability. Trans. GIS **16**(4), 523–544 (2012)
19. Landauer, T.K., Nielsen, J.: A mathematical model of the finding of usability problems. In: Interchi 1993, ACM Computer–Human Interface Special Interest Group (1993)
20. Longley, P., Goodchild, M., Maguire, D., Rhind, D.: Geog. Information Systems and Science. Wiley, Chichester (2010)
21. MacEachren, A.M., Kraak, J.A.: Research challenges in geovisualization. Cartography Geograph. Inf. Sci. **28**, 3–12 (2001)
22. MacEachren, A.M., Cai, G., Sharma, R., Rauschert, I., Brewer, I., Bolelli, L., Wang, H.: Enabling collaborative geoinformation access and decision-making through a natural, multimodal interface. Int. J. GIS **19**(3), 293–317 (2005)
23. MacEacheren, A.M., Cia, G., Brewer, I., Chen, J.: Supporting map-based geocollaboration through natural interfaces to large screen displays. Cartographic Perspect. **54**, 16–34 (2006)
24. Maquil, V.: Towards understanding the design space of tangible user interfaces for collaborative urban planning. Interact. Comput. (2015). doi:10.1093/iwc/iwv005
25. Maquil, V., De Sousa L., Leopold U., Tobias E.: A geospatial tangible user interface to support stakeholder participation in urban planning. In: Proceedings of GISTAM 2015 (2015, to be published)
26. Maquil, V., Ras, E.: Collaborative problem solving with objects: physical aspects of a tangible tabletop in technology-based assessment. In: COOP 2012 (2012)
27. Marsh, S.L.: Using and evaluating HCI techniques in geovisualization: applying standard and adapted methods in research and education. In: Proceedings of GIS Research UK, pp. 33–38 (2008)
28. Marshall, P.: Do tangible interfaces enhance learning. In: Proceedings of TEI 2007, Baton Rouge, LA, USA (2007)
29. Montello, D.: Spatial cognition. In: Smelser, N., Baltes, P. (eds.) International Encyclopedia of the Social and Behavioural Sciences, pp. 14771–14775. Pergamon Press, Oxford (2001)
30. Nagel, T., Maitan, M., Duval, E., Moere, A.V., Klerkx, J., Kloeckl, K., Ratti, C.: Touching transport-a case study on visualizing metropolitan public transit on interactive tabletops (2014)
31. Nivala, A.M., Brewster, S., Sarjakoski, T.L.: Usability evaluation of web mapping sites. Cartographic J. **45**(2), 129–138 (2008)

32. Parsons, E.: The map of the future may not be a map! Cartographic Rev. **50**(2), 182–186 (2013)
33. Price, S.: A representation approach to conceptualizing tangible learning environments. In: Proceedings of TEI 2008, p. 151. ACM Press, New York
34. Roth, R.E.: Interactivity and cartography: a contemporary perspective on user interface and user experience design from geospatial professionals. Cartographica **50**, 2 (2015)
35. Schneider, B., Jermann, P., Zufferey, G., Dillenbourg, P.: Benefits of a tangible interface for collaborative learning and interaction. IEEE Trans. Learn. Technol. **4**(2), 222–232 (2011)
36. Scott, S., Allavena, A., Cerar, K., Mcclelland, P., Cheung, V.: Designing and assessing a multi-user tabletop interface to support collaborative decision-making involving dynamic geospatial data. Technical report, University of Waterloo, Canada (2010)
37. Shneiderman, B., Plaisant, C.: Designing the User Interface, 4th edn. Pearson, Addison Wesley, USA (2005)
38. Sui, D.: The wikification of GIS and its consequences: or Angelina Jolie's new tattoo and the future of GIS. Comput. Environ. Urban Syst. **32**, 1–5 (2008)
39. Ullmer, B.: Emerging frameworks for tangible user interfaces. IBM Syst. J. **39**, 915–931 (2000)
40. Tversky, B., Morrison, J.B., Betrancourt, M.: Animation: can it facilitate? Int. J. Hum Comput Stud. **57**, 247–262 (2002)
41. Jones, C.E., Maquil, V.:. Twist, shift, or stack? Usability analysis of geospatial interactions on a tangible tabletop (2015)
42. Fernaeus, Y., Tholander, J., Jonsson, M.: Towards a new set of ideals: consequences of the practice turn in tangible interaction. In: Proceedings of TEI 2008 (2008)
43. Maquil, V., Ras, E.: Collaborative problem solving with objects: physical aspects of a tangible tabletop in technology-based assessment. In: Proceedings of COOP 2012 (2012)

Integration of a Real-Time Stochastic Routing Optimization Software with an Enterprise Resource Planner

Pedro J. S. Cardoso[1,2(✉)], Gabriela Schütz[1,3], Jorge Semião[1,5], Jânio Monteiro[1,5], João Rodrigues[1,2], Andriy Mazayev[4], Emanuel Ey[1], and Micael Viegas[1]

[1] Instituto Superior de Engenharia, University of the Algarve, Faro, Portugal
{pcardoso,eevieira,mimviegas}@ualg.pt
[2] LARSys, University of the Algarve, Faro, Portugal
jrodrig@ualg.pt
[3] CEOT, University of the Algarve, Faro, Portugal
gschutz@ualg.pt
[4] Depart. de Eng. Eletrónica e Informática, University of the Algarve, Faro, Portugal
amazayev@ualg.pt
[5] INESC-ID, Rual Alves Redol, 9, Lisbon, Portugal
{jsemiao,jmmontei}@ualg.pt

Abstract. In order to manage their activities in a centralized manner, an Enterprise Resource Planning (ERP) software is a fundamental tool to many companies. As a generic software, many times it's necessary to add new functionalities to the ERP in order to improve and to adapt/suite it to the companies' processes. The Intelligent Fresh Food Fleet Router (*i3FR*) project aims to meet the needs expressed by several companies, namely the usefulness of a tool that makes "intelligent" management of the food distribution logistics. This "intelligence" presupposes interconnection capacity of various platforms (e.g., fleet management, GPS, and logistics), and active communication between them in order to optimize and enable integrated decisions.

This paper presents a multi-layered architecture to integrate existing ERPs with a route optimization and a temperature data acquisition module. The optimization module is prepared to deal with dynamic scenarios, as new demands may appear during the optimization process and the routes will admit several states (e.g., open, locked and closed), according with the ERP manager instructions. The data aquisition module implements the retrieve of some vehicles parameters (e.g., chambers' temperatures and vehicle's global positioning system data), used to validate the routes and provide information to the company's manager.

A distribution company was selected as case-study, having up to 5000 daily deliveries and a fleet of 120 vehicles. The integration of the developed modules with the company's ERP allowed the maintainance of most of the existing procedures, avoiding routines disruption.

C. Grueau and J.G. Rocha (Eds.): GISTAM 2015, CCIS 582, pp. 124–141, 2016.
DOI: 10.1007/978-3-319-29589-3_8

Keywords: Enterprise resource planning · Vehicle routing problem · Geographical information · Application programming interface · Data acquisition

1 Introduction

i3FR (Intelligent Fresh Food Fleet Router) is an ongoing project of the University of the Algarve and X4DEV Business Solutions, with the main objective of building a system to manage and to optimize the distribution of fresh products by private fleets. The system will be integrated with an existing ERP, namely the SAGE ERP X3 [22], and will compute routes from the depots to the delivery points using cartographic information and taking into consideration multiple objectives.

The intelligent management of the distribution fleet is assured not only by optimizing costs, but also by monitoring what is being distributed. In the case of refrigerated vehicle fleets, for the transportation of fresh and frozen goods, the specificity is high. In this particular case it is necessary to maintain the quality of the products, to satisfy the legal criteria for food transportation, as well as the goods' safety and hygiene as required by suppliers and customers.

The product being developed can be divided in three main components: (1) a data acquisition system for the distribution vehicles; (2) an intelligent routing optimization platform, and (3) an Application Program Interface (API) to integrate the new modules with the existing ERP.

The first component will acquire statistical data from the vehicles (e.g., fuel consumption and speeds obtained from "control units"), from the real delivery tours (e.g., using GPS data or track obstructions reported by drivers), along with data from wireless sensors placed inside and outside the transportation chambers. In the case-study, the vehicles' transportation chambers are divided into three subtypes, namely: frozen, chilled and ambient/dry goods categories. The acquired data is then used in the second component, which optimizes the routes, taking into account multiple constraints (e.g., vehicle capacities, time windows for deliveries, route duration, balance between route duration and driver working periods) and objectives (e.g., minimize the number of routes, minimize the total distance, minimize the total time to perform the deliveries and maximize customers satisfaction). The route computation will be endowed with intelligence based on data acquired during previous routes, leaving open the possibility of integrating other sources. The third module is fundamental as a bridge between the existing ERP and what is proposed to be conducted in the *i3FR* project.

Some related works and products exist on the market. An algorithm for the distribution of fresh vegetables in which the perishability represents a critical factor was developed in [17]. The problem was formulated as a VRPTW with time-dependent travel-times (VRPTWTD), where the travel-times between two locations depend on both the distance and the time of the day. The problem was solved using a heuristic approach based on the Tabu Search [7]. The performance of the algorithm was verified using modified Solomon's problems. A somehow

similar work was proposed in [26], which deals with distribution problem formulated as an open multi-depot vehicle routing problem (OMDVRP) encountered by a fresh meat distributor. To solve the problem, a stochastic search metaheuristic algorithm, termed as the list-based threshold accepting algorithm, was proposed. In [2] was considered a generalization of the asymmetric capacitated vehicle routing problem with split delivery. The solution determines the distribution plan of two types of products, namely: fresh/dry and frozen food. The problem was solved using a mixed-integer programming model for the problem, followed by a two-step heuristic procedure. In [1] the distribution of fresh vegetables was addressed. The focus of the study was the delivery of fresh vegetables selecting the best routes particularly for urban areas such as Kuala Lumpur city, which faces traffic problems. There are also a relatively large number of companies providing commercial software which is similar to the one developed in the *i3FR* project [10,13–15,20].

The main contribution of this paper is the proposal of a multi-layered architecture to integrate existing ERPs with a route optimization and a temperature data acquisition module. The optimization module is prepared to deal with dynamic scenarios, where new demands may appear during the optimization process, which will integrate them in the iteration procedure. Furthermore, computed routes will have several states, e.g., (1) "open" meaning that new orders can be added, removed or swapped inside that route, (2) "locked" in which case new orders may still be inserted but swapping is no longer allowed, and (3) "closed" meaning that no more changes can be made to the route. Furthermore, the system is prepared to acquire geographical data from several sources, namely from Google Maps, from an Open Street Maps router and from data retrieved from previous deliveries/routes. The system is also prepared to accept several optimizers, if necessary. All this maintaining the company's core ERP software.

A distribution company of fresh, frozen, ultra-frozen and dried products was selected as case-study, having up to 5000 daily deliveries and a fleet of 120 vehicles. Therefore, the system must be able to manage thousands of customers, each with different demand characteristics, matching the company's delivery resources with the customer's delivery requirements. The use of the company's ERP allows for maintaining most of the existing procedures, which avoids disruption to the company's routines.

The remaining document is structured as follows. Section 2 presents the steps necessary to integrate the existing ERP with the new modules, namely the Optimization module (*i3FR-Opt*), the Hub module (*i3FR-Hub*), the Database module (*i3FR-DB*) and the Maps module (*i3FR-Maps*). The section will also include some of the results returned by the system. The data acquisition module is addressed in Sect. 3. The fourth and final section presents some conclusions.

2 Steps in the Integration of an ERP with a VRP Optimization Module

A typical enterprise has many departments or business units (e.g., sales, inventory, finance, and human resources departments), that have to be continuously

communicating and exchanging data with each other. For instance, whenever the sales department makes a sale, it has to check the inventory department and send information relative to the sale to the finance department. In the middle is the human resources department, which will be informed to ensure man power to process the picking and sending of the items. Therefore, the success and efficiency of the organization lies, although not exclusively, in the successful communication between those departments.

A decentralized system would maintain data spread through the local departments. The local departments, in general, do not have access to the data of the other departments, which turns obtaining real data (e.g., stock inventory) into a time-consuming procedure, leading to inefficiency, loss of revenues and possible customer dissatisfaction. Simultaneously, the lack of communication makes it hard to do integrated business intelligence procedures, due to disparate, redundant and, most probably, inconsistent enterprise data.

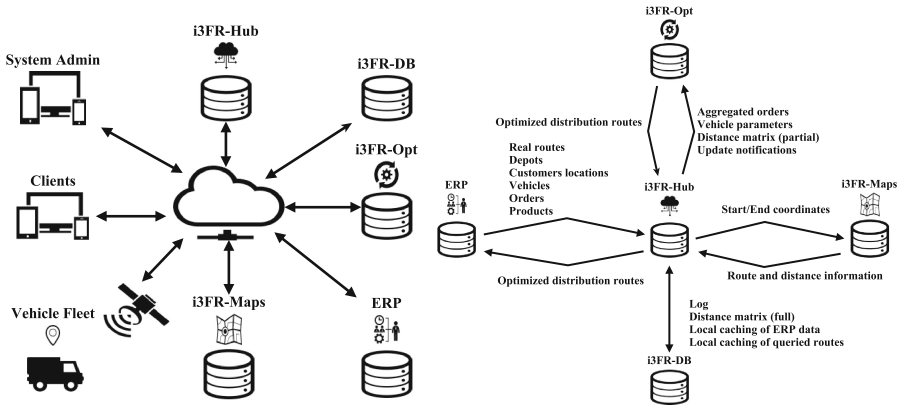
In a centralized system, often simply called ERP, data is maintained at a central location and shared with other departments in real time. Sensitive data sharing is supported on views to the various users and departments. This strategy provides more accurate data by removing redundancy and real time updates.

In our case, the developed work was made for a SAGE ERP X3 management system [22], which stores all vital information of the corporation activities, like customers, orders, suppliers, vehicles, employees, etc. The corporation management is done using interfaces/forms properly developed for the retrieving and filling of data stored on the ERP database.

The *i3FR* Routing System, used to compute the distribution routes, was designed to be independent of the ERP, allowing future integration with other ERP systems or even autonomous operation, given the implementation of the proper interfaces. Under that independence perspective, the communications between the systems were implemented as web services. More precisely, a Web API with simple representational state transfer (REST) based communications was considered [19]. The advantage is in the fact that RESTful APIs do not require XML-based Web service protocols to support their interfaces, which in the present case was done using JSON [9].

The web services on the ERP side, provide the necessary data through GET requests/methods (e.g., customers addresses, orders, available vehicles, and vehicles capacities) and receive the results generated by the *i3FR* Routing System through POST methods (e.g., computed routes).

The *i3FR* Routing System is composed of several interdependent submodules which also communicate via HTTP, thus allowing for the system to be scaled from a single-machine environment to a large multi-server production deployment. In more detail, the system is composed of the following modules: (a) *i3FR-Opt* – The actual route optimization algorithm is implemented in the optimization module designated *i3FR-Opt*. The optimization process relies on heuristics and meta-heuristics [3] that should run in real-time, taking into consideration the dynamics of the overall system (more details in Sect. 2.1); (b) *i3FR-Hub* – All communications occurring inside the *i3FR* Routing System go



(a) System's global diagram (b) Inputs and outputs of *i3FR-Hub*.

Fig. 1. System's global diagram and communications between modules.

through *i3FR-Hub* which acts, as the name suggests, as a hub. For instance, the optimization module, *i3FR-Opt*, sends and receives all the data it requires through a RESTful API provided by the *i3FR-Hub*. Furthermore, all communication to the external systems also goes through this hub, which acts as a "gateway", a security access provider, and a data validator (more details in Sect. 2.2); (c) *i3FR-DB* – The *i3FR* Routing System operates on top of a non-relational database. The database provides local storage of information relevant to the optimization procedure, namely: data retrieved from the ERP (avoiding overloading the system with constant requests), cartographic information, computed routes, etc. (see Sect. 2.3); and (d) *i3FR-Maps* – A cartography subsystem was also implemented. The system retrieves routing informations from other systems (e.g., Google Maps) and stores it on the *i3FR-DB* (see Sect. 2.4).

The overall system is depicted in Fig. 1(a) which includes, besides the presented modules, interfaces developed on the side of the ERP, the ERP, and the connections to the data acquired (e.g., real routes and chambers temperatures) and send to the vehicles of the fleet (e.g., the optimized routes).

In the next sections, we will explain in more details each of the developed modules.

2.1 *i3FR-Opt* Module

i3FR-Opt module is the optimization unit, responsible for finding a suitable set of routes for the distribution of fresh goods. The optimization procedure computes routes from one or more depots that visit each customer once, within given time intervals, without violating the vehicle capacities, and other legal constraints (e.g., maximum consecutive driving time). The most similar problem to the one to be solved is the Vehicle Routing Problem with Time Windows (VRPTW), common to the majority of the distribution fleets [3,5,6,8,24].

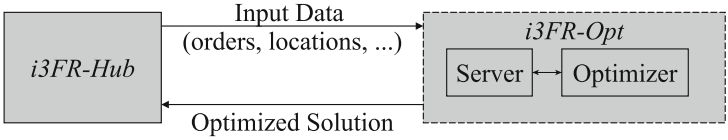


Fig. 2. Communication between Hub and Opt modules.

In its simplified form, the VRPTW can be stated as the problem of designing an optimum set of routes that deliver a set of goods, to a set of customers, within predefined time windows.

Within the VRPTW perspective, for the time being, the main objectives of the computational process are to obtain solutions which minimize the number of necessary delivery vehicles/drivers, minimize the total delivery distance, and maximize the minimum load between all vehicles. These objectives optimize the operational costs (e.g., fuel and worker) and give more equilibrated routes, in terms of man workload.

The optimization module can be seen as a composition of two sub-modules (see Fig. 2). The *server* sub-module is responsible for the communication with the external services, in particular with the *i3FR-Hub* module (see Sect. 2.2). This sub-module acts as a server, in the sense that it waits for notifications from the hub. Whenever a notification arrives, the server sub-module gets the corresponding data, which includes the tasks of checking if the received data is consistent and sends it to the optimizer sub-module. This strategy allows to continuously maintain the optimization sub-module iterative process, avoiding blocking communication. Those notifications (e.g., new/edited orders) are passed through shared memory to the optimizer, which in turn computes the best possible routes.

In the other direction, whenever significant improvements to the solution are obtained, they are communicated to the server, which posts them to the *i3FR-Hub*, and which in turn sends them to the company's ERP to be visualized and possibly adopted as a solution. Figure 3(a) shows an excerpt of a particular route, as a JSON document, return by the optimization module to the hub. Among other things, the route document has information about the start and end times, total distance, orders and path identifiers. The document is then expanded in the *i3FR-Hub* by dereferencing the identifiers, and the result is stored on the *i3FR-DB* (see Sect. 2.3) and sent to the ERP.

At this point, since it is unfeasible for the company to process all routes simultaneously, a method for locking/constraining optimization of certain routes was devised. As vehicles are loaded in inverse delivery order, this locking state prohibits the optimization module from changing the delivery sequence but allows other changes as the insertion of new deliveries along the computed path. This allows the company to lock a route to start picking goods and loading the corresponding vehicle. This locking command, as all communications, is relayed through the *i3FR-Hub*. This led to the need to implement a rollback procedure,

(a) Excerpt of route generate by the *i3FR-Opt*.

(b) Example of an expanded path stored on the database.

Fig. 3. Excerpts of JSON documents.

such that the system could recover to the point where the locked solution was sent to the decision maker at the Interface and Management System (i.e., the ERP).

Optimization Process. Mathematically, we have considered an instance of the VRPTW composed of a set of customer locations, C, a set of depot locations, D, and distances $d(i,j) \in \mathbb{R}$ and times $t(i,j) \in \mathbb{N}$ between each pair of locations $i,j \in C \cup D$. Each vehicle has a capacity q_τ for τ in Υ which is the set of storage types (e.g., dry, refrigerated, frozen). Each customer, i, has a time window $[a_i, b_i]$, a service time s_i and a demand $d_{i,\tau}$, for $i \in C$ and $\tau \in \Upsilon$. A solution is a set of m routes, $T_i = (l_{i,1}, l_{i,2}, l_{i,3}, \ldots, l_{i,m_i})$ with $i \in \{1,2,\ldots,m\}$, such that a route starts and ends at the same depot, $l_{i,1} = l_{i,m_i} \in D$, $l_{i,2}$, $l_{i,3}$, $\ldots$, $l_{i,m_i-1} \in C$ are customers, and a customer is served by a single route (i.e., $\{l_{i,2}, l_{i,3}, \ldots, l_{i,m_i-1}\} \cap$

$\{l_{j,2}, l_{j,3}, \ldots, l_{j,m_j-1}\} = \emptyset$ for $i \neq j$). The feasible routes take into consideration the travel time between customers and the associated service time, such that each vehicle leaves and arrives at the depots and clients in their time windows (in the case of the clients it can arrive before the time window opening, in which case it is forced to wait until the opening moment – the maximum waiting time is an additional parameter). Another restriction states that the capacity of the vehicle, in the different transportation types should not be exceeded, i.e., $\sum_{l \in T_i} d_{l\tau} \leq q_\tau$, $i \in \{1, 2, \ldots, m\}$ and $\tau \in \Upsilon$.

Taking the formulation into consideration, two criteria where minimized: the total number of vehicles required to achieve the service within the restrictions and the total traveled distances (i.e., the sum of the distances made by each vehicle).

The optimization procedure is a hybrid adapted Push Forward Insertion Heuristic (PFIH) with solution seeding and post-optimization. The process, implemented in the *i3FR-Opt*, can be divided into the following stages.

Stage 1. Solution Seeding. The optimization procedure starts by computing a seed partial solution, S_{seed}. The partial solution takes into consideration customers' demands and the vehicle capacities to estimate the minimum required number of routes. Once the number of initial routes is computed it is necessary to choose a set of customers and assign them to the new routes. This is done by consecutively choosing the farthest customer from all the routed customers to start a new route. On other words, the first customer to be given a new route is the one farthest away from the depot, c_1. The second route will be started with the customer farther away from the depot and the customer in the first route, c_2. The $n-th$ route will be started with the customer further away from all the previous customers and depot, i.e., $c_n = \arg\max_{i \notin C-C'} \sum_{j \in C'} d(j,i)$, where $C' = \{depot, c_1, c_2, \ldots, c_{n-1}\}$. The motivation to implement this procedure comes from the fact that the selected locations are not, in general, satisfied by the same routes. Possible bad decisions can be later repaired through post-optimization, namely the ejection operator (see Stage 3.). Please refer to [3] for a more detailed explanation of the procedure.

Stage 2. Complete the Seeded Routes. The second stage consists in the completion of the seeded routes using a PFIH. The PFIH is a greedy constructive heuristic [24,28] proposed by M. Solomon. The method has been implemented and tested by several authors [21,25,27]. In general, the PFIH tour-building procedure sequentially inserts customers into the solution. The procedure can be described by the following steps: (A) Using the seed procedure described in Stage 1 instantiate a set of routes, $S = S_{seed}$, which will contain the final solution; (B) If all customers were placed in a route, then stop the procedure and return the built solution; Otherwise (C) for all non inserted customers compute their PFIH cost and choose the one with smallest value; (D) Try to insert the customer into an existing route, minimizing the traveled distance and taking into consideration the constraints (customer's time windows and vehicle's capacities); (E) If the insertion of customers is impossible without violating the constraints, start a

new route ($depot - customer - depot$) and add it to S; (F) Update the distances, delivery times and vehicle capacities. Return to step (B).

As seen in step (C), the computation of the PFIH cost sets the order in which the customers are inserted in the solution. In our case, the i-customer's cost, $PFIHCost_i$, was defined as $PFIHCost_i = -\alpha d(i,o) + \beta b_i + \lambda(b_i - a_i)^{-1}, i \in C$ where o is a depot, and α, β, and λ are parameters such that $\alpha + \beta + \lambda = 1$. Different α, β and λ allow to give more or less importance to formula parcels, resulting in distinct orderings of the customers. For instance, large values of α will prioritize the insertion of customers near the depot. Larger values of β will make customers with earlier closing window preferable. Finally, larger values of λ will prefer customers with smaller time windows to be inserted first. The steps described are repeated for a set of α, β, and λ parameters (namely, for $(\alpha, \beta, \lambda) \in \{(\alpha, \beta, \lambda) : \alpha, \beta, \lambda \in \{0, 0.1, 0.2, \ldots, 1\} \wedge \alpha + \beta + \lambda = 1\}$) creating a collection of solutions, $PFIHSet$, for the next stage.

Stage 3. Post optimization. The next stage is a cycle which is computed until all routes are closed by the ERP/administrator or all PFIH parameters are tested. The cycle starts by getting (and removing) the most promising solution from $PFIHSet$ and setting an ejection rate value. A tabu list, T, is started which will contain all computed solutions before applying post-optimization, for each ejection rate. Then try at most $MaxTries$ times to improve the solution by applying: (a) a 2-Opt operator (which iterates through all routes, one by one, and tries to rearrange the sequence by which the customers are visited in order to reduce the route distance, maintaining feasibility [4]); (b) a cross route operator (similar to the One Point Crossover operator of the Genetic Algorithms [11], receives two paths as input, and tries to find a point where the routes can be crossed, thus improving the total distance, without losing feasibility); and (c) a band ejection operator which is a generalization of the radial ejection [23] (selects a route and, based on the proximity and similarity of the nodes, for each customer located in the route ejects it and a certain number of geographical neighbors which are then reinserted in other routes, without violating the problem's constraints). Please refer to [3] for a more detailed explanation.

The first two operators are capable of diminishing the total distance, i.e., doing route optimization. However, they are not capable of reducing the number of routes present in the original solution, which can be achieved using the third operator.

The first two stages are quite fast. Therefore it is during the last stage that new orders arriving from the *i3FR-Hub* to the *i3FR-Opt* are treated. On other words, the *i3FR-Opt/Server* thread is responsible for the continuous communications with the *i3FR-Hub*, and whenever new orders arrive they are placed in shared memory. After each cycle the *i3FR-Opt/Optimizer* checks the shared memory for new orders that will be treated as ejected customers, i.e., it tries to insert them in the existing routes or creates a new route if that is not possible. As mentioned, during the process, improved solutions are sent from the *i3FR-Opt* to the *i3FR-Hub* which in turn resends them to the ERP/administrator.

2.2 *i3FR-Hub* Module

As already introduced, *i3FR-Hub* is the routing system's central communications hub, which fuses data from several sources to be pre-processed and forwarded, for instance to the *i3FR-Opt* module. *i3FR-Hub* accesses the RESTful-API provided by the Interface and Management System (the ERP) to obtain data on depots, customers' locations, vehicles, customer orders and product details. Vehicle data includes information about the maximum (legal) transportation mass and volume capacities. In the case-study, it was considered that the vehicles have multiple transportation categories, namely: frozen, chilled and ambient/dry goods categories. Product information includes mass per unit, transportation category (e.g., temperature range) as well as individual product dimensions, from which package volumes are computed. Customer orders are pre-processed/aggregated before being sent to *i3FR-Opt*, i.e., the *i3FR-Hub* uses the product information to aggregate products by transportation categories. This way, the *i3FR-Opt* receives the totals for each transportation category (namely, mass and volume), allowing *i3FR-Opt* to focus on improving the solutions. From the list of customer locations and depots, a distance matrix composed of several layers is generated (see Sect. 2.4) and stored on *i3FR-DB*. An excerpt of this data is passed to *i3FR-Opt* where it is used for computing the routes. A different excerpt, containing full human-readable path descriptions is later passed to the Interface and Management System allowing the representation of the computed routes on the management interface and later use on the vehicle's navigation systems (see Fig. 3(b)).

A full log of operations is maintained by *i3FR-Hub*, which allows storing intermediate solutions from the beginning to the end of the optimization process, as well as rolling back and resuming optimization from previous states.

Finally, *i3FR-Hub* also provides a RESTful API for an in-development web-based back office user interface. An overview of the data going in and out of *i3FR-Hub* is shown in Fig. 1(b).

2.3 *i3FR-DB* Module

The *i3FR* routing system stores all data in a network connected high performance database which relies on a MongoDB server, with an ODM (Object-Document Mapper) layer implemented in Python for convenience. MongoDB is a non-relational database, also known as a NoSQL database [12,18]. This is a document-oriented database with high performance and high reliability. Other characteristics include easy scalability (vertically and horizontally through replication and auto-sharding, respectively) and map-reduce.

A MongoDB database is structured as a set of collections, which store documents. These documents are BSON objects (a binary JSON [9] document format), allowed to have a dynamic schema, i.e., documents in the same collection are not forced to have the same structure. This schema-less property is particularly important in the present problem, since some of the data stored in the database does not follow a common design (e.g., data retrieved from the

cartographic system). Other important factors led us to use MongoDB: (1) the possibility to embed in a single document a set of data (e.g., a solution of the problem or an order with multiple products), avoiding intricate and expensive cross-table join procedures; (2) the geospatial engine capable of geographic storage and geographic search features. Many of these features are supported on 2D and 2Dsphere indexes. The first one calculates geometries over a flat surface while the second one does it over an Earth-like sphere. In this case, location data is stored in GeoJSON objects with longitude and latitude. The engine supports fundamental operations like "inclusion" to query for locations contained entirely within a specified polygon, "intersection" to query for locations that intersect with a specified geometry, and "proximity" to query for the points nearest to another point. The last one is used for instance to set the priorities of the access to the cartographic API (see Sect. 2.4). Finally, (3) JSON was defined as the language used to communicate between all modules, which simplifies storing and generating documents, avoiding costly conversions to other data representations.

The decision of using MongoDB combines the high performance of the document storage database with the convenience of using document-like instances. At the same time, overloading the existing Interface and Management System infrastructure with high volumes of database queries is avoided, because a local cache of all relevant ERP data is maintained in the MongoDB database, along with a full operations and error log.

2.4 *i3FR-Maps* Module

The database stores a multi-layered distance matrix built with data obtained from several mapping sources which contains full road information at the highest layer. The first (lowest) layer is built using the geospatial MongoDB features to compute a complete distance matrix of geographic distances over a sphere with approximately the Earth's curvature. This is easily computed from the geographical coordinates, providing a first approximation to the real distance. However, these values are only for newer customers since more exact values will be computed later using cartographic services.

An asynchronous process then makes use of this first layer to incrementally build a second layer from a locally maintained mapping system based on the OpenStreetMaps router [16]. This process will work towards generating a complete matrix of queried routes, prioritizing its queries by nearest geographic distance between locations, which are more probable of belonging to a route.

Initially, this 2nd layer of the distance matrix is provided to *i3FR-Opt* to generate initial delivery routing approximations. These initial approximations are then analyzed in *i3FR-Hub* in order to gradually build a 3rd layer of the distance matrix by selectively querying an up-to-date commercial routing service such as GoogleMaps (the one in use), Bing or similar.

Further iterations of the routing optimization are then based on the 3rd and most precise layer of the distance matrix. The higher layers of the distance matrix include information on distance, travel time and toll information for multiple alternative routes between locations. Full human-readable point by point route

information is also maintained which can be displayed to drivers on the in-vehicle user interface.

2.5 Overall System Dynamics

In summary, the problem in question has a constraint that the solution should be achieved in near real-time since customers tend to send their orders very near to the loading time of the vehicles. In some extreme cases the orders are even made after the beginning of the vehicles' loading, which in those cases will not be unloaded. These last minute orders are then engaged in an existing route, or into a new one if they are not engageable without violating feasibility (e.g., time windows or vehicles capacity). Therefore, from the order formulation by the clients to the delivery of the goods, the process is quite dynamic. These dynamics are supported by human interventions which have a set of actors: clients which send the orders, ERP manager which controls the overall process and remaining workers which, for instance, do the picking of the goods from the warehouse and load the vehicles.

In a sequence line the process can be described as follows (see Fig. 4 for a sequence diagram of the procedure described next). First a client or a seller sends an order to the ERP. Typically the order is for the next day although it can be for any future request. Then, at a convenient moment, the ERP manager sends a signal to the *i3FR* system in order to start the optimization process. This signal goes to the *i3FR-Hub* which begins by requesting the necessary data from the ERP. In general, the requested data is information relevant to the optimization process which is not yet present in the *i3FR-DB*. On other words, static data (e.g., client delivery locations, vehicle data) was already fetched by *i3FR-Hub* and stored in the *i3FR-DB*, which means that the *i3FR-Hub* requests the order information (e.g., volumes and transportation categories). Nevertheless, after receiving the data, the *i3FR-Hub* checks that all necessary data for the optimization process is present. For instance, if a new customer or delivery location is present in the orders then the new information is retrieved from the ERP (delivery location) and the *i3FR-Maps* is updated. The *i3FR-Maps* stores the routes between all possible delivery locations, that is, it has a $n \times n$ matrix (n being the number of delivery locations) of routes including distances, travel time and corresponding routing directions details.

On receipt of the orders, the *i3FR-Hub* posts them to the *i3FR-Opt* module which is divided in two submodules: *i3FR-Opt/Server* and *i3FR-Opt/Optimizer*. During the process, improved solution obtained by the optimizers are sent to the *i3FR-Hub* which stores them in the *i3FR-DB* and routes that information to the ERP.

Two main factors contribute to the dynamism of the process: (a) the arrival of new orders and (b) partial solution locking by the ERP administrator/decision maker. The first factor was a requirement made such that it would be possible to receive late orders while already optimizing. Late orders should be integrated in the already existing solutions without requiring the re-initialization of the

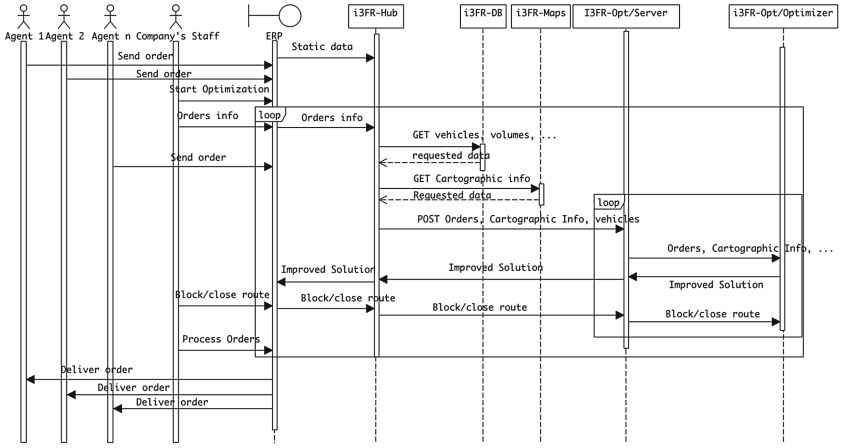


Fig. 4. Sequence diagram of the flow from the customers to the optimizer and reverse.

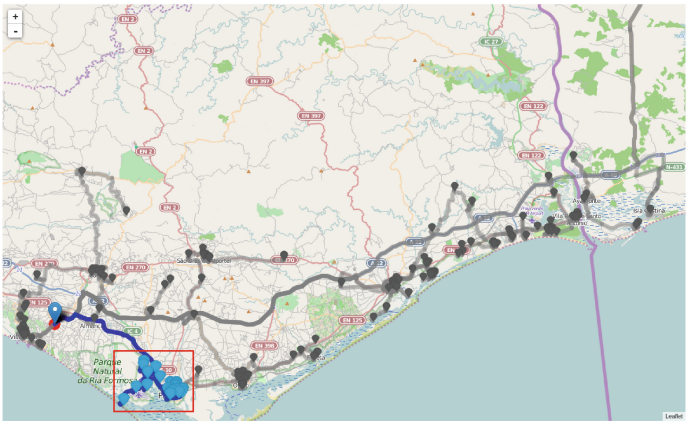
optimization procedure. The second factor has to do with the human intervention of the ERP administrator/decision maker. The ERP does not receive a final solution since the optimization process, being an heuristic, does not guarantee an optimal solution (see Sect. 2). Instead, it receives the best solution computed until the moment by the *i3FR-Opt/Optimizer*. In presence of those solutions, the decision maker can block or close certain routes and prepare the picking and loading of the corresponding vehicles. Since the picking and loading is time consuming, this allows a phased load of the vehicles while continuing the optimization procedure and accepting new orders. In this sense three route states were conceived: "open" meaning that all operations can be made to the route; "blocked" which does not allow altering the order in which the clients are visited, motivated by the fact that the vehicles are loaded in the reverse order of the deliveries, but allows for new orders to somehow be accommodated in the vehicle; and "closed" which does not allow any changes to the route.

2.6 Experimental Results

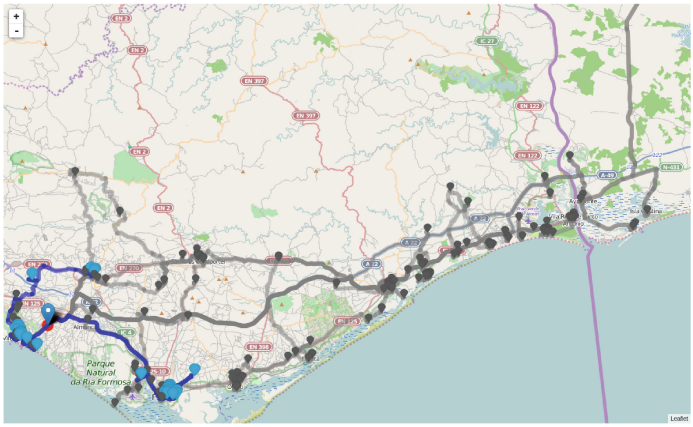
In this section we present a small set of results achieved by the *i3FR* system. For the sake of simplicity, only two scenarios, both with 204 deliveries, for distinct days in the Algarve region where considered (the insertion of new orders after the optimization procedure starts was not considered). In the first scenario all deliveries had time windows between 9:00 and 17:00. In the second scenario, 52 % of the delivery time windows were between 9:00 and 17:00, 28 % between 9:00 and 12:00 and the remaining 20 % between 14:00 and 17:00. For both cases the routes were obtained after 15 s of computation on a commodity computer (Intel i7-4770 processor, with 16 GB of RAM and Kubuntu 14.04).

Figure 5 presents screen shots of the back-office interface were the routes where drawn over the Open Street Maps (OSM) API. The depot is marked with

the symbol, the active route drawn in blue with delivery locations marked with the symbols, and the the remaining routes in gray with delivery location marked with the symbols.



(a) A result for the 1st scenario.



(b) A result for the 2nd scenario.

Fig. 5. Example of the computed routes represented over the OSM API.

i3FR-Opt returned 7 routes for the less restrictive scenario (the first one, Fig. 5(a)). For instance, in this case it is observable that all considered customers from Faro's city, inside the red rectangle, are served by a single route. The second scenario (Fig. 5(b)) led to a rather different solution with 9 routes where, for instance, more than one route was necessary to serve the same customers in Faro. The additional routes in the second scenario are due to non-intersection of the time windows. Serving clients in the marked area, such as presented in first scenario, would imply considerable waiting times, which is not acceptable

in the present business model. In both scenarios the maximum wait time allowed was set to 5 min, i.e., a vehicle can arrive up to 5 min before the opening of the customers time window.

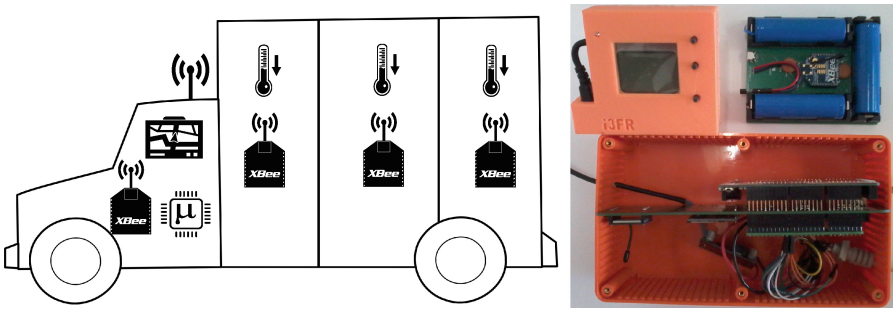
3 Data Acquisition Module

The *i3FR* project also implements a data acquisition module from the vehicles included in a more general system. The acquired data is used to validate the routes and provides information about other parameters, such as the vehicles chambers' temperatures, vehicles' fuel cost, traveled kilometers, and vehicle's Global Positioning System (GPS) data.

The overall system is based on a main controller module, powered by the vehicle's battery, which is based on an ARM microcontroller unit core. This module is responsible for gathering the information from the external sensors and devices, and to communicate it with the main ERP system, through a Global System for Mobile Communications (GSM) protocol. Hence, the main controller of the data acquisition system is connected inside the vehicle to several devices, as follows: (i) to the vehicle's electronic control unit (ECU), using the OBDII standard; (ii) to an external GPS module, using a serial interface; (iii) to temperature sensors placed in the refrigerated chambers, using the ZigBee protocol (based on the IEEE 802.15.4 standard); (iv) to a user's display, to show the basic information such as connected sensors' configuration data, or temperature information; and (v) to an external GPRS (General Packet Radio Service) module, to connect via GSM to the world wide web and send the gathered data to the ERP's database. Figure 6(a) presents a diagram of the vehicle layout showing the sensors, wireless communication and users interface, while Fig. 6(b) presents the developed prototypes for the main controller (the lower module), the user's display (the upper-left module) and the wireless temperature sensor's module (the upper-right module).

An important feature of the data acquisition system is the use of wireless temperature sensors, specially developed for this project, to allow the easy implementation of the system in any existing distribution fleet. As vehicles in this case are, typically, trucks with triple refrigerated cabins or chambers, each one with its unique temperature characteristics and categories (frozen, chilled and ambient/dry) and certified for food transportation, it is necessary to use wireless sensors to not violate the walls of the chambers by installing a wired communication with the sensors.

Hence, the ZigBee protocol was chosen to allow the wireless interface between the microcontroller unit and the sensors. The ZigBee devices use a Coordinator/End-Device configuration, that allows setting the ZigBee on the sensor nodes into a sleep mode. When the ZigBee End-Device (sensor) wakes up, it sends the temperature data to the coordinator. This method allows battery savings while the ZigBee is asleep. Moreover, each external temperature module has a ZigBee device with four ADC (Analog to Digital Converter) inputs, where two are used to acquire the temperature information, from an integrated circuit



(a) (b)

Fig. 6. (a) Diagram of the vehicle layout showing the sensors, wireless communication and users interface; (b) Photo of the developed prototypes for the main controller (the lower module), the user's display (the upper-left module) and the wireless temperature sensor's module (the upper-right module).

(IC) temperature sensor, and a signal for the power-supply voltage obtained from the batteries. The MCP9700 [12] chip is used as temperature sensor, which allows acquiring temperatures in the defined range with a low budget. Note that the signal for the batteries' power-supply voltage is acquired to allow monitoring the battery discharge, remotely in the ERP, during on-field operation. As a long battery life is required and the operating environment can have extreme temperatures from −20°C to +60°C, three 3.2 V LiFePo4 (lithium iron phosphate) type batteries were used, which ensure long lifetimes. Note that these rechargeable batteries, although they have somewhat lower energy density than other common batteries found in consumer electronics (e.g., LiCoO2 type), they offer longer lifetimes, better power density (the rate that energy can be drawn from them) and are inherently safer when working at the considered negative temperature range.

The data acquired by the data acquisition system, is sent to the ERP's database to be used by the *i3FR-Opt* module. It is also stored in the data acquisition module in a SD (Secure Digital) memory card, to ensure that it is not lost during field operations. Moreover, dedicated printed circuit boards for both, the main control module and the temperature sensor module were specially developed and were already tested in real vehicles, sending the acquired information to the ERP's database, through the Internet.

4 Conclusions

The steps to the integration of an ERP with a route planning and optimization module, and a data acquisition module were presented in this work. The route planning and optimization module, as the name suggests, is used to compute the distribution routes being supported on several processes, namely: optimization, maps, database and a hub/communication processes. The optimization

process implements a hybrid Push Forward Insertion Heuristic with pre and post-optimization, prepared to deal some dynamics such as late clients orders and multi-state routes. The maps process is responsible for the retrieve and return of that data relative to the estimated navigation of the vehicles between customers and depots. The database process is responsible for storing all the necessary data to the processes. Finally, the hub process is the core of the communication between all the modules and the ERP. Although integrated with SAGE ERP X3, the route planning and optimization module was designed to be independent of it, allowing the future incorporation with other ERP systems or even autonomous operation. Under that independence perspective, the system was implemented over multi-layered architecture with the communications supported on web services.

The data acquisition module is composed of software and hardware and responsible for the retrieving of data from the fleet's vehicles. Composed of a main controller and a set of wireless communicating sensors, the module sends the acquired data to the ERP allowing the observation of distribution parameters, such as: food transportation chambers' temperatures, vehicles' navigation data, vehicles' consumptions, etc.

Both modules were applied to a company, showing that it was possible to integrate them without causing major disruptions to the business core process.

Acknowledgements. This work was partly supported by project *i3FR*: Intelligent Fresh Food Fleet Router – QREN I&DT, n. 34130, POPH, FEDER, the Portuguese Foundation for Science and Technology (FCT), project LARSyS PEstOE/EEI/LA0009/2013. We also thanks to project leader X4DEV, Business Solutions, http://www.x4dev.pt/.

References

1. Abousaeidi, M., Fauzi, R., Muhamad, R.: Application of geographic information system (gis) in routing for delivery of fresh vegetables. In: 2011 IEEE Colloquium on Humanities, Science and Engineering (CHUSER), pp. 551–555. IEEE (2011)
2. Ambrosino, D., Sciomachen, A.: A food distribution network problem: a case study. IMA J. Manage. Math. **18**(1), 33–53 (2007)
3. Ey, E., Schütz, G., Cardoso, P.J.S., Mazayev, A.: Solutions in under 10 seconds for vehicle routing problems with time windows using commodity computers. In: Gaspar-Cunha, A., Henggeler Antunes, C., Coello, C.C. (eds.) EMO 2015. LNCS, vol. 9019, pp. 418–432. Springer, Heidelberg (2015)
4. Carić, T., Galić, A., Fosin, J., Gold, H., Reinholz, A.: A modelling and optimization framework for real-world vehicle routing problems. In: Caric, T., Gold, H. (eds.) Vehicle Routing Problem, pp. 15–34. InTech (2008)
5. Chen, H.K., Hsueh, C.F., Chang, M.S.: Production scheduling and vehicle routing with time windows for perishable food products. Comput. Oper. Res. **36**(7), 2311–2319 (2009)
6. Faulin, J.: Applying MIXALG procedure in a routing problem to optimize food product delivery. Omega **31**(5), 387–395 (2003)
7. Glover, F., Laguna, M.: Tabu Search. Springer, New York (1999)

8. Hsu, C.I., Hung, S.F., Li, H.C.: Vehicle routing problem with time-windows for perishable food delivery. J. Food Eng. **80**(2), 465–475 (2007)
9. JSON: Javascript object notation, June 2015. http://www.json.org
10. Logvrp.com: Logvrp.com, June 2015. http://logvrp.com
11. Magalhães Mendes, J.: A comparative study of crossover operators for genetic algorithms to solve the job shop scheduling problem. WSEAS Trans. Comput. **12**(4), 164–173 (2013)
12. MongoDB, Inc.: MongoDB, June 2015. http://www.mongodb.com
13. Newronia.com: Newronia.com, June 2015. http://en.newronia.com
14. Optimoroute.com: Optimoroute.com, June 2015. http://optimoroute.com
15. Optrak.com: Optrak.com, June 2015. http://optrak.com
16. OSRM: OSRM – Open Source Routing Machine, June 2015. http://project-osrm.org
17. Osvald, A., Stirn, L.Z.: A vehicle routing algorithm for the distribution of fresh vegetables and similar perishable food. J. Food Eng. **85**(2), 285–295 (2008). http://www.sciencedirect.com/science/article/pii/S0260877407004141
18. Redmond, E., Wilson, J.R.: Seven databases in seven weeks: a guide to modern databases and the NoSQL movement. Pragmatic Bookshelf (2012)
19. Richardson, L., Ruby, S.: RESTful Web Services. O'Reilly Media Inc., Sebastopol (2008)
20. Routyn: Routyn. http://www.routyn.com, June 2015
21. Russell, R.A.: Hybrid heuristics for the vehicle routing problem with time windows. Transp. Sci. **29**(2), 156–166 (1995)
22. SAGE, ERP X3: SAGE ERP X3. http://www.sageerpx3.com/, June 2015
23. Schrimpf, G., Schneider, J., Stamm-Wilbrandt, H., Dueck, G.: Record breaking optimization results using the ruin and recreate principle. J. Comput. Phys. **159**(2), 139–171 (2000)
24. Solomon, M.M.: Algorithms for the vehicle routing and scheduling problems with time window constraints. Oper. Res. **35**(2), 254–265 (1987)
25. Tan, K., Lee, L., Ou, K.: Artificial intelligence heuristics in solving vehicle routing problems with time window constraints. Eng. Appl. Artif. Intell. **14**(6), 825–837 (2001)
26. Tarantilis, C., Kiranoudis, C.: Distribution of fresh meat. J. Food Eng. **51**(1), 85–91 (2002). http://www.sciencedirect.com/science/article/pii/S0260877401000401
27. Thangiah, S.R.: A hybrid genetic algorithms, simulated annealing and tabu search heuristic for vehicle routing problems with time windows. Pract. Handb. Genet. Algorithms **3**, 347–381 (1999)
28. Thangiah, S.R., Osman, I.H., Sun, T.: Hybrid genetic algorithm, simulated annealing and tabu search methods for vehicle routing problems with time windows. Technical report SRU CpSc-TR-94-27 69, Computer Science Department, Slippery Rock University (1994)

Using Conditional Probability and a Nonlinear Kriging Technique to Predict Potato Early Die Caused by *Verticllium Dahliae*

Luke Steere(✉), Noah Rosenzweig, and William Kirk

Michigan State University, East Lansing, MI, USA
{steeregr, rosenzw4, kirkw}@msu.edu

Abstract. *Verticillium dahliae* is a plant pathogenic fungus that can be devastating to commercial potato production. Potato growers in the state of Michigan have experienced yield declines and decreased marketability as a direct result of the persistence of *V. dahliae* in soil. A team of researchers at Michigan State University conducted a soil evaluation using geostatistics and geographic information systems (GIS). The use of a nonlinear Kriging method allowed the team to predict where infection may occur. Nonlinear Kriging is a useful tool for creating conditional probability maps based on a threshold, which can be built into the equation. *Verticillium dahliae* has an inoculum threshold needed to cause infection in a potato plant. Using this threshold, maps can be created based on a probability of any point in space being greater than the threshold. The methods used in this paper show how geostatistics can be a valuable tool for commercial growers.

Keywords: Geostatistics · Indicator · Kriging · Potato early die · Soilborne disease

1 Introduction

Potatoes (*Solanum tuberosum*) are mainly consumed as fresh, chipping, frozen, or starch products and require tubers that meet a high quality standard in either cosmetic appearance or structural integrity from producers. Potato is one of the most intensively managed crops, and cultivation using vegetative tubers as seed makes the crop vulnerable to several recurrent and persistent soilborne diseases. Commercial potato production has long been plagued by a number of persistent soilborne diseases [1]. Nearly 90 % of major diseases that impact crops (including potato) are caused by soilborne pathogens [2]. Soilborne diseases in potato production are currently managed using combinations of chemical fungicides, biological fungicides, fumigation, crop rotation, soil amendments, and other cultural practices. Soil fumigation is not consistently effective against soilborne diseases, but is more consistent against nematodes and some insects, but is cost prohibitive especially at labelled rates and creates environmental concerns [3].

In 2012, a team comprised of potato growers and university researchers was formed to address the issue of declining yields and decreased tuber quality in some areas in

C. Grueau and J.G. Rocha (Eds.): GISTAM 2015, CCIS 582, pp. 142–151, 2016.
DOI: 10.1007/978-3-319-29589-3_9

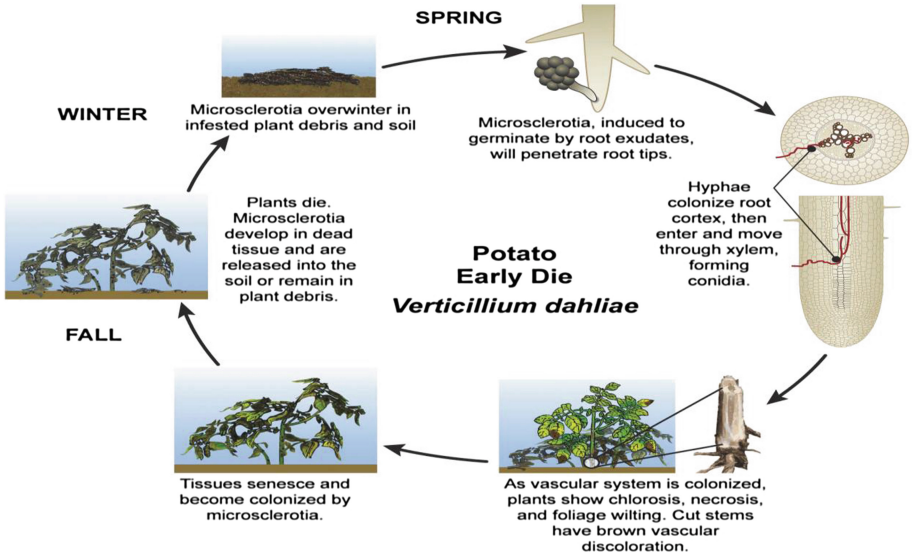


Fig. 1. The disease cycle for potato early die shows how direct penetration of the root cortex leads to vascular blockage and plant death. The dead plant tissue serves as an overwintering structure for new microsclerotia. Image is reproduced with permission, from Michigan State University Extension Bulletin E-3207 © 2015 Michigan State University. All rights reserved.

Michigan dedicated to potato production. The goals of the research were 1. to better understand the spatial variability of soilborne pathogen inoculum levels in potato fields; 2. to better understand the soil biology and quantify soil microbial diversity and 3. to predict where in the field an infection may occur based on pathogen levels determined by conditional probability.

Verticillium dahliae is a soilborne pathogen that is particularly significant and, in conjunction with *Pratylenchus penetrans* (root-lesion nematode), can cause potato early die (PED) [4]. *Verticillium dahliae* has a wide host range including bell pepper, eggplant, mint, potato, and tomato. Potato plants are infected directly via penetration of root hairs by the fungus. Once the fungus has penetrated the root cortex it enters the xylem where it quickly plugs the vascular system leading to premature senescence (Fig. 1). PED is an annual production concern for commercial potato growers and impacts plant health and subsequently, crop yield. The Ascomycota fungus *Verticillium dahliae* is a well-documented pathogen of potato plants [5–7]. The use of conditional probability may better determine where infection by *V. dahliae* might occur based on inoculum levels at sampled locations.

This research used geographic information systems (GIS) and geostatistics to create predictive maps of entire fields from known sample points. GIS technology has proven to be a successful tool in precision agriculture. GIS and global positioning satellite (GPS) technology helps growers to identify problem areas across a field and make management decisions to target those specific locations. This means fields no longer need to be managed as a whole, but that problem areas can be managed as separate entities [8].

GIS has enabled growers to improve productivity, apply fertilizers and pesticides at variable rates, and precisely guide equipment across fields [9]. Since the 1980s when precision agriculture first increased in use, tools such as GIS and GPS have increased grower's understanding of the complex relationships that exist across their fields. The use of linear Kriging methods in soil science has been well documented [10–13]. This project evaluated a nonlinear Kriging model to interpolate the data for *V. dahliae*. Nonlinear Kriging techniques have advantages over linear Kriging techniques due to their ability to account for uncertainty and therefore are often used to predict the conditional probablity for categorical data at non-sampled locations [14, 15].

Indicator Kriging is a nonlinear Kriging technique that is flexible and can be modified to fit specific management or research goals by modifying the critical threshold criteria [16]. Conditional probability maps generated using indicator Kriging can be used to visualize the probability of any point in space (within the field of interest) being greater than a set threshold. When known threshold values are available for certain pathogens and insects, a conditional probability map can be a valuable agronomic crop management tool.

2 Materials and Methods

2.1 Study Areas and Collection of Data

Three field sites located in a commercial potato production area were established for this study in Saint Joseph County in the Southwestern corner of Michigan. Each field was ~30 ha. Each field was on a two-year rotation, alternating between round white potatoes used for chipping and seed corn (*Zea mays*). 20 soil cores were collected from each field, on a grid-sampling scheme to obtain samples proportionally throughout the entire field, with a 25 mm JMC soil corer (Clements Assoc., Newton, IA) to a depth of ~100 mm around a central point in each grid (10 cores and mixed). The position of each point was recorded using a Trimble Juno 3D Handheld GPS device (Trimble Navigation Limited, Sunnyvale, CA). Soil samples were placed in separate labelled plastic bags and stored at 4 °C pending further analysis. Soil data were entered relative to their geographical coordinates and plotted and analysed using ArcGIS 10.1 (ESRI Inc., Redlands, CA).

2.2 Quantification of Verticillium Dahliae Colony Forming Units

To estimate *V. dahliae* colony forming units (CFU), 1.0 g of soil from each sample point was prepared using the wet sieving method [17]. Soil left in the 37 μm sieve was plated onto an NP-10 medium [18] which served as a selective nitrogen source and promoted the development of CFU of *V. dahliae* while inhibiting the growth of other soilborne fungi and bacteria. Isolates were stored at 20 °C for 14–21 days and observed at 4× magnification under a dissecting microscope (Leica Microsystems Inc., Buffalo Grove, IL) and the number of microsclerotia (CFU) were recorded. Each sample point was replicated five times to confirm the accuracy of the initial CFU enumeration.

2.3 GIS Analysis

Point data were collected in the field using a Trimble Juno 3D Handheld GPS device (Trimble Navigation Limited, Sunnyvale, CA) and were uploaded into ArcGIS for further analysis. Data from CFU counts were imported to ArcGIS as data tables then linked to the point data information from the field to create a database from which predictive maps could be developed. Geostatistics were used to predict and interpolate the values of spatially distributed data. Spatial interpolation is based on Tobler's First Law of Geography, that data is often spatially dependent or autocorrelated, or a value of a variable at one location will be similar to values of nearby variables [19, 20].

General Interpolation. In most interpolation methods, predicted values can be estimated by weighted averages from the surrounding areas. The general equation for the interpolation of non-sampled locations is computed as follows:

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{1}$$

where $Z^*(x_0)$ is the non-sampled location that is being predicted, $Z(x_i)$ are the values at n sampled locations and λ_i are the weights assigned to each sampled data point [21]. The difference between interpolation methods is dependent on how λ_i is calculated and what their respective values are.

Indicator Kriging Interpolation Method. The indicator Kriging model assumes an unknown, constant mean. The technique has been well documented [22, 23] and the general form can be computed as follows [14]

$$I(s) = \mu + \varepsilon(s) \tag{2}$$

where μ is an unknown constant and $I(s)$ is a binary variable. The indicator function under a desired cut-off value z_k is computed as

$$I(x, z_k) = \begin{cases} 1, \, if \, z(x) \geq z_k \\ 0, otherwise \end{cases} \tag{3}$$

The indicator Kriging model estimator $I(x_i, z_k)$ at the location can be calculated using

$$I^*(x_o; z_k) = \sum_{i=1}^{n} \lambda_i I(x_i; z_k) \tag{4}$$

and the indicator Kriging, given $\Sigma\lambda = 1$, is

$$\sum_{j=1}^{n} \lambda_j \gamma_I(x_j - x_i) = \gamma_I(x_o - x_i) - \mu \tag{5}$$

Model Evaluation. The accuracy of the indicator Kriging model was evaluated by using the root mean square error (RMSE) cross-validation calculated as [24]

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\hat{Z}(x_i) - Z(x_i)\right]^2} \tag{6}$$

where $\hat{Z}(x_i)$ is the predicted value at the cross-validation point, $Z(x_i)$ is the measured value at point x_i and N is the number of data sets measured. The successfulness of the model in assessing the variability was evaluated by using the root mean squared standardized error (RMSSE) cross-validation statistic calculated as [24]

$$RMSSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\hat{Z}(x_i) - Z(x_i)}{\sigma^2(x_i)}\right]^2} \tag{7}$$

where $\hat{Z}(x_i)$ is the predicted value at the cross-validation point, $Z(x_i)$ is the measured value at point x_i, N is the number of data sets measured, and $\sigma^2(x_i)$ is the variance at cross-validation point x_i.

3 Results and Discussion

Cross-validation statistics analysis was performed on data for the three fields with a low-, high- and variable-risk based on spatial distribution of CFU (Table 1). These cross-validation statistics are used to determine how well the indicator Kriging equation interpolated the *V. dahliae* CFU numbers for each of the three fields. The closer the RMSE is to zero, the closer the prediction is to the measured values [25]. All three fields had RMSE values relatively close to zero meaning that the model derived from the data points in each of the respected fields accurately predicted the probability of any point in space within the field being greater than the threshold of 5 CFU/g of soil.

Table 1. Cross-validation parameter root mean squared error (RMSE) and root mean squared standardized error (RMSSE) are used to assess the accuracy of the models predictions and the model's successfulness in assessing variability.

Field	RMSE[a]	RMSSE[b]
1	0.1133264	0.953032
2	0.3442308	1.145598
3	0.4960541	1.034625

[a]Root mean squared error, the root value of the mean squared error

[b]Root mean squared standardized errors The closer to 1, the more accurate the prediction of variability for that model

The RMSSE shows the model's successfulness in assessing variability. The closer the RMSEE is to 1, the more successful the prediction of variability for that model was [25]. The calculations using the indicator Kriging equations above for each of the three fields of interest showed high levels of accuracy in predicting and assessing variability. Each of the three equations preformed well in regards to how accurate the predictions of the established threshold probability (CFU > 5 CFUs/g of soil) at points that were not sampled.

Conditional probability maps were generated for the three individual fields (Fig. 2). These maps spatially represented the probability of PED incidence based on a 5 CFU/g of soil threshold. A conditional probability map was generated of the low-risk field (Fig. 2A). Based on the 20 original *V. dahliae* CFU values and a threshold value of 5 CFU/g of soil, the indicator Kriging model developed for this field predicts a low incidence of PED. The small portion of the field colored red had a probability from 0.95 to 1 of PED. The majority of the field, colored in blue had a probability between 0 and 0.1 for PED. A conditional probability map was generated of the high-risk field (Fig. 2B). The majority of this field had a probability between 0.95 and 1 for PED. This is quite a contrast from the low-risk field. Finally a conditional probability map was generated of the variable-risk field (Fig. 2C). The result is a map where the probability of being above the established PED threshold varied throughout the field.

The visualized differences among these three maps shows how the use of conditional probability can be used to predict the spatial distribution of plant diseases in the soil and provide and informational tool for commercial potato growers. In an effort to help reduce inoculum levels of *V. dahliae* and other soilborne pathogens, growers will often elect to use soil fumigants. For many years, soil fumigants such as methyl bromide were used, with great effectiveness, to eliminate soilborne plant pathogens such as *V. dahliae* [26–28]. More recently, the commercial agriculture industry has phased out the use of methyl bromide due to its negative effect on the environment [29]. New soil fumigants such as metam sodium and chloropicrin have taken the place of methyl bromide but as researchers begin to better understand the role of beneficial soil microorganism related to plant health [30] the use of any broad-spectrum fumigant is being re-evaluated in a new context. While these soil fumigants may control soilborne pathogens, they may be, in effect, reducing the beneficial soil microorganism populations that assist in plant growth and natural defence against plant pathogenic bacteria and fungi.

The accessibility of conditional probability maps could become a useful informational tool for growers implementing integrated pest management. Rather than making crop management decisions for a field's acreage as a whole, a grower would be able to assess each field individually, or even at the sub-field level to determine problem fields or areas of the field that would be the field that would benefit from soil fumigation. If the grower maintained a low-risk field (Fig. 2A), they could use conditional probability as a holistic management tool to determine no need for fumigation in that field based on the PED risk. Conversely, if the grower assesses the conditional probability for PED and the results indicate a high-risk for PED above the established threshold (Fig. 2B), the grower may elect to treat with applications of soil fumigants. Lastly, if a grower is managing a variable-risk field for PED (Fig. 2C), this would allow the grower to make

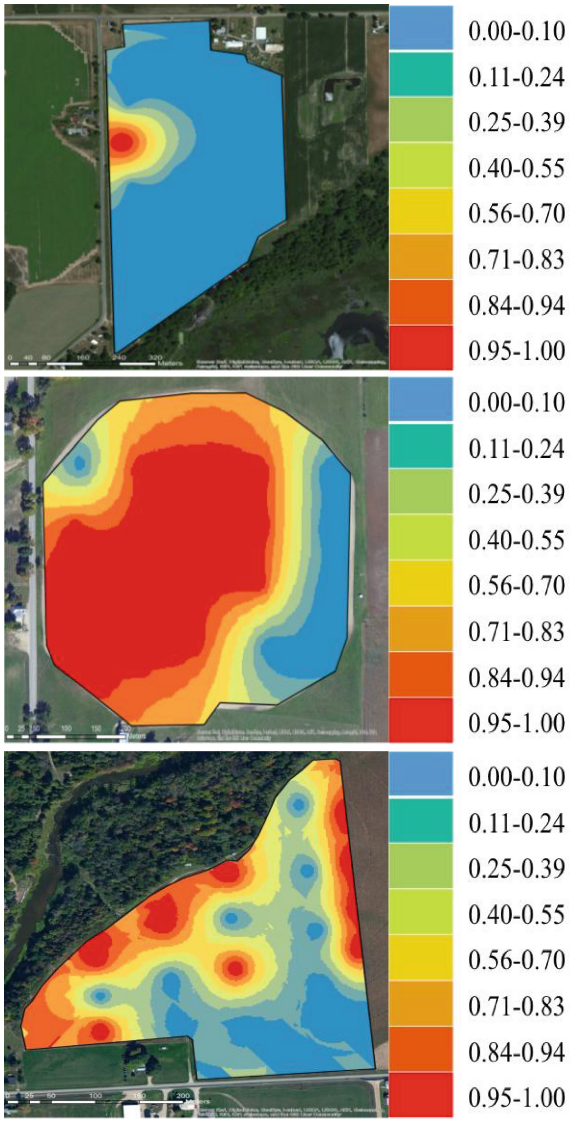


Fig. 2. Conditional probability maps developed for low-risk field (A), high-risk field (B), and variable-risk field (C) using the indicator Kriging method of interpolation with the threshold set at 5 CFUs/g of soil. The conditional probability map for each field represents the risk for the development of potato early die (PED) based on the probability of that area in space having greater than 5 CFUs/g of soil with the color red representing a high probability and the color blue representing a low probability based on predicted values of *Verticillium dahliae* CFUs at that location in the field (Color figure online).

decisions based on a sub-field management approach and only apply fumigant to the portions of the field that present a greater probability of PED. By moving away from generalized, large-scale management practices and into single field and sub-field management strategies with the incorporation of geostatistics and GIS, growers have the potential to greatly decrease input cost and negative environmental effects brought on by heavy regimens of soil fumigants and pesticides, and other inputs.

4 Conclusions

This research shows how geostatistics has the potential to inform management decisions in commercial potato production. Over the last 50 years, commercial potato growers have applied large amounts of soil fumigants to combat soilborne pathogens. New research has indicated that these soil fumigants are disrupting the soil ecology and often times removing beneficial microbes from the soil. When beneficial microorganisms are removed from the soil they are likely to be replaced by pathogenic microorganisms due to the pathogenic microorganism's ability to adapt to environmental stresses. Conditional probability maps will allow growers to make more informed decisions on when soil fumigation is necessary. By decreasing the amount of amendments made to the soil it is possible that beneficial microorganisms will survive and over time a balanced soil environment will be attained. Though this research studied only one pathogen in one cropping system, the methods and tools used have the potential to be incorporated into any cropping system that is plagued by soilborne pests. Most soilborne pathogens that affect commercial crops have been researched for decades. Information on pathogen thresholds needed to cause infection is available for most soilborne pests. Conditional probability mapping can be inserted into most integrated pest management plans for crops that have soilborne pests.

5 Future Plans

The long-term goals of this project include using the baseline data presented here to develop a trans-disciplinary tool combining DNA technologies, GIS and computational biology at the subfield management scale so that growers can easily monitor soil conditions, soil biodiversity, and pathogen levels. This technology will improve productivity, reduce chemical inputs, and improve soil quality for disease-free and sustainable high-quality crop production. The research team is currently developing a high throughput DNA sequencing protocol to more rapidly quantify *V. dahliae* CFU in hopes of getting information to back to growers in a timely manner. By eliminating the 21-day growth period for analysing CFU, a grower would be able to make management decisions earlier in the season. The ultimate goal is to build a framework for the development of subfield management tools that may be used in precision production systems to provide a disease risk advisory for commercial potato growers.

Acknowledgements. This research was supported by funding provided by the Michigan Potato Industry Commission through a USDA NIFA Specialty Crop Block Grant Program (Grant #791N1300). Additional funding and resources were provided by the Michigan Potato Industry Commission and the Michigan State University Project GREEEN (Generating Research and Extension to Meet Economic and Environmental Needs). The authors wish to thank Rob Schafer and the potato growers of Michigan.

References

1. Miller, J., Hopkins, B., Johnson, D.: Checklist for a holistic potato health management plan. In: Potato Health Management, Second Edition, pp. 7–10 (2008)
2. Wilson, C.M.: Roots: Miracles Below. Doubleday, New York (1968)
3. González-Rodríguez, R.M., et al.: Determination of 23 pesticide residues in leafy vegetables using gas chromatography–ion trap mass spectrometry and analyte protectants. J. Chromatogr. A **1196**, 100–109 (2008)
4. Stevenson, W.R., et al.: Compendium of Potato Diseases. American Phytopathological Society, St. Paul (2001)
5. Martin, M., Riedel, R., Rowe, R.: Verticillium dahliae and Pratylenchus penetrans: interactions in the early dying complex of potato in Ohio. Phytopathology **72**(6), 640–644 (1982)
6. Nicot, P., Rouse, D.: Relationship between soil inoculum density of Verticillium dahliae and systemic colonization of potato stems in commercial fields over time (1987). http://apsjournals.apsnet.org/loi/phyto
7. Powelson, M.L., Rowe, R.C.: Biology and management of early dying of potatoes. Annu. Rev. Phytopathol. **31**(1), 111–126 (1993)
8. Heermann, D., et al.: Interdisciplinary irrigated precision farming research. Precision Agric. **3**(1), 47–61 (2002)
9. Adrian, A.M., Dillard, C., Mask, P.: GIS in agriculture. Geograph. Inf. Syst. Bus., 324–342 (2004)
10. Kerry, R., et al.: Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at the regional scale. Geoderma **170**, 347–358 (2012)
11. Kravchenko, A., Bullock, D.G.: A comparative study of interpolation methods for mapping soil properties. Agron. J. **91**(3), 393–400 (1999)
12. Mueller, T., et al.: Map quality for ordinary kriging and inverse distance weighted interpolation. Soil Sci. Soc. Am. J. **68**(6), 2042–2047 (2004)
13. Yost, R., Uehara, G., Fox, R.: Geostatistical analysis of soil chemical properties of large land areas. II. Kriging. Soil Sci. Soc. Am. J. **46**(5), 1033–1037 (1982)
14. Eldeiry, A.A., Garcia, L.A.: Using nonlinear geostatistical models in estimating the impact of salinity on crop yield variability. Soil Sci. Soc. Am. J. **77**(5), 1795–1805 (2013)
15. Goovaerts, P.: Comparative performance of indicator algorithms for modeling conditional probability distribution functions. Math. Geol. **26**(3), 389–411 (1994)
16. Smith, J.L., Halvorson, J.J., Papendick, R.I.: Using multiple-variable indicator kriging for evaluating soil quality. Soil Sci. Soc. Am. J. **57**(3), 743–749 (1993)
17. Nicot, P., Rouse, D.: Precision and bias of three quantitative soil assays for Verticillium dahliae. Phytopathology **77**(6), 875–881 (1987)
18. Kabir, Z., Bhat, R., Subbarao, K.: Comparison of media for recovery of Verticillium dahliae from soil. Plant Dis. **88**(1), 49–55 (2004)

19. Rogerson, P.A.: Statistical Methods for Geography: a Student's Guide. Sage Publications, London (2010)
20. Tobler, W.R., A computer movie simulating urban growth in the Detroit region. Economic geography, 1970: p. 234–240
21. Goovaerts, P.: Geostatistics for Natural Resources Evaluation. Oxford University Press, Oxford (1997)
22. Journel, A.G.: Nonparametric estimation of spatial distributions. J. Int. Assoc. Math. Geol. **15**(3), 445–468 (1983)
23. Solow, A.R.: Mapping by simple indicator kriging. Math. Geol. **18**(3), 335–352 (1986)
24. Ramos, P., Monego, M., Carvalho, S.: Spatial distribution of a sewage outfall plume observed with an AUV. In: Oceans 2008: Proceedings of the MTS-IEEE Conference, Quebec City, QC, Canada (2008)
25. Robinson, T., Metternicht, G.: Testing the performance of spatial interpolation techniques for mapping soil properties. Comput. Electron. Agric. **50**(2), 97–108 (2006)
26. Wilhelm, S., Paulus, A.O.: How soil fumigation benefits the California strawberry industry. Plant Dis. **64**(3), 264–270 (1980)
27. Wilhelm, S., Storkan, R., Sagen, J.: Verticillium wilt of strawberry controlled by fumi-gation of soil with chloropicrin and chloropicrin-methyl bromide mixtures. Phytopathology **51**(11), 744 (1961)
28. Ebben, M.H., Gandy, D.G., Spencer, D.: Toxicity of methyl bromide to soil-borne fungi. Plant. Pathol. **32**(4), 429–433 (1983)
29. Thomas, W.: Methyl bromide: effective pest management tool and environmental threat. J. Nematology **28**(4S), 586 (1996)
30. Hayat, R., et al.: Soil beneficial bacteria and their role in plant growth promotion: a review. Ann. Microbiol. **60**(4), 579–598 (2010)

Using Linked Open Data in Geographical Information Systems

Patricia Carolina Neves Azevedo[1,3], Vitor Afonso Pinto[3], Guilherme Sousa Bastos[2], and Fernando Silva Parreiras[3(✉)]

[1] CPRM, Companhia de Pesquisa de Recursos Minerais, Av. Brasil 1731, Belo Horizonte, MG 30140-002, Brazil
patricia.neves@cprm.gov.br

[2] IESTI, Institute of System Engineering and Information Technology, UNIFEI, Av. BPS 1303, Itajubá, MG 37500-903, Brazil
sousa@unifei.edu.br

[3] LAIS, Laboratory of Advanced Information Systems, FUMEC University, Av. Afonso Pena 3880, Belo Horizonte 30130-009, Brazil
vitor.afonso.pinto@gmail.com, fernando.parreiras@fumec.br

Abstract. Linked Open Data is becoming increasingly important for Geographical Information Systems because most of the sources available on the Web are free of charge. In this work, we present an approach for integrating heterogeneous data located in various public organizations. We address the concepts and technologies which allow for visualizing flood information available from linked open data sources using geographical information systems. The proposed approach adds to the decision-making process, specially in the context of minimizing damage caused by floods. This work also contributes to reducing costs to obtain information beyond organization boundaries by using Semantic Web technologies.

Keywords: Linked open data · Geographical information system · Flood · Semantic web

1 Introduction

The Brazilian federal government, through responsible agencies, adopts actions to minimize the damage caused by floods in river basins, such as collecting and analyzing data. However, despite the amount of information available, these are spread out over several data sources in multiple institutions (e.g., government agencies, private companies and academic institutions), databases, schemas and heterogeneous formats. Some data are available only in PDF or scanned image files in non-compliance to the Brazilian Information Access Law (Law No. 12,527) and are causing rework in agencies and entities that use these files. The diversity of formats and data models hampers the interpretation, integration and reuse. Moreover, there is not possible to display them for a interested user in following up the history of water levels in the rivers of the Rio Doce basin.

C. Grueau and J.G. Rocha (Eds.): GISTAM 2015, CCIS 582, pp. 152–166, 2016.
DOI: 10.1007/978-3-319-29589-3_10

In this context, the following question unfolds: What are the concepts and technologies that for Using Linked Open Data in Geographical Information Systems?

When dealing with floods or competitive intelligence, one realizes that visualization, interaction and dissemination of these data can assist in disaster management. In this context, the principles of linked data [1] are a means to make the information shared on the web available in an standardized way, publishing and linked datasets.

This paper presents a framework able to (1) receive, from different sources, data about specific applications, i.e., floods in Rio Doce Basin, (2) integrate them using semantic web [2] technologies and standards and (3) make them available visually to interested users.

One application is to to identify vulnerable communities and develop emergency and preventive actions, contributing to disaster management on the basin of the Rio Doce. The Brazilian government encourages the publication of data to the public through the Internet, aiming to inform the population and support the transparency of government data. However, the publication of unstructured data is insufficient to achieve the goals of efficiency, transparency and accountability. Semantic Web technology can contribute to achieving these goals by providing data integration of heterogeneous sources.

The paper is organized as follows: Sect. 2 contextualizes the research problem. Section 3 describes the background. Section 4 details the proposed solution for visualizing linked data, presenting the conceptual framework. Section 5 describes the validation with two case studies. Section 6 discusses the related work and Sect. 7 concludes the paper by highlighting its contribution and future lines of action.

2 Scenario

2.1 Case Study 1: Analyzing Flooding Facts and Figures

When analyzing the current situation of data from the Rio Doce basin, we observed that they are not in a format available for reuse. Nowadays, only reports with measurement data from sensors installed along the Rio Doce basin are available on the Internet, using technical language, not appropriate for lay users. In this scenario, in which citizens do not have access to information about the historical and monitoring of water levels from the Rio Doce basin, answering the following questions can expand the vision of managers and interested citizens:

S1Q1: What was the level of the river the monitoring points which recorded flood in the day X?
S1Q2: What is the region with the largest population affected by floods in the day X?
S1Q3: What are the municipalities most affected by rain with HDI below of X?
S1Q4: The works against floods from Brazilian Acceleration Plan (PAC) are developed in the most affected areas?

S1Q5: On which areas occurs more floods and diseases related to floods?
S1Q6: In which areas the occurrence of flooding happened in areas of low altitude?

2.2 Case Study 2: Supporting Competitive Intelligence with Linked Data

Mining companies are responsible for producing ores which are ingredients for various items, such as: cell phones, airplanes, building structures, coins, among others. To achieve an organic growth, mining companies usually have a Mineral Exploration Department who is in charge of a comprehensive geological research program all over the world focused on the discovery of large mineral deposits. As large deposits are difficult to be discovered, many years of study are required until reserves areas can be discovered. The work of Mineral Exploration Department comprises geological, technological and engineering studies to promote discoveries and evaluate the feasibility of the extraction, beneficiation and transportation of the ore, considering technical, economic, environmental and social features. All disciplines must work in an integrated way so as to create long-term value and provide shared benefits to all stakeholders. The competency questions presented next were considered.

S2Q1: What are the countries with the highest human development index (HDI)?
S2Q2: What are the countries with the highest numbers of hospital beds?
S2Q3: What are the countries with the highest export trade volume?
S2Q4: What are the countries with the lowest tax burdens?
S2Q5: What are the countries with the lowest time for starting a business?

3 Background

3.1 Linked Data

According to [3], for the purposes of the Semantic Web, linked data is usually represented in RDF. But the authors recognize that, in general, linked data is just data that is interlinked in some way to make it more useful. Normalized relational databases are a form of linked data, and there is a variety of methods for transmitting such structured data over the Internet, including the open standard Javascript Object Notation (JSON) and eXtensible Markup Language (XML).

Linked Data (LD) is a term describing a set of best practices to facilitate the publishing, accessing and interlinking of the data of the Semantic Web. Thus, Linked Data is an open framework for the loose integration of data in the Internet, where data sources can easily cross-link [4]. LD is the data exposed, shared, and connected via URIs on the Web. It uses URLs to identify things as resources to facilitate people to dereference them. It also provides useful information about

these resources, as well as links to other related resources which may improve information discovery [5].

According to [6], the Linked Data is based on the idea of using the Web to create typed links between data from different sources. It is described in a machine-readable way with an explicitly defined meaning. LD can be linked to and linked from other external data. Linked Data is a promising concept for providing and retrieving widely distributed structured data. It uses the principles and technologies of the Semantic Web to publish, interlink and query data. Using a subset of the Semantic Web stack, Linked Data can provide a common framework for inter-organizational collaboration and data integration. Based on formal ontologies, different information spaces can be integrated in a large data cloud using semantic relations. The cloud can then be accessed using standard methods and web technologies [7].

Linked data technology uses web standards in conjunction with four basic principles for exposing, sharing and connecting data [4,6,8]. These principles are:

1. Use URIs as names for things;
2. Entity URIs should be dereferenceable via HTTP;
3. For a given entity URI, the Linked Data provider should respond with "useful information" about the entity in the form of RDF triples;
4. The provider should include links to additional, related data URIs in the response in order to maximize the interlinking of the web of data;

There are roughly two approaches to fulfill the LD vision, namely community-driven and data-driven. **Community-driven** tries to fulfill the data request of a community, e.g. movie fans, gene researchers, etc. **Data-driven** starts with a set of core data and tries to establish connections with as many relevant data sets as possible to emerge patterns not possible to individual data sets alone [9]. The Linked Data paradigm involves practices to publish, share, and connect data on the Web, and offers a new way of data integration and interoperability. The basic principles of the Linked Data paradigm are: (a) use the RDF data model to publish structured data on the Web, and (b) use RDF links to interlink data from different data sources [10].

3.2 Linked Open Data

A pragmatic vision of the Semantic Web has emerged via the Linking Open Data project (LOD), focusing on translating various datasets available on the Web into RDF and interlinking them, following the Linked Data principles. Lots of different datasets have been provided via this LOD initiative, such as DBpedia (the RDF export of Wikipedia) or Geonames (a large geolocation database). All together, they form a complete Web-scaled graph of interlinked knowledge, commonly known as the Linked Open Data Cloud [11].

For [3], Open data is data that is freely available for anyone to use and re-publish. An open data environment is essential if the Semantic Web is to provide

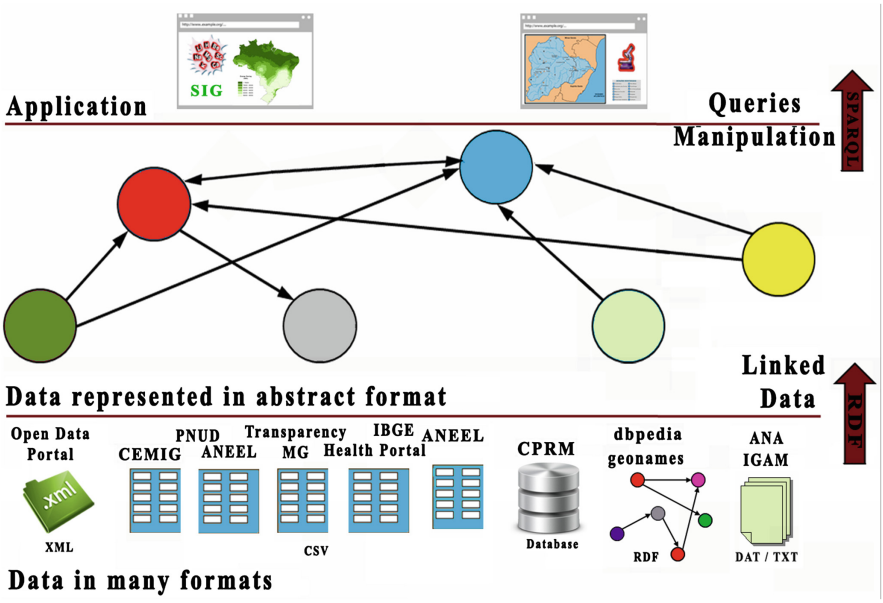


Fig. 1. Overview of the proposed architecture, based on [13].

a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. Open linked data, usually referred to as Linked Open Data (LOD), is just linked data that is open data. Linking Open Data (LOD) initiative is a community project focusing on providing interlinked RDF data from existing open sources, leading to the availability of billion of resources and triples on the Web, and based on the Linked Data principles [12].

4 Conceptual Framework

The Fig. 1 depicts the decomposition of components that are part of the proposed solution, and the relations between them. With this architecture, it is possible, through linked data technologies and principles [1], to receive data from different organizations, to integrate them and to make them available visually.

The Fig. 1 is divided according to the following layers:

(a) Data: The data were obtained from various public agencies in different formats (txt, dat, csv, xml, rdf), and open data from the Linked Data Community available on the Internet. These data were stored in a database and converted to standard RDF [14].

(b) Dataset: The dataset generated from the conversion is already one of the results of this research. It concerns, for example, any information of the levels

of the rivers that comprise the Rio Doce Basin, as well as levels of attention and alert and information from municipalities connected. To answer the research questions, the SPARQL queries [15] were engineered and the result forwarded to GIS.

(c) Visualization in a GIS: The application layer is on top of the architecture, where the information is displayed through the GIS, in a friendly interface and able to answer the questions suggested initially.

The Fig. 1 shows three layers of the proposed solution architecture, where the first layer are the datasets. These data are in different formats and will be converted to standard RDF with the aim of being interconnected and thereby generate the RDF graph, which is illustrated in the second layer of the architecture. In the last layer, we will use SPARQL[1] to query on this data. The result is the combination of all data, and a geographic visualization in a GIS. Geographic information is distinguished from other information by referring to objects or phenomena in a specific location in space and, therefore, has an spatial address [16].

After structuring the data, RDF is used[2] to represent the information, as proposed by the W3C to publish linked data on the web.

5 Validation

In order to validate the proposed approach, a proof of concept through competency questions was conducted, as presented in Sect. 2. Following, the demonstration of queries use in the application and its results.

5.1 Case Study 1: Analyzing Flooding Facts and Figures

Data. As one of our goals is to create a new dataset with data from flooding from the Rio Doce basin, it has become necessary to collect data from different sources, including government databases. In this case, there were collected data from ANA (National Water Agency), ANEEL (National Energy Agency), Cemig (Energy Company), IGAM (State Institute for Water Management) and CPRM (Mineral Resource Research Company) through FTP sites or directly through the organization's Web site. These data were in unstructured formats and have undergone harmonizing, rescaling, and cleaning before its use in the prototype. Given the effort to promote the semantic web [2], we tried to follow open standards as recommended by W3C, representing datasets as linked data [1].

[1] As systems databases make use of SQL to query records in databases, SPARQL is a query language for retrieving information in RDF graphs [15].

[2] Resource Description Framework (RDF) is a language for representing information on the Web and designed for situations where information needs to be processed by applications, rather than simply being shown to people [14].

The dataset creation involved two lines of action: the extraction of collected data through FTP sites or HTML pages of organizations and data conversion from relational databases to RDF model.

Several measurement stations operate in different parts of the rivers and municipalities. Data extracted from these were in TXT format and were converted to CSV using MS Excel software. Other data also related to river levels were collected directly from a CPRM's server, with an employee help. These were in DAT format and were also converted into CSV using the same software. Data were collected from the government website in XML format[3], regarding the Growth Acceleration Program (PAC) in Minas Gerais. Data about HDI of the municipalities was obtained through the PNUD website[4], and were in CSV format. In Brazilian Health Portal website, we collected data about the occurrence of the following diseases related to floods: tetanus, dengue, leptospirosis, malaria, hepatitis A and C, typhoid and cholera. These data were also found in CSV format. Population and altitude data of each municipality was collected directly from the Brazilian Statistics Bureau (IBGE)[5] website in CSV format.

Other sources of data were used aiming to aggregate information, as shown in Table 1.

Table 1. Source, description and format of data used in the study.

Source	Description	Format
ANA	Precipitation and Rivers Levels	DAT
ANEEL	Precipitation and Rivers Levels	CSV
CEMIG	Precipitation	CSV
IGAM	Precipitation	TXT
CPRM	Rivers Levels	Database
Transparency Portal of MG	Onlending of Investments	CSV
IBGE	Population and Altitude	CSV
Health Portal	Diseases	CSV
PNUD	HDI	CSV
Open Data Portal	PAC Works	XML
Geonames	Geographic Names, latitude and longitude	RDF
DBPEDIA	General data of the cities	RDF

With the RDF dataset created about the floods in the Rio Doce Basin and aggregate data, such information becomes part of the Web of Data, where machines and humans can search and use this data set as one of its data sources.

[3] http://dados.gov.br/.
[4] http://www.pnud.org.br/.
[5] http://cidades.ibge.gov.br/.

Fig. 2. RDF graph representing the dataset created.

Is believed that the availability of open and standardized data enables discovery of new knowledge through reuse of this data in new applications. Publishing data about floods in Rio Doce Basin follows the principles of linked data and enables discovery, integration and searches for other sources of data.

The Fig. 2 shows the RDF graph, generated from the RDF file, where the classes Town and River inherit from the upper class Thing. Table 2 relates the concept name depicted in Fig. 2 with the namespace of each concept.

After generating data in RDF model, the file is validated according to Linked Data principles. This verification was taken through the online validation tool W3C RDF Validation Service which was executed successfully. RDF files are available in RDF/XML and N/Triple formats on the following links: RDF/XML: https://db.tt/pJ0r78qw - N/Triple: https://db.tt/DKx7dkK4. Thus, data are ready to be consumed as linked data through browsers, search engines or applications for specific domains.

Table 2. Concepts and namespaces used in the *dataset*.

Conceito	URI
Municpio	http://purl.org/ontology/places#Town
nome_municpio	http://www.geonames.org/ontology#name
populao	http://dbpedia.org/ontology/populationTotal
altitude	http://dbpedia.org/ontology/elevation
lat_long	http://www.georss.org/georss/point
idh	http://dbpedia.org/ontology/humanDevelopmentIndex
doena	http://dbpedia.org/ontology/Disease
investimento_pac	http://paoli.open.ac.uk/watson-cache#Government_aid
Rio	http://dbpedia.org/ontology/River
Estao	http://paoli.open.ac.uk/Open_stream_water_level_recorders
cod_municpio	http://loki.cae.drexel.edu/~wbs/ontology/2004/01/iso-metadata#identCode
cota	http://www.loa-cnr.it/ontologies/OWN/OWN.owl#WATER_LEVEL_2
nivel_alerta	http://kmi-web05.open.ac.uk:81/cache#ALERT_TIME_DUR
nivel_ateno	http://www.loa-cnr.it/ontologies/OWN/OWN.owl#FLOOD_INUNDATION_DELUGE
nivel_enchente	http://ontosem.org/#flood
longitude	http://dbpedia.org/resource/Longitud
latitude	http://dbpedia.org/resource/Latitude
nome_estao	http://xmlns.com/foaf/0.1/
cod_estao	http://www.geonames.org/ontology#featureCode

Table 3. Query S1Q1.

Which stations recorded flood on 01/09/2012?
SELECT ?resource ?cod_station ?station ?measure ?alert_level ?date
WHERE { ?resource geonames:featureCode ?cod_station
?resource paoli:Open_stream_water_level_recorders ?station
?resource dbpprop:date ?date
?resource loa:WATER_LEVEL_2 ?measure
?resource ontosem:flood ?alert_level
FILTER (?date =“2012-09-01”^^xsd:date)}

Dataset. To convert spreadsheet, CSV files, XML files, relational databases and other documents to RDF format we used D2RQ platform [17].

The D2RQ was chosen for use in this study because of some factors, among which stands out the flexibility of mapping language, the simplicity of the commands and the generation of RDF dumps, making possible the reuse of the dataset created.

The next step was to generate RDF dump file from the mapping file and through the dump-rdf D2RQ platform tool. The command provides the following types of output format: Turtle, RDF/XML, RDF/XML-Abbrev, N3 or N-Triple. In this work, the RDF/XML has been used.

Table 3 shows queries and results, limited to 10 lines and no sorting, of the competency question Q1, as a form of validation of the concepts mentioned before.

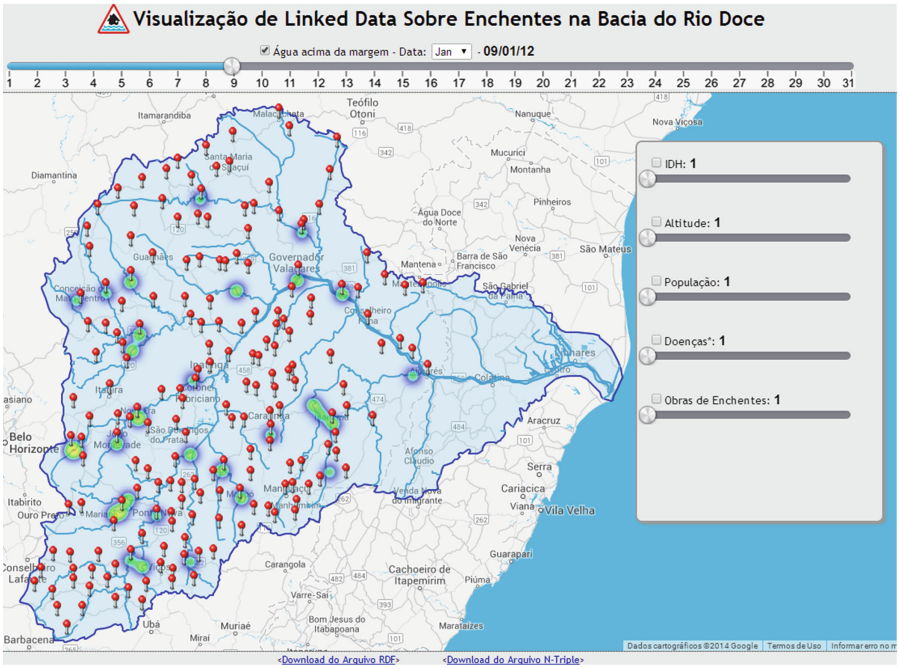


Fig. 3. Visualization of query S1Q1.

Visualization in a GIS. The result of SPARQL queries was displayed into the GIS, a web application implemented using JavaScript language, where the user selects data about the Rio Doce basin, to be viewed on the map. Different combinations can be made with the objective of linking data from multiple sources simultaneously, for example, you can see if the places with high occurrence of floods are the same with occurrences of diseases related to floods or low HDI.

Visualization and interaction of linked data is a question that has been recognized since the beginning of the Semantic Web [18]. When applying techniques of information visualization, semantic beb assists users in exploration and interaction data. The processing and visual presentation of these data are the main goals of information visualization, so that users can get a better understanding of the data [19]. Visualizations are useful for obtaining an overview of the datasets, their main types and the relationships between them.

This application of data visualization provides two main contributions: the visualization of information into a map and the proof that it is possible to make consistents applications from the dataset created in this research. Figure 3 illustrates the resulting prototype, which presents the return of SPARQL queries on a map.

Question #1

SPARQL results:

country	idh
"Norway"	1
"Australia"	0.97
"United States"	0.97
"Germany"	0.94
"Ireland"	0.94
"Netherlands"	0.94
"New Zealand"	0.94
"Sweden"	0.94
"Canada"	0.93
"Japan"	0.93
"Switzerland"	0.93
"Iceland"	0.92
"Korea, South"	0.92
"Hong Kong"	0.92
"Denmark"	0.91

Question #2

SPARQL results:

country	HospitalBeds
"Japan"	1
"Belarus"	0.8
"Korea, South"	0.75
"Ukraine"	0.63
"Germany"	0.59
"Kazakhstan"	0.55
"Austria"	0.55
"Azerbaijan"	0.54
"Czech Republic"	0.51
"Hungary"	0.51
"France"	0.5
"Lithuania"	0.49
"Barbados"	0.49
"Poland"	0.48
"Bulgaria"	0.47

Question #3

SPARQL results:

country	ExportTrade
"China, People's Republic of"	1
"United States"	0.75
"Germany"	0.69
"Japan"	0.38
"France"	0.27
"Netherlands"	0.27
"Korea, South"	0.26
"Russia"	0.25
"Italy"	0.24
"Hong Kong"	0.24
"United Kingdom"	0.23
"Canada"	0.22
"Belgium"	0.21
"Singapore"	0.19
"Mexico"	0.18

Question #4

SPARQL results:

country	MarginalTaxRates
"Bahamas, The"	1
"Bahrain"	1
"Vanuatu"	1
"Montenegro"	0.83
"Bosnia and Herzegovina"	0.81
"Bulgaria"	0.81
"Albania"	0.81
"Macedonia"	0.81
"Paraguay"	0.81
"Qatar"	0.81
"Oman"	0.78
"Cyprus"	0.77
"Ireland"	0.77
"Liechtenstein"	0.77
"Jordan"	0.74

Question #5

SPARQL results:

country	TimeToStartBusiness
"New Zealand"	1
"Georgia"	0.99
"Australia"	0.99
"Macedonia"	0.99
"Portugal"	0.99
"Rwanda"	0.99
"Singapore"	0.99
"Hong Kong"	0.99
"Albania"	0.98
"Iceland"	0.98
"Armenia"	0.98
"Liberia"	0.98
"Netherlands"	0.98
"Belgium"	0.98
"Burundi"	0.97

Fig. 4. SPARQL queries results.

5.2 Case Study 2: Supporting Competitive Intelligence with Linked Data

In order to test if linked open data could be used to support competitive intelligence in mining companies, a prototype for integrated data visualization was built. The prototype was implemented following the framework identified in the literature review and had four layers: data, wrapper, integration and presentation.

Data. In order to identify competency questions, that is, variables and data sources that could be used to build and validate the proposed framework, we applied a focus group methodology, which is understood as a way of collecting qualitative data, and involves engaging a small number of people in informal group discussions, 'focused' around a particular topic or set of issues [20].

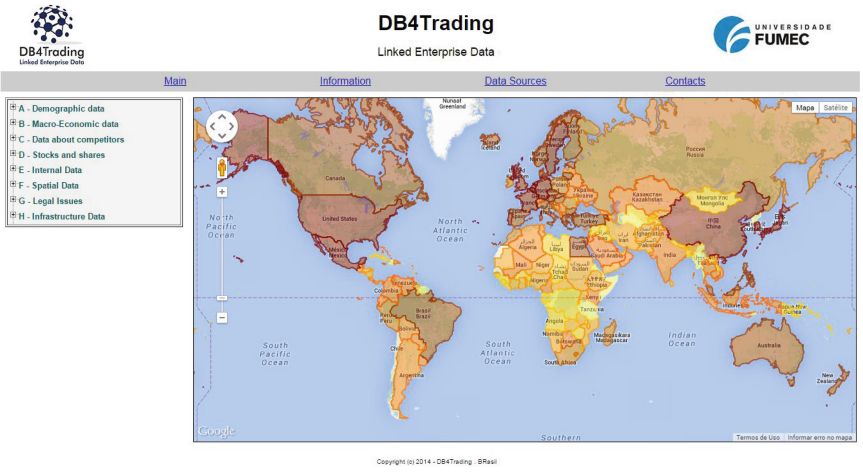


Fig. 5. DB4Trading perspective 1 - results.

Table 4. Generic datasources.

Datasource	Website Address
GOOGLE	https://www.google.com/
DOING BUSINESS PROJECT	http://portugues.doingbusiness.org/
UNITED NATIONS	http://www.un.org/
WORLD HEALTH ORGANIZATION	http://www.who.int/en/
WORLD TRADE ORGANISATION	http://www.wto.org/
CENTRAL INTELLIGENCE AGENCY	https://www.cia.gov/index.html
DELLOITE	http://www.deloitte.com/
LONDON METAL EXCHANGE	http://www.lme.com/
IMF	http://www.imf.org/external/index.htm
EDGARD	https://www.sec.gov/edgar/aboutedgar.htm
SEDAR	http://www.sedar.com/homepage_en.htm
REUTERS	http://br.reuters.com/
ABNT	http://www.abnt.org.br/
CRU	http://www.crugroup.com/
SNL	http://www.snl.com/
BROOKHUNT	http://www.brookhunt.com
SCIENCEDIRECT	http://www.sciencedirect.com/
SCOPUS	http://www.scopus.com/home.url
U.S. CENSUS BUREAU	http://www.census.gov/
U.S. GEOLOGICAL SURVEY	http://www.usgs.gov/
STEEL BUSINESS BRIEFING	https://www.steelbb.com/pt/?PageID=1
NASA	http://www.nasa.gov/
ESRI	http://www.esri.com/
GREENPEACE	http://www.greenpeace.org/brasil/pt/
FRASER INSTITUTE	https://www.fraserinstitute.org/
WIKIPEDIA	http://www.wikipedia.org/

Participants mentioned the existence of specific and generic datasources. 'Specific Datasources' may vary depending on the country, business or competitor being analyzed. 'Generic Datasources' can be used in multiple analysis, independently of country, business or competitor. Participants also separated datasources applicable to multiple segments from those applicable uniquely to their segment. Table 4 presents a list of the identified datasets.

The **data layer** was implemented in MS-SQL Server 2012. This implementation enabled creation of the dataset which was used later to validate competency questions. Dataset was populated with information gathered from various sources, as identified on focus group. From all information identified during focus group, only those available on web in open data format were selected to compound this dataset.

Dataset. In this study, both the **wrapper** and the **integration** layers were implemented through D2RQ Platform, which is a system for accessing relational databases as virtual, read-only RDF graphs. D2RQ Platform has D2R Server, which is a tool for publishing the content of relational databases on the Semantic Web [21]. After installing and configuring D2RQ Platform, the dataset presented in Data Layer section was made available through SPARQL endpoint.

SPARQL queries were performed against dataset to validate the framework. The competency questions presented in Sect. 2 were considered. Figure 4 presents the SPARQL queries results.

Visualization. Finally, the **presentation layer** was implemented. A web application called **DB4Trading** was built so that users could validate data from Semantic Repository using their own criteria. From the weight defined for each one of the variables in the categories area, the application was designed to identify countries whose information were more adherent. In order to identify countries, we used a specific Google API capable to highlight the country polygon. DB4Trading was designed to calculate the color of countries, marking in red those classified as more relevant and leaving in blank those classified as less relevant.

In order to address requirements of competitive intelligence professionals who attended the focus group, we created three visualizations according to the focus group professionals specialties: Engineering perspective, considering infrastructure variables as more relevant; Strategic Management perspective, considering projection variables as more relevant; Health and Environment perspective, considering human variables as more relevant. Figure 5 presents results obtained for the first perspective.

6 Related Work

An early example of using geographic information system was performed by John Snow showing relation between water supply and cholera outbreaks in London in 1854, achieved by linking public data about contaminated water and disease [22].

In the research of Nurefşan Gr, Laura Diaz e Tomi Kauppinen, was used linked open data to publish health-related data, such as diseases, disorders, genes, and drugs into a technology of visualization referred to geo web. For this, the use case studied was RCPH - Research Center of Public Health, based on three conceptual domains: health, spatial and statistical and following the linked data principles. Finally was used an infrastructure integrating geospatial and semantic web technologies to show mortality rates for specific diseases in a spatio-temporal format [23].

Finally, [24] presented a sequence of procedures used to develop an application that used multiple heterogeneous public datasets, about Spain, which are specifically related to administrative units, hydrography and statistical units. The application aims to analyze existing relations between the Spanish coastal

area and different statistical variables such as population, unemployment, housing, industry, commerce and construction. Besides providing methodological guidelines for the generation, publishing and exploitation of Linked Data from these datasets, it was used resources to handle the geometric information of data.

Can be observed that all related work generate an RDF file and visualization in a GIS, however none combined data from a specific topic with statistical data from the location involved, as seen in this study. It is important to note the use of government data in all related work.

7 Conclusion

In this study we analyzed applications of linked data technologies into gathering and integrating data generated both inside and outside an organization for supporting the competitive intelligence process. Firstly, we identified competency questions, that is, variables and data sources that were used to build and validate the framework. We outlined a conceptual framework and we built a prototype for integrated data visualization, based on proposed framework. Finally, we tested and validated conceptual framework, using competency questions.

The contribution of this experiment encompasses the use of methods and tools for publishing data as the principles and standards linked data. It is believed that the availability of open and standardized data enables discovery of new knowledge, of this data through reuse in new applications. For the citizen, the developed application allows a user-friendly visualization of data involved in the research and knowledge discovery from them.

In future, the following lines action can be explored: (a) Expansion of the Dataset: The inclusion of pertinent data improves the relevance, especially when link with existing data; (b) Improvements in data visualization application: Extending the dataset enables new ways of representing data more user friendly way. Therefore, the visualization of information can be enhanced with a larger amount of data, making it more dynamic and interactive for the end user application. (c) Visualization Application (DB4Trading) generated in this research could incorporate some improvements: The tool should allow users to assign scores to each of the data sources because there is a consensus between participants that credibility may vary according to the data source. The tool also could save different scenarios's settings, allowing users to retrieve their preferences.

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. Int. J. Semant. Web Inf. Syst. **5**, 1–22 (2009)
2. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Sci. Am. **284**, 28–37 (2001)
3. Willer, M., Dunsire, G.: Bibliographic Information Organization in the Semantic Web. Elsevier, Amsterdam (2013)
4. Feridun, M., Tanner, A.: Using linked data for systems management. In: NOMS, pp. 926–929. IEEE (2010)

5. Mi, J., Chen, H., Lu, B., Yu, T., Pan, G.: Deriving similarity graphs from open linked data on semantic web. In: IRI, pp. 157–162 (2009). IEEE Systems, Man, and Cybernetics Society
6. Thoma, M., Sperner, K., Braun, T.: Service descriptions and linked data for integrating WSNs into enterprise it. In: SESENA, pp. 43–48. IEEE (2012)
7. Ziegler, J., Graube, M., Urbas, L.: RFID as universal entry point to linked data clouds. In: RFID-TA, pp. 281–286. IEEE (2012)
8. Curry, E., O'Donnell, J., Corry, E., Hasan, S., Keane, M., O'Riain, S.: Linking building data in the cloud: integrating cross-domain building data using linked data. Adv. Eng. Inform. **27**, 206–219 (2013)
9. Svensson, G., Hu, B.: A case study of linked enterprise data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part II. LNCS, vol. 6497, pp. 129–144. Springer, Heidelberg (2010)
10. Galiotou, E., Fragkou, P.: Applying linked data technologies to greek open government data: a case study. Procedia Soc. Behav. Sci. **73**, 479–486 (2013). Proceedings of the 2nd International Conference on Integrated Information (IC-ININFO 2012), Budapest, Hungary, August 30 September 3, 2012
11. Passant, A., Laublet, P., Breslin, J.G., Decker, S.: Semslates: improving enterprise 2.0 information systems using semantic web technologies. In: CollaborateCom, pp. 1–10. IEEE (2009)
12. Breslin, J.G., O'Sullivan, D., Passant, A., Vasiliu, L.: Semantic web computing in industry. Comput. Ind. **61**, 729–741 (2010)
13. Herman, I.: Tutorial on semantic web technologies. Presentation (2012). http://www.w3.org/People/Ivan/CorePresentations/SWTutorial/
14. Manola, F., Miller, E. (eds.): RDF Primer W3C Recommendation. W3C (2004)
15. PrudHommeaux, E., Seaborne, A., et al.: SPARQL query language for RDF. W3C recommendation (W3C)
16. Kraak, J., Ormeling, F.J.: Cartography: Visualization of Geospatial Data. Prentice Hall, New York (2003)
17. Bizer, C., Seaborne, A.: D2RQ-treating non-RDF databases as virtual RDF graphs. In: Proceedings of ISWC2004, vol. 2004 (2004)
18. Geroimenko, V., Chen, C.: Visualizing the Semantic Web: XML-Based Internet and Information Visualization. Springer-Verlag GmbH, London (2003)
19. Card, S.K., MacKinlay, J.D., Schneiderman, B.: Readings in Information Visualization: Using Vision to Think. Interactive Technologies Series. Morgan Kaufmann Publishers Inc., San Francisco (1999)
20. Hyland, S., Haugen, A.S., Thomassen, O.: Perceptions of time spent on safety tasks in surgical operations: a focus group study. Saf. Sci. **70**, 70–79 (2014)
21. Cyganiak, R.: Accessing relational databases as virtual RDF graphs (2012). http://www.d2rq.org. Accessed 01 October 2014
22. Johnson, S.: The Ghost Map: The Story of London's Most Terrifying Epidemic-and How it Changed Science, Cities, and the Modern World. Riverhead Books, New York (2006)
23. Gür, N., Díaz, L., Kauppinen, T.: GI systems for public health with an ontology based approach. In: AGILE 2012, Avignon, France (2012)
24. Vilches-Blázquez, L.M., Villazón-Terrazas, B., Saquicela, V., de León, A., Corcho, O., Gómez-Pérez, A.: Geolinked data and inspire through an application case. In: SIGSPATIAL 2010, GIS 2010, pp. 446–449. ACM, New York (2010)

Author Index