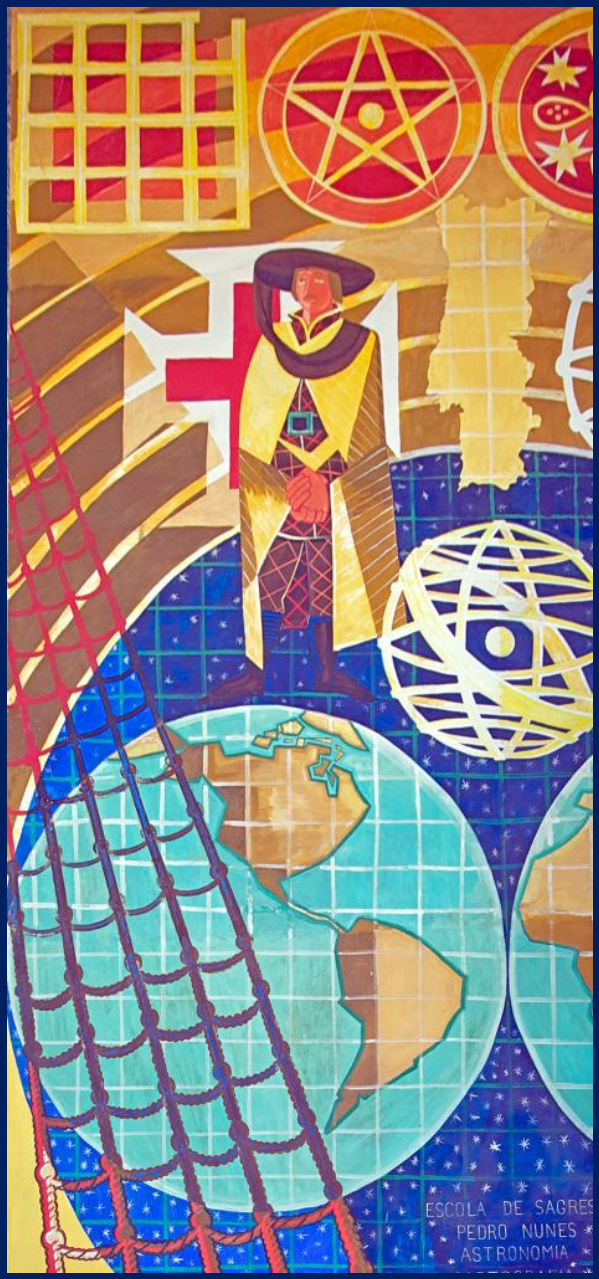
Proceedings of the 7th International Symposium on Spatial Data Quality

Coimbra – Portugal, October 2011



Editors:

Cidália Fonte
Luísa Gonçalves
Gil Gonçalves

Proceedings of the 7th International Symposium on Spatial Data Quality

12 – 14 October, 2011
University of Coimbra, Coimbra, Portugal

Edited by:

Cidália C. Fonte
Luísa M.S. Gonçalves
Gil Gonçalves

Proceedings of the
7th International Symposium on Spatial Data Quality (ISSDQ 2011)
Coimbra - Portugal, October 2011

Cover ilustration:

Panel located at the entrance of the Department of Mathematics - Faculty of Sciences and Technology of the University of Coimbra, where the conference was held.

Editors:

Cidália C. Fonte
Luísa M.S. Gonçalves
Gil Gonçalves

Disclaimer

The Instituto de Engenharia de Sistemas e Computadores de Coimbra and Faculdade de Ciências e Tecnologia da Universidade de Coimbra accept no responsibility for errors or omissions in the papers and shall not be liable for any damage caused by them.

Published and distributed by:

Instituto de Engenharia de Sistemas e Computadores de Coimbra (INESC Coimbra)
Rua Antero de Quental, Nº199
3000 - 033 Coimbra, Portugal

Tel.: + 351 239 851040/9
Fax: + 351 239 824692
http://www.inescc.pt/

ISBN: 978-989-95055-8-2

Printed in Portugal by Redhorse – Coimbra

Preface

The International Symposium on Spatial Data Quality is held every two years and is concerned with all aspects of Spatial Data Quality. The 7th edition of this symposium, held in at the University of Coimbra, from October 12 to 14 2011, was organized by the Institute for Systems and Computers Engineering at Coimbra and the Department of Mathematics of the University of Coimbra. This volume contains 35 papers selected throught a reviewing process including at least two reviewers per paper. The selected papers were scheduled for oral or poster presentation. The contributions reflect the richness of research on topics within the scope of the conference and represent several important developments, specifically focused on methods for assessment of error and uncertainty as well as uncertainty modeling and propagation.

We would like to thank Professor Alexis Comber, Professor Edzer Pebesma and Doctor Luísa Bastos for accepting our invitations to present keynote lectures. We would also like to acknowledge the help and support of several individuals and organizations, which made the organization of this symposium possible, namely the Direction of the Institute for Systems and Computers Engineering at Coimbra and the Direction of the Department of Mathematics of the University of Coimbra, as well as the sponsors indicated below.

SPONSORS (at the date of going to the press):

International Society for Photogrammetry and Remote Sensing
Ordem dos Engenheiros - Região Centro
Leica Geosystems
Turismo do Centro
Turismo de Coimbra

Local Organizing Committee

Cidália C. Fonte (University of Coimbra, INESC Coimbra), chair
Luísa M. S. Gonçalves (Polythecnic Institute of Leiria, INESC Coimbra)
Gil Gonçalves (University of Coimbra, INESC Coimbra)
Jorge Santos (University of Coimbra, INESC Coimbra)
José Paulo Almeida (University of Coimbra, INESC Coimbra)
António Samagaio (Polythecnic Institute of Leiria)

Scientific Committee

Ola Ahlqvist (USA)
José Paulo Almeida (Portugal)
Yvan Bédard (Canada)
Mário Caetano (Portugal)
Ismael Colomina (Spain)
Alexis Comber (UK)
Mahmoud Delavar (Iran)
Rodolphe Devillers (Canada)
Peter Fisher (UK)
Cidália C. Fonte (Portugal)
Giles Foody (UK)
Andrew Frank (Áustria)
Gil Gonçalves (Portugal)
Luísa M.S. Gonçalves (Portugal)
Michael Goodchild (USA)
Gerard Heuvelink (The Netherlands)
André Jalobeanu (Portugal)
Wenzhong Shi (Hong Kong)
Amilcar Soares (Portugal)
Stephen Stehman (USA)
Alfred Stein (The Netherlands)
Nicholas Tate (UK)
Robert Weibel (Switzerland)

Contents

Assessment of Error and Uncertainty

Assessing the positional uncertainties of geometric corrected remotely sensed imagery 3
Carlos A. O. Vieira, Giuliano S. Marotta & Ricardo S. Brites

Evaluation of classifications obtained from lossy compressed remote sensing images 9
Alaitz Zabala & Xavier Pons

Comparison of Mapping Methods for Plumes Using Prior Knowledge from Simulations 15
Kristina B. Helle, Poul Astrup, Wolfgang Raskob & Edzer Pebesma

Multivariate spatial outlier detection using geographically weighted principal component analysis.....21
Paul Harris, Chris Brunsdon & Martin Charlton

Assessing the debris around glaciers using remote sensing and random sets.....27
Mus Bandishoev, Arta Dilo & Alfred Stein

Assessing the spatial variability of the accuracy of multispectral images classification using the uncertainty information provided by soft classifiers.....33
Cidália C. Fonte & Luísa M. S. Gonçalves

New Horizons for Spatial Data Quality Research.....39
Suzie Larrivée, Yvan Bédard, Marc Gervais & Tania Roy

Predicting spatial uncertainties in stereo photogrammetry: achievements and intrinsic limitations 45
André Jalobeanu

Indicators of spatial autocorrelation for identification of calibration targets for remote sensing 51
Nicholas A.S. Hamm, E.J. Milton & V.O. Odongo

Uncertainty Modeling and Propagation

Area Measurement Error Caused by Rasterization 59
Qianxiang Xu & Wenzhong Shi

Multiphase sampling using expected value of information.....65
Sytze de Bruin, Daniela Ballari & Arnold Bregt

Which Spatial Quality can be Meta-Propagated?.....71
Didier G. Leibovici, Amir Pourabdollah & Mike Jackson

Model Parameter Uncertainty Assessment in the Land Transformation Model.....77
Amir Hossein Tayyebi, Saeid Homayouni, Jie Shan, Mohammad Javad Yazdanpanah, Bryan Christopher Pijanowski & Amin Tayyebi

A model to estimate length measurements uncertainty in vector databases 83
Jean-François Girres

A test for stationarity of aggregated spatio-temporal point processes 89
Alfred Stein & Marie-Colette N.M. van Lieshout

Ignoring Correlation Leads to Bone Shaped Confidence Regions and other Counter-Intuitive Aspects of Spatial Data Quality 95
Andrew U. Frank & Gerhard Navratil

Applications

GSM data analysis for tourism application 103
Ana-Maria Olteanu, Roberto Trasarti, Thomas Couronné, Fosca Giannotti, Mirco Nanni, Zbigniew Smoreda & Cezary Ziemlicki

A Quality approach to Volunteer Geographic Information 109
Pau Aragó, Laura Díaz & Joaquín Huerta

Quality Assessment for Cadastral Geometry 115
Gerhard Navratil

Data quality of free of charge climate datasets: A comparison of NOAA temperature and precipitation data with validated sources 121
Péter Zalavári & Hermann Klug

Searching for spatial data resources by fitness for use 127
Ivana Ivánová, Javier Morales, Mesele Atshbeha Gebresilassie & Rolf A. de By

GEOVIQUA: a FP7 scientific project to promote spatial data quality usability: metadata, search and visualization 133
Joan Masó, Ivette Serral & Xavier Pons

Gas pipeline route selection using Dempster Shafer theory of evidence 139
Zahra Bahramian & Mahmoud R. Delavar

Propagation of spatial imprecision in imprecise quantitative data in agronomy 145
Karima Zayrit, Eric Desjardin, Cyril de Runz & Herman Akdag

Spatial Data Model for Local Government with the Inspire Rules 151
Jose Carlos Martinez Llario, Rafael Sierra Requena & Eloina Coll Aliaga

Posters

Analyzing the most adequate GPS kinematic observable for linear elements control methodologies of cartographic products 159
Antonio T. Mozas-Calvache, Manuel A. Ureña-Cámara & Juan J. Ruiz-Lendínez

Handling imperfect spatiotemporal information from the conceptual modeling to database structures 165
Asma Zoghlami, Cyril de Runz, Herman Akdag, Montaceur Zaghdoud & Henda Ben Ghezala

Quality Control of Fieldwork for Estonia's Topographic Mapping 171
Kiira Mõisja, Tõnu Oja

Automated verification of road database in digital images 177
Aluir Porfírio Dal Poz & Marco Aurélio Oliveira da Silva

Construction of an elevation model for the evaluation of estuarine flooding frequency: the case of Lima River, Portugal.....183
Ana Paula Falcão, Alexandre B. Gonçalves, Nuno Silva, Nádia Braz, José Nuno Lima, Maria Amélia V. C. Araújo & António Trigo Teixeira

Partitioning of cadastre features based on straight skeletons for uncertain boundaries.....189
Sunghwan Cho, Jonggun Gim, Gyoungju Lee

Preliminary assessment of the positional accuracy of a QuickBird ortho image 195
Nuno Afonso, Ana Fonseca, José Nuno Lima, Teresa Santos, Sérgio Freire, Ana Navarro & José António Tenedório

Colour Coded Traffic Light Labeling: An Approach to Assist Users in Judging Data Credibility in Map Mashup Applications.....201
Nurul Hawani Idris, Mike J. Jackson & Robert J. Abrahart

Multi-Scale Analysis Approach of Simulating Urban Growth Pattern using a Land Use Change Model.....207
Amir Hossein Tayyebi, Saeid Homayouni, Jie Shan, Mohammad Javad Yazdanpanah, Bryan Christopher Pijanowski & Amin Tayyebi

Uncertainty Framework in Land Use Change Models: An Application of Data, Model Parameter and Model Outcome Uncertainty in Land Transformation Model.....211
Amir Hossein Tayyebi, Saeid Homayouni, Jie Shan, Mohammad Javad Yazdanpanah, Bryan Christopher Pijanowski & Amin Tayyebi

Contributing authors

Abrahart, Robert J. 201
Afonso, Nuno 195
Akdag, Herman 145, 165
Aliaga, Eloina Coll 151
Aragó, Pau 109
Araújo, Maria Amélia V. C. 183
Astrup, Poul 15
Bahramian, Zahra 139
Ballari, Daniela 65
Bandishoev, Mus 27
Bédard, Yvan 39
Braz, Nádia 183
Bregt, Arnold 65
Brites, Ricardo S. 3
Brunsdon, Chris 21
Charlton, Martin 21
Cho, Sunghwan 189
Couronné, Thomas 103
Dal Poz, Aluir Porfírio....177
de Bruin, Sytze 65
de By, Rolf A. 127
de Runz, Cyril....145, 165
Delavar, Mahmoud R. 139
Desjardin, Eric 145
Díaz, Laura 109
Dilo, Arta 27
Falcão, Ana Paula 183
Fonseca, Ana 195
Fonte, Cidália C. 33
Frank, Andrew U. 95
Freire, Sérgio 195
Gebresilassie, Mesele Atshbeha 127

Gervais, Marc 39
Ghezala, Henda Ben 165
Giannotti, Fosca 103
Gim, Jonggun 189
Girres, Jean-François 83
Gonçalves, Alexandre B. 183
Gonçalves, Luísa M. S. 33
Hamm, Nicholas A.S. 51
Harris, Paul 21
Helle, Kristina B. 15
Homayouni, Saeid 77, 207, 211
Huerta, Joaquín 109
Idris, Nurul Hawani 201
Ivánová, Ivana 127
Jackson, Mike J. 71, 201
Jalobeanu, André 45
Klug, Hermann 121
Larrivée, Suzie 39
Lee, Gyoungju 189
Leibovici, Didier G. 71
Lima, José Nuno 183, 195
Llario, Jose Carlos Martinez 151
Marotta, Giuliano S. 3
Masó, Joan. 133
Milton, E.J. 51
Mõisja, Kiira 171
Morales, Javier 127
Mozas-Calvache, Antonio T. 159
Nanni, Mirco 103
Navarro, Ana 195
Navratil, Gerhard 95, 115
Odongo, V.O. 51
Oja, Tõnu 171
Olteanu, Ana-Maria 103
Pebesma, Edzer 15
Pijanowski, Bryan Christopher 77, 207, 211

Pons, Xavier....9, 133
Pourabdollah, Amir 71
Raskob, Wolfgang 15
Requena, Rafael Sierra 151
Roy, Tania 39
Ruiz-Lendínez, Juan J. 159
Santos, Teresa 195
Serral, Ivette....133
Shan, Jie 77, 207, 211
Shi, Wenzhong 59
Silva, Marco Aurélio Oliveira da 177
Silva, Nuno 183
Smoreda, Zbigniew 103
Stein, Alfred....27, 89
Tayyebi, Amin 77, 207, 211
Tayyebi, Amir Hossein 77, 207, 211
Teixeira, António Trigo 183
Tenedório, José António 195
Trasarti, Roberto 103
Ureña-Cámara, Manual A. 159
van Lieshout, Marie-Colette N.M. 89
Vieira, Carlos A. O. 3
Xu, Qianxiang 59
Yazdanpanah, Mohammad Javad 77, 207, 211
Zabala, Alaitz 9
Zaghdoud, Montaceur 165
Zalavári, Péter 121
Zayrit, Karima 145
Ziemlicki, Cezary 103
Zoghlami, Asma 165

ASSESSMENT OF ERROR AND UNCERTAINTY

Assessing the positional uncertainties of geometric corrected remotely sensed imagery

Carlos A. O. Vieira[1], Giuliano S. Marotta[2] & Ricardo S. Brites[2]

[1] Federal University of Santa Catarina, Geocience Department, Florianópolis, Trindade, SC Brazil 88040-900
carlos.vieira@ufsc.br
[2] Institute of Geosciences, University of Brasília, Campus Universitário Darcy Ribeiro, Brasília, DF, Brazil 70910-900
marotta@unb.br; brites@unb.br

Abstract

A new methodology to evaluate and to visualize positional uncertainties that occur through the geometric correction process of remotely sensed imagery is successfully presented. Five different transformation methods are described. Results show that the best RMS error was obtained by 3D the projective modified model and the worst one was obtained by the 2D affine model. The 3D projective model and 3D projective modified one were very similar in performance. These results also point out the importance in choose a better transformation model in order to perform the geometric corrections in remote sensed data and emphasize the importance of select a number of Ground Control Points (GCP) spread all over the study area.

Keywords: Positional Accuracy, Error Propagation, Geometric Correction, Least Mean Square, visualizing uncertainty.

1 Introduction

High spatial resolution of orbital sensors provides greater facility on the collection of control points to perform the geometric correction of remotely sensed imagery. However, it is important to be careful and precise during the process of obtain these reference coordinates, because inherent errors to the reference coordinates, as well as the processes of obtain them, can propagate uncertainty to derived products. Therefore, in order to evaluate the quality of these geometrically corrected images, there is a need to involve techniques that put in evidence the positional uncertainty on explicit and spatial form.

Thus, the objective of this article is to evaluate and to visualize uncertainties that occur through the geometric correction process of remote sensing images and its effect into the final coordinates of geometrically corrected images. Moreover, it is also noticed that knowing the positional coordinates, and its uncertainties, allows the analyst to determine the potentialities and applications of the geometrically corrected image, related to the positional aspect.

2 Material and Methods

2.1 Study area and material

The study area is located in the southeast of Brazil, in the Minas Gerais State, at the Campus of the Federal University of Viçosa (Figure 1). The region has a dramatic landscape, with mountains all over the region.

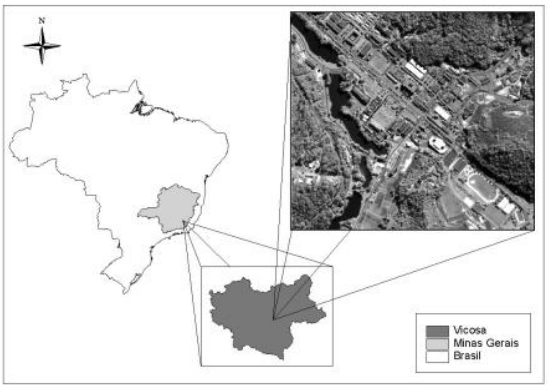


Figure 1. Study area location and a sample of the Quickbird image used.

It was used a QuickBird image in order to carry out the experiments, with spatial and radiometric resolution of 0.60m and 11 bits respectively. It was also used the system ERDAS Imagine 8.3.1 version, for visual control points extractions and visualizations.

Three Ashtech Promark II GPS receivers were used to collect Ground Control Points (GCP) and the Trimble Geomatics Office 1.63 software was also used to process and adjust the coordinates.

It was collected 13 very well defined and distributed ground control points and their homologous on the Quickbird image. The ground control points have their location on the vicinities of VICO GPS reference station, which belongs to the Brazilian Geodesic Network of Continuous Monitoring (RBMC).

2.2 Methodology

The first step was the identification of control features, followed by the extraction of the coordinates in a Quickbird image and the determination of their homologous reference coordinates on the ground, using a GPS Receiver (Vieira *et al.*, 2002).

A topographical polygon was implanted on the field, with many vertices. This polygon was linked to three known GPS points of the Brazilian National Grid, covering 1.5 km^2 inside the study area at the campus of the Federal University of Viçosa.

The polygon computation was carried out using the parametric model *Least Mean Squares* (LMS), method that provided the positional covariance for every polygon vertices (i.e., GCP). This process allowed to get known the positional uncertainties associated to all the used GCP, needed to perform the orbital image geometric correction (Baltsavis *et al.*, 2001).

Using the reference coordinates and their homologous image extracted coordinates, the transformation parameters were calculated using two different transformation models: affine and projective as well their variants for two dimensions (2D) and three dimension (3D) transformations. Such procedure allows the association of a positional uncertainty of the corrected image as a function of the inherent uncertainties from the reference coordi-

nates. In addition, these uncertainties had been propagated, through the transformation parameters, to the rectified image coordinates (Fraser and Amakawa, 2004).

Transformation models are mathematical equations used to transform coordinates between projection systems. The affine and projective transformation models and their variants can be found in Vieira *et al.* (2008).

The determination of these parameters, for all the transformations models, used the parametric *Least Mean Square* (LMS) method, in which the observations (L_b) plus the observation residuals (V) must be equal to the function $F(X_a)$ of the mathematical model used (Gemael, 1994), in relation to adjusted parameters:

$$L_b + V = F(X_a) \tag{1}$$

A stochastic model was set up based on variances of the observed screen coordinates (C_{Lb}). These variances were arbitrary chosen and all of them had the same standard deviations, as following:

$$C_{Lb} = \mathrm{diag}[\ \sigma^2_{C_1}\ \sigma^2_{L_1}\ ..\ \sigma^2_{C_N}\ \sigma^2_{L_N}\ ] \tag{2}$$

This arbitrary procedure of choosing the variances *a priori* is utilized for the weights computations on the observations, which are recomputed due to a variance in the *a posteriori* adjustment, in order to adequate to the adjusted parameters.

After the computation of parameters and weight fitness, the variance-covariance matrix (VCM) of the observed values and of the adjusted parameters is computed, as follows:

$$C_{PAR} = \hat{\sigma}^2_0.(A^t.P.A)^{-1} \tag{3}$$

Using these adjusted parameters and the extracted image coordinates (C and L), it is possible to perform the transformation between systems through the inverse model and obtain the coordinates (X, Y and Z) for every image pixel in the geodesic system.

The precision of the adjustment model was performed using the RMS error analysis and the verification of the *a posteriori sigma 0*, which were evaluated through the X^2 statistical test, using a significance level of 0.05, for each transformation model. Thus, to perform these computations, it is necessary to multiply the *a posteriori* reference variance by the degrees of freedom and compare it with the tabulated value. If the computed value is inside the tabulated interval, the null hypothesis is accepted, implying that the *a priori* reference variance is statistically equal to the *a posteriori* reference variance.

The *a posteriori* adjustment variance was calculated by:

$$\hat{\sigma}^2_0 = \frac{V^t.P.V}{GL} \tag{4}$$

$$P = \sigma^2_0.C^{-1}_{Lb} \tag{5}$$

where P is the observations residuals and GL is the degrees of freedom (the difference between the number of observations and the number of parameters).

After the acceptance of the precision test, the uncertainties of the transformation parameters and that of the reference coordinates were propagated into the geometrically corrected image and, then, RMS errors were generated for every image pixel. As the next step, a *positional error map* could be generated, using the individual RMS, in meters, for each pixel of the image.

3 Results

Table 1A presents the set of 13 coordinates, which were collected directly from the satellite image. It is observed that the collection error for each coordinate was assumed to be 0.5 pixels for every collected point and the altitudes were not collected.

Table 1. Image coordinates (A) and geodetic coordinates - UTM projection system (B).

Points	C (pixel)	C error (pixel)	L (pixel)	L error (pixel)	h (pixel)	h error (pixel)
1	708	0.500	1324	0.500	0	0.000
2	278	0.500	184	0.500	0	0.000
3	1719	0.500	1600	0.500	0	0.000
4	335	0.500	867	0.500	0	0.000
5	1043	0.500	879	0.500	0	0.000
6	1462	0.500	1278	0.500	0	0.000
7	690	0.500	568	0.500	0	0.000
8	1227	0.500	207	0.500	0	0.000
9	1723	0.500	354	0.500	0	0.000
10	1446	0.500	510	0.500	0	0.000
11	369	0.500	1627	0.500	0	0.000
12	1587	0.500	871	0.500	0	0.000
13	1083	0.500	1378	0.500	0	0.000

C = Column, L = Line, h = altitude.

(A)

Points	X (m)	X error (m)	Y (m)	Y error (m)	h (m)	h error (m)
1	721833.861	0.001	7702378.716	0.001	650.998	0.002
2	721585.024	0.001	7703060.306	0.001	647.328	0.003
3	722441.781	0.001	7702224.542	0.001	656.256	0.003
4	721614.844	0.001	7702652.644	0.001	647.328	0.003
5	722039.943	0.002	7702650.328	0.001	654.516	0.004
6	722289.909	0.001	7702415.010	0.001	656.254	0.002
7	721829.400	0.000	7702834.126	0.000	651.286	0.001
8	722154.039	0.001	7703049.621	0.001	709.544	0.002
9	722451.907	0.009	7702966.922	0.004	677.397	0.023
10	722283.740	0.002	7702872.384	0.002	665.209	0.005
11	721630.932	0.079	7702198.404	0.018	661.354	0.073
12	722366.429	0.001	7702656.512	0.001	695.651	0.003
13	722060.586	0.001	7702350.609	0.001	691.864	0.004

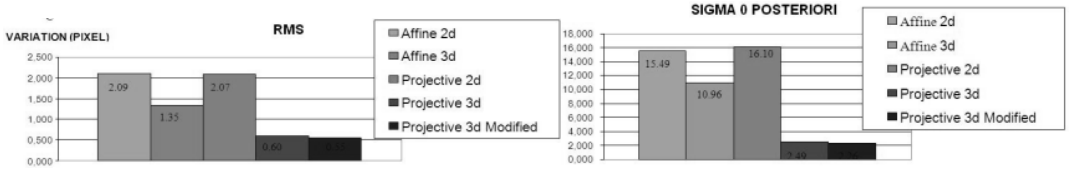
X = East (E), Y = North(N), h = geometric altitude.

(B)

In addition, it was performed a field collection of homologous points, identified in the image, using GPS receivers. The data were processed using the RBMC VICO as the reference GPS station, with the results presented on the Table 1B.

The next step was to perform the observations adjustment using the *Least Mean Square* method for each transformation model and then to evaluate the precision of every model through the RMS error and the verification of the *a posteriori sigma 0* analysis (Marotta and Vieira, 2005).

The global RMS error for each transformation model is presented in Figure 2A. It is shown that the best RMS error was obtained by the 3D projective modified model (0.55 pixels) and the worst one was obtained by the 2D affine model (2.09 pixels), as expected. The 3D projective model presented a very similar RMS error (0.60 pixels). These results point out the importance of choosing a better transformation model in order to perform the geometric corrections in remotely sensed data.



(A) **(B)**

Figure 2. The RMS error (A) and *a posteriori* variances (B) behaviors for each transformation model using LMS adjustment.

Figure 2B shows a graph with the *a posteriori* variances behavior for each transformation model after the LMS adjustment. The results were very consistent with the RMS error analysis. Observing the results, it is possible to conclude that both the 3D projective and the 3D projective modified models obtained the best adjustment results (2.49 and 2.26, respectively), compared to the other ones. Moreover, it is observed that both models presented very similar values too.

Although these measures indicate the best and the worst transformation model to be used, they are non-spatial statistic measures. Therefore, they do not consider the positional aspect of the uncertainty. In order to spacially evaluate the uncertainty, it was pro-

posed a method in which the variances are propagated through the inverse transformation models to obtain the RMS errors, from residual, for every image pixel, bringing adjusted parameters uncertainty into the transformed coordinates.

The Figure 3 presents a *positional error map* for each transformation model. After the uncertainty process spacialization, it is possible to evaluate the areas where the positional accuracy is less precise than the others. The analyst could consider the possibility of control points densification in this area in order to improve the positional accuracy and to obtain a better transformation model adjustment.

Although the non-spatial statistic measures (RMS error and sigma *a posteriori*) pointed out the 3D projective modified model as the best one it is possible, through the propagation of the uncertainty, to conclude that the 3D projective model presented a better result.

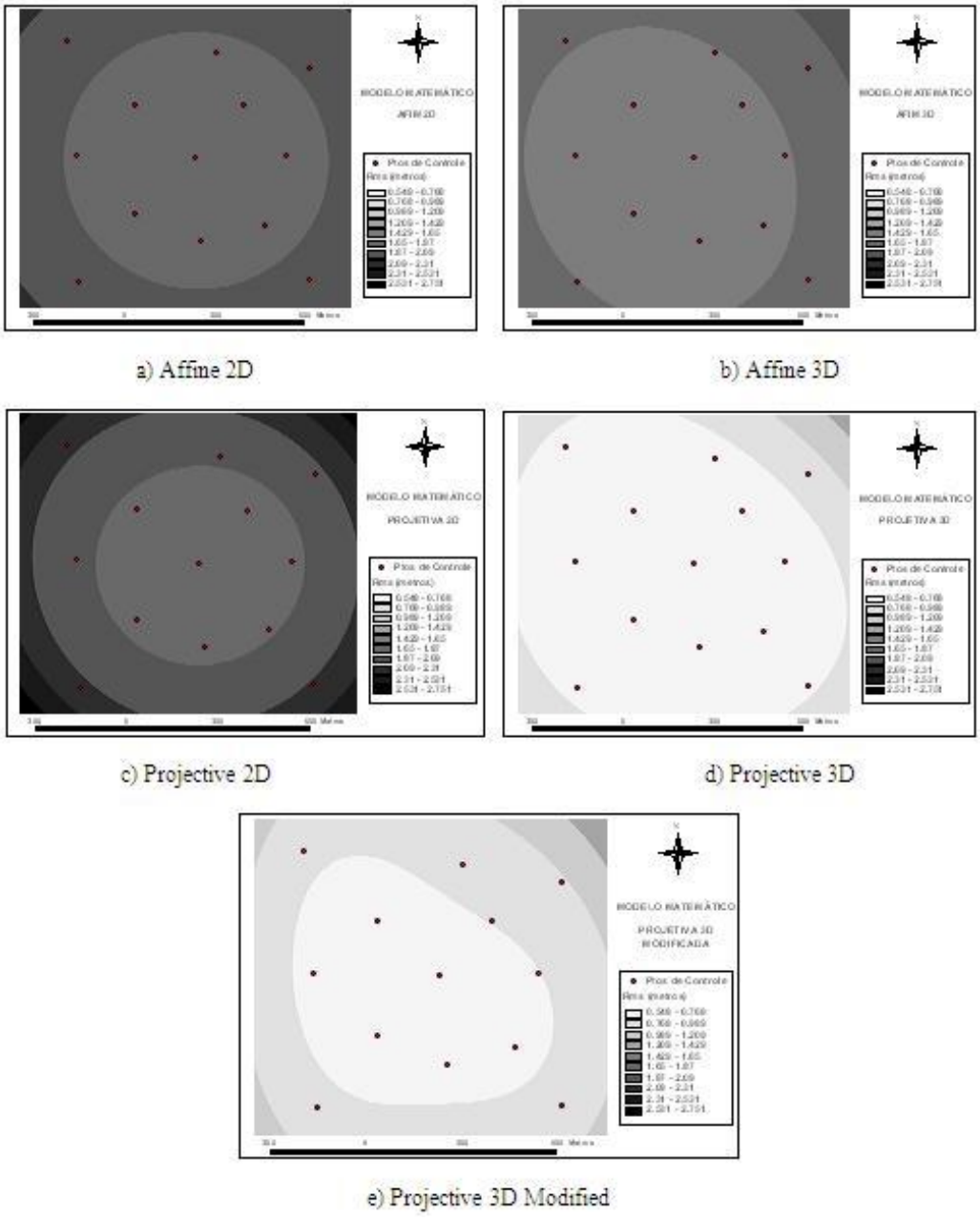


Figure 3. The errors spacialization for each transformation model.

4 Conclusion

After the experiments, it is concluded that the best RMS error was obtained by the 3D projective modified model and the worst one was obtained by the 2D affine model. The 3D projective model and 3D projective modified were very similar in performance. These results point out the importance in choosing a better transformation model in order to perform the geometric corrections in remote sensed data. In addition it was shown that the uncertainties increase as the pixels get far away from the support polygon (i.e., GCP), which emphasizes the importance of selecting a number of GCPs that spread all over the study area. Moreover, using the proposed positional error map, it was possible to evaluate every observation, with its precisions, offering high confidence in the transformed image coordinates. It is also important to mention that the use of the variance propagation rules allowed to analyze the residual uncertainties of the transformation parameters, spatially, in the entire image. Research needs to be developed in order to verify the potential use of these tools for uncertainties visualization of geometrically corrected remotely sensed imagery. Considering the positional error map the 3D projective model presented a better transformation result.

Acknowledgments

The research conducted by Mr. Marotta was supported by the Brazilian Research Councils (Capes, CNPq e FAPESC) performed in collaboration with Civil Engineering Post-Graduation Program (UFV). We are grateful to PSA Research Group for permission to use their images. We would like to thank the Civil Engineering Department and Surveying Engineering Sector (UFV) for the use of their equipments and software.

References

Baltsavias, E., Pateraki, M., Zhang, L. (2001), "Radiometric and Geometric Evaluation of Ikonos Geo Images and Their Use For 3d Building Modelling". *ISPRS Workshop High Resolution Mapping From Space*, Hannover, Germany, pp. 19-21.

Fraser, C. S., Amakawa, Y. (2004), "Insights into the affine model for hight-resolution satellite sensor orientation". *Journal of Photogrammetry & Remote Sensing* – ISPRS, 58, pp. 275-288.

Gemael, C. (1994), *Introdução ao ajustamento de observações: aplicações geodésicas*. Curitiba – PR, Universidade Federal do Paraná - UFPR, 319p.

Marotta, G. S., Vieira, C. A. O. (2005), "Aplicação Do Padrão De Exatidão Cartográfica Em Imagens Orbitais Aster Para Fins De Atualizações De Mapeamentos". In: *XXII Congresso Brasileiro de Cartografia*, Macaé – RJ.

Vieira, C. A. O, Mather, P. M., Borges, P. A., (2002) "Positional accuracy of remotely sensed products". *Proceedings of the 5th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Science*, Melbourne, Australia, July 10-12.

Vieira, C. A. O., Marotta, G., Rodrigues, D. D., Andrade, R. (2008) "Visualizing positional uncertainties of geometric corrected remote sensing images". *Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Science*, Shanghai, P. R. China, June 25 – 27, pp. 327-334.

Evaluation of classifications obtained from lossy compressed remote sensing images

Alaitz Zabala & Xavier Pons

Department of Geography, Universitat Autònoma de Barcelona. 08193 Cerdanyola del Vallès. Spain.
alaitz.zabala@uab.cat, xavier.pons@uab.cat

Abstract

The main aim of this paper is to quantitatively show the effects of several lossy compression techniques on the cartography resulting from remote sensing data, taking into account several image types, geographical scenarios, compression options and evaluations methods. Among the main contributions of the paper, we found that the effect of compression depends on the methodology used to obtain the cartography, on the compression standard employed and on the fragmentation of the study area. More fragmented areas cannot be compressed as much as less fragmented areas, especially if less efficient standards are used. JPEG 2000 obtains better results than the classic JPEG, especially if 3D JPEG 2000 is used. On the other hand, the paper concludes that different quality results can be obtained using several evaluations methods (e.g. independent test areas, ground-truth layer), thus the selection of an evaluation method is of prime importance.

Keywords: Classification, Segmentation, Lossy compression, JPEG, JPEG 2000.

1 Introduction

Remote Sensing (RS) images are used for many applications, including land cover mapping and analysis, disaster management, climate modelling and agricultural and forest management. Data are continually generated, and the amount of information we are acquiring is growing. This process clearly provides enormous application potential, but there is also an important handling problem and a growing need for compression formats that allow the volume of stored data to be decreased without significantly reducing the quality of images used in applications.

The new Spatial Data Infrastructures (SDI) paradigm developed over recent years promotes the establishment of web data services, usually in terms of the Open Geospatial Consortium proposals. These services require compression strategies in order to transfer images repetitively to environments with restricted bandwidth (especially in emergency situations in which mobile devices with low bandwidth are usually the only option). It is unavoidable to standardize data compression and transmission formats in SDI environments in order to achieve interoperability.

In the RS field, and in spite of the spectacular compression ratios reached, there has been little quantitative analysis on the implications of these compressions for classification. Indeed, previous research has generally focused on compression itself, and explored

modifications to compression techniques in order to improve them (Qian *et al.*, 2005; Penna *et al.*, 2007; Du and Fowler, 2007; Carvajal *et al.*, 2008; Choi *et al.*, 2008); however, only some of these studies regard the effect of compression on classification (Qian *et al.*, 2005; Penna *et al.*, 2007; Carvajal *et al.*, 2008; Choi *et al.*, 2008; Zabala and Pons, 2011a; Zabala and Pons, 2011b).

The main aim of this paper is to quantitatively show the effects of several lossy compression techniques on the thematic cartography resulting from RS images, taking into account several image types, geographical areas and compression options. Moreover, several evaluation methods are compared, in order to be aware of the prime importance of evaluation methodologies used. The paper summarizes results from other, more specific, studies so providing a valuable reference for a quantitative overview of the effects of lossy compression in different RS scenarios.

2 Methodology

2.1 Compression

JPEG 2000 and JPEG were considered in order to measure the effect of compression on the classification results. The compression algorithms used were the JPEG 2000 implementations of Kakadu and BOÍ, and the JPEGIMG module implemented in the MiraMon v.6 software based on the JPEG public libraries (hereafter JPG). JPEG2000 can be applied using its special features (*i.e.* multiple-component and the third dimension decorrelation transformation; hereafter J2Km) or in a JPEG-comparable approach (hereafter J2K).

In this paper, the compression ratio (CR) is computed as the ratio between the size of the original file and the size of the compressed file, and is expressed as, *e.g.*, 10:1 for a file that is compressed to a tenth of the original file size.

2.2 Cartography generation

To obtain classifications from remote sensing images, several methods can be used. On this paper, several scenarios are studied to cover pixel by pixel and object-based classification methods.

a) Forest pixel-by-pixel classification

Two forested areas located in Catalonia (NE Spain) were selected: Garrotxa-Ripollès and Maresme-Vallès. The first area is less fragmented than the second one. Both zones were analyzed using four Landsat-7 (ETM+) and 5 (TM) images within 2002-2003, selected taking into account the annual vegetation dynamics to take advantage of the different spectral responses during the year. Additionally to images, other variables were used to improve classification accuracy: NDVI vegetation index, climatic variables and topographic slope.

In order to classify only the forest areas, a mask obtained from the Land Cover Map of Catalonia was applied over the original and the compressed/decompressed images. The classification method used is a hybrid classification that has been designed to improve the accuracy of these classifications (details in Serra *et al.*, 2009) and that combines an unsupervised classifier and a supervised one.

b) Crops pixel-by-pixel classification

We chose two medium-sized zones located in two different agricultural regions in Catalonia: Segrià and Pla d'Urgell. The Segrià zone is richer in fruit trees and slightly

less fragmented. The Pla d'Urgell zone is richer in maize and somewhat more fragmented; it is a zone in which irrigated herbaceous crops and dry permanent crops predominate. Both zones were analyzed using five Landsat-5 (TM) images: 16-05-2004, 17-06-2004, 19-07-2004, 23-10-2004 and 08-11-2004. As in the previous case, images were masked as well as classified with a hybrid classifier.

c) Urban area object-based classification

In this scenario we used four RGB orthophotos and one Quickbird 4-band image (including near infrared, NIR, and RGB bands) to asses the effect of compression on object-based classifications. 3 of the orthophotos are located over Catalonia (Sant Cugat, Vallvidrera and Olot areas) and the other one over Navarre (N Spain: Zizur mayor area). Quickbird image used is the so called Boulder Standard, covering the western side of Boulder, Colorado (USA). All areas have a low density urban landscape. A multiresolution segmentation process was carried out using Definiens professional 5. The segmented image was then classified using a user-defined fuzzy classification.

2.3 Classification evaluation

The accuracy assessment of all classifications was computed by generating con-fusion matrices using ground-truth. The accuracy reports overall accuracy, describing the percentage of well classified pixels and the total amount of classified pixels, as well as the overall kappa index for each classification. User's and producer's accuracy for each class were also obtained.

It is very important for studies that assess compression effects (as well as for most land cover change studies) to be able to evaluate classifications with the highest possible reliability. The best option is to analyze the obtained maps with an "absolute" ground-truth layer (for example coming from a photointerpretation). This approach is not widely used (in fact, if there is a large and absolute ground truth layer, why would we need the result obtained by the classification process?) but it allows us to evaluate the appropriateness of the proposed methodology to be confident in applying it to other areas with similar characteristics. A second option is the use of independent test areas. This is a very common approach (*e.g*, Qian *et al.*, 2005), because it employs independent information (not used on the classification) to test the results, and because it is not very time consuming. Finally, if we are evaluating the effects of compression on classification, we can use the original classification (over non compressed images) as ground truth. (*e.g.*, Choi *et al.*, 2008). This last approach has the advantage of a wide classified area (the full classification) and thus fewer problems appear especially with those categories covering a small area on the classification. On the other hand, the main drawback is that it does not consider that the original classification may have some errors and, thus, it is not able to evaluate the possible positive effects of compression in some circumstances (as found by Qian *al.*, 1997). By comparing the three evaluation approaches we can determine some recommendations on how classifications should be evaluated.

3 Results

This section evaluates the effect of lossy compression on image application, trying to define optimum compression ratios for each application. On the other hand, a comparison among several evaluation methods is presented.

Figure 1 shows the overall accuracy computed using independent test areas of classifications over forest areas, regarding compression ratio. Continuous line show results for the less fragmented areas, and dotted line the more fragmented one. JPEG and J2K results are depicted with a triangle and a square respectively. At low compression ratios, compression does not have a serious effect on the classification results. If the optimum compression level is exceeded, adverse effects appear on the classification accuracy. The inflection point is located at different CR depending on image fragmentation (more fragmented images accept less compression) and on the compression method (JPEG2000 obtains better results). In less fragmented zones, both JPG and J2K high compressions (CR 10:1 or CR 20:1) are possible. However, in fragmented zones it is not advisable to compress images using JPG, and when J2K is used, only medium compression is possible (CR 3.33:1 or CR 5:1).

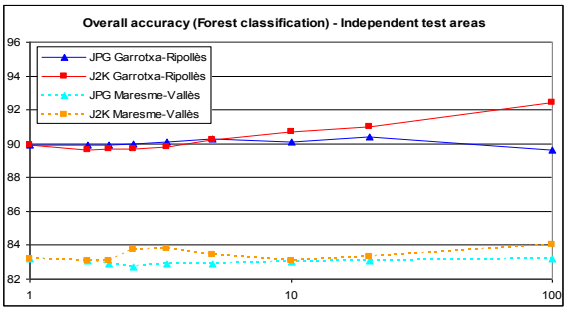


Figure 1: Overall accuracy computed using independent test areas over a pixel-by-pixel classification in forest areas.

Figure 2 shows the overall accuracy of classifications over forest areas, regarding compression ratio. This overall accuracy is computed using independent test areas (left) or using the classification over non compressed images (right) as ground truth. Continuous line show results for the less fragmented areas, and dotted line the more fragmented one. JPEG, J2K and J2Km results are depicted with a triangle, a square and a rhombus respectively. Optimum compression depends on image fragmentation and compression standard. If compression is applied over the less fragmented area, a compression ratio up to 20:1 (JPEG) or 100:1 (J2Km) can be used. If a more fragmented are is classified, the optimum compression is 10:1 for both compression standards.

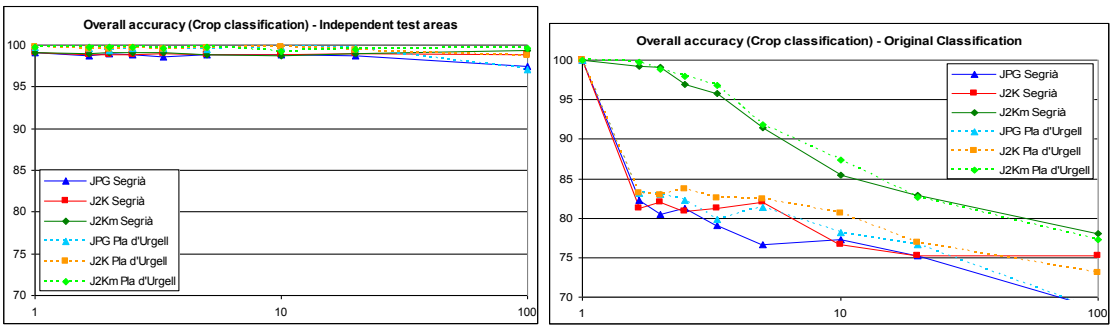


Figure 2: Overall accuracy computed using independent test areas (left) and the original classification (right) over a pixel-by-pixel classification in crop areas.

Regarding the differences among both evaluations methods, it is interesting to note that results of independent test areas are more variable with compression ratio than those based on the original classification. This is probably due to the fact that the area covered by test polygons is only a small proportion of the total area and, thus, results are less stable. Moreover, in some cases results of independent test areas are higher that those

obtained with original images generally due to the lower salt and pepper effect present on images produced with compressed images.

Figure 3 shows the overall accuracy of classifications over urban areas, regarding compression ratio. This overall accuracy is computed using independent test areas (left) or using a photointerpretation (covering a quarter of each image) (right) as ground truth. In this case only J2K results are present, thus different lines and symbols depict several studied areas. Although for all uncompressed images similar overall accuracies were obtained (*e.g.*, 69.40% in the case of Vallvidrera), in more fragmented areas the decrease in accuracy at the first compression level was more significant than in the less fragmented areas. Similar overall accuracies were obtained for classifications for less fragmented areas in the uncompressed and the first three compression levels (up to 20:1), and this decreases more significantly if more compression (40:1 or higher) is applied.

Although pointing to qualitatively similar conclusions, results obtained with an independent test areas are clearly more optimistic quantitatively than those obtained when the analysis uses a ground-truth layer. Assuming that the ground-truth layer is fully representative, this illustrates the importance of selecting the independent areas, and outlines the risk of not to succeed when trying to obtain a representative set of test polygons for all the study area. To avoid this risk, the size of the independent test areas used to test the image should be big enough to be representative. In other words, underestimating or overestimating the accuracy, and therefore obtaining unstable results, is a large risk when the effect of compression is evaluated.

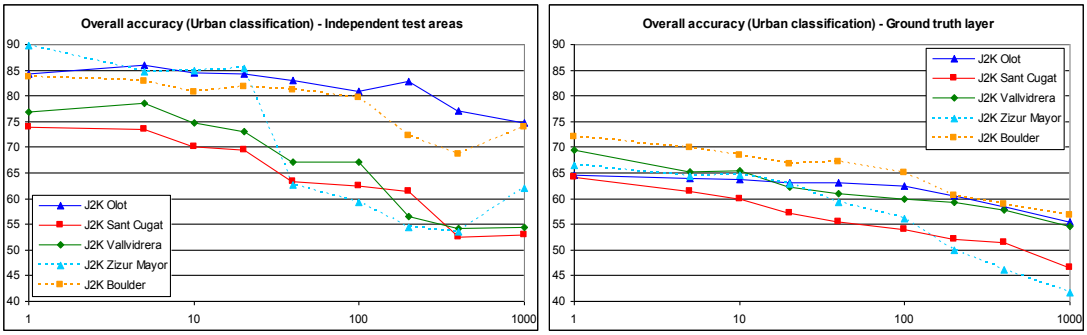


Figure 3: Overall accuracy computed using independent test areas (left) and a ground truth layer (right) over an object-based classification in urban areas.

4 Conclusion

This paper researches how compression affects image classification (several image types, geographical scenarios and compression options), and how several methods to evaluate classification can reflect these effects on the classification.

We found that the effects of compression depend on the methodology used to obtain the cartography. If the objective is pixel-by-pixel classification of a multitemporal RS dataset, the appropriate compression depends on the main land cover (forest or crop areas) and on the area fragmentation. The JPEG 2000 compression ratio varies from 3.33:1 or 5:1 in the worst case (fragmented forest areas) to 100:1 (less fragmented crop areas). JPEG 2000 obtains better results than the classic JPEG, especially if a 3D compression is applied. If object-based classification using image segmentation is the mapping procedure, results show that when images are compressed up to a ratio of 20:1 (depending on the image fragmentation), the classification obtained is similar to the original classification.

On the other hand, the paper concludes that different quality results can be obtained using several evaluations methods. Generally, evaluation results obtained through a ground truth layer are the most solid and are those able to better explain the effects produced due to compression. The difficulty of having those maps usually bring scientist to use other approaches to evaluate compression effects. Among those, the use of independent test areas is the most common approach, but our results show that it can generate important quantitative discrepancies. The use of the original classification as ground truth can not reflect the benefits of compression in some situations. Therefore, in order to rigorously estimate the effects of compression, the selection of the ground truth has to be very accurate.

Acknowledgments

This work was supported in part by the Spanish Ministry of Science and Innovation and the FEDER funds (TIN2009-14426-C02-02), by the Catalan Government (SGR2009-1511), as well as by the European Commission (FP7-ENV-2010-1-265178, FP7-SPACE-2009-1-242390). Xavier Pons is recipient of an ICREA Acadèmia Excellence in Research grant (2011-2015). We would like to thank DigitalGlobe® Inc. for their policy of distributing sample imagery products, which allowed us to include a portion of the Boulder Standard 4-Band, 8-bit, image as a test (© [2006] DigitalGlobe, Inc. All rights reserved).

References

Carvajal, G., Penna, B., Magli, E. (2008), "Unified Lossy and Near-Lossless Hyperspectral Image Compression Based on JPEG 2000". *IEEE Geosc. Rem. Sens. Letters*, 5: 593-597.

Choi, E.; Lee, S., Lee, C., (2008), "Effects of Compression on the Classification of Hyperspectral Images". *In:* N.E. Mastorakis *et al.* (eds.), *Mathematics and Computers in Science and Engineering,* Heraklion, pp. 541-546.

Du, Q.; J.E. Fowler, J.E. (2007), "Hyperspectral Image Compression Using JPEG2000 and Principal Component Analysis". *IEEE Geosc. Rem. Sens. Letters*, Vol.4:201-205.

Penna, B; T. Tillo, E. Magli, G. Olmo (2007), "Transform coding techniques for lossy hyperspectral data compression", *IEEE Transactions Geosc. Rem. Sens.*, 45: 1408-1421.

Qian, S.E.; Hollinger, A.B.; Williams, D.; Manak, D. (1997), "3D data compression system based on vector quantization for reducing the data rate of hyperspectral imagery". *Applications of Photonic Technology*, Vol. 2: 641–654.

Qian, SE; A. Hollinger, M. Bergeron, I. Cunningham, C. Nadeau, G. Jolly, H. Zwick (2005), "A multidisciplinary user acceptability study of hyperspectral data compressed using an on-board near lossless vector quantization algorithm". *Int. J. Remote Sens*., Vol. 26: 2163-2195.

Serra, P., G. Moré, X. Pons (2009), "Thematic accuracy consequences in cadasterland-cover enrichment from a pixel and from a polygon perspective". *Photogramm. Eng. Remote Sens.*, Vol. 75(12): 1441-1449.

Zabala, A.; X. Pons (2011a), "Effects of lossy compression on remote sensing image classification of forest areas", *Int. J. Appl. Earth Obs. Geoinf.*, Vol. 13: 43–51.

Zabala, A.; X. Pons (2011b), "Segmentation and thematic classification of color orthophotos over non-compressed and JPEG 2000 compressed images", *Int. J. Appl. Earth Obs. Geoinf.*, in press, doi:10.1016/j.jag.2011.05.017.

Comparison of Mapping Methods for Plumes Using Prior Knowledge from Simulations

Kristina B. Helle[1], Poul Astrup[2], Wolfgang Raskob[3] & Edzer Pebesma[1]

[1] Institute for Geoinformatics, University of Muenster, Germany
kristina.helle@uni-muenster.de, edzer.pebesma@uni-muenster.de
[2] Risø National Laboratory for Sustainable Energy, Technical University of Denmark
poas@risoe.dtu.dk
[3] Karlsruhe Institute of Technology, Germany
wolfgang.raskob@kit.edu

Abstract

Maps of plumes are required to support decision making in nuclear and radiological emergencies. Such plumes can be modelled with atmospheric transport and dispersion models, but in operational situations the input parameters may not be known sufficiently therefore methods based on measurements may compose a useful complement. We compare contrasting mapping approaches: interpolation, fitting of a parametric trend, and simulations from uncertain parameters. This study is based on simulations. First we derive global properties from a training subset. Using this prior information, maps are produced from some extracted values and compared to the original data. The aim is to minimise absolute residuals and to improve delineation of affected areas. We discovered, that inverse distance weighted interpolation an kriging perform equally well to delineate threshold exceedance, and trend fitting can reduce absolute residuals, especially if few measurements are available.

Keywords: interpolation, trend fitting, plumes

1 Introduction

Maps of plumes are a fundamental tool for decision making in emergencies. The methods to obtain such maps use contrary information: physical models simulate the plume development, based on the wind field and the source term of the released materials, interpolation uses measurements of the resulting plume at several sensor locations. Thus the methods are not interchangeable, but rather complement each other. In this study we compare inverse distance weighted interpolation (IDW), ordinary kriging, fitting of a parametric trend, and simulation. The study area is the surroundings of a nuclear power plant, the potential source of radioactive plumes.

We built our research on a set of simulated plumes, which can be taken as realisations of a random field. These simulations are regarded as "reality". For each of the methods, we first derive global properties, which can improve mapping, from some training plumes. These are optimal neighbourhood size for IDW, the subregional variogram for kriging, and the form and initial parameters of the trend. For simulation we perturb the parameters to account for incomplete knowledge. In the second step we make maps for the other plumes. We extract some "measured" values of these "real" plumes and turn

them into maps, employing the results of the first step. These maps are compared to the original ones to assess absolute residuals and delineation of areas above a given threshold (Figure 1).

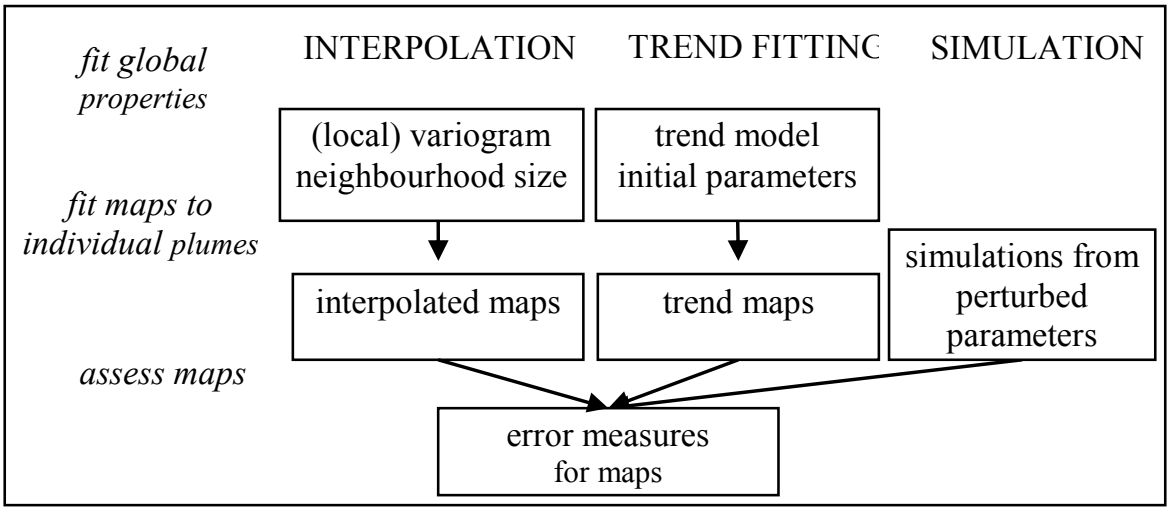


Figure 1. Workflow

2 Methods

This study was based on simulated plumes, made with the RIMPUFF dispersion model (Mikkelsen *et al.,* 1984, 2007). In total 219 one hour releases from the nuclear power plant at Chooz, France, have been simulated. The input meteorological data with one hour resolution was calculated with the Weather Research and Forecast Model (WRF), based on NCEP reanalysis data (Kalnay *et al.,* 1996; NOAA). During January to September 2007 a new one hour release was started every 30 h and simulated for 24 h. The source term was taken from the real-time on-line decision support model RODOS (Ehrhardt and Weis, 2000). It was the same for all simulations, except from scaling by a random factor 10^x with $x \sim N(0,1)$. For further research, we used the cumulative doses of gamma radiation from deposition and puff. These were provided at 7561 locations on a telescopic grid of 200 km x 200 km with a resolution of 0.5 km close to the source and up to 3 km farther away. The first third of the plumes were used for training, to derive global properties. The others were used for testing, to make maps and assess them.

All computations except plume simulations were carried out with the statistical programming language *R* (R Development Core Team, 2010).

2.1 Mapping methods

As an example for a simple interpolation method, we used **inverse distance weighted interpolation** (IDW) with power $p = 2$ applying *idw* from package *gstat* (Pebesma, 2004). To improve it, the optimal number of neighbours was determined first from the training data.

Ordinary kriging is a classical interpolation method that uses the spatial autocorrelation of the data via the variogram. This is assumed to be stationary. To achieve this, we split up the area into 8 x 8 subregions, based on distance and direction from the source. We also assumed the data to vary slower in radial direction, therefore we first derived the anisotropy with *estimateAnisotropy* from *intamap* (Pebesma *et al.,* 2010). To test, if the anisotropy ratio depends on the distance from the source and if the main

anisotropy axis is radial, we used *aov.circular* and *mle.vanmises* (*circular*, Lund and Agostinelli, 2011). Then variograms were fitted for all 64 bins separately using *fitvario* (*RandomFields:* Schlather, 2010). For the interpolation of the test data we used *krige* (*gstat*).

As source and main shape of plumes were known, we developed a **trend model** (1). The trend at location x is determined by the distance to the source $d(x)$ and the angle difference $a(x) = |g(x) - \alpha_5|$ from the main plume direction α_5.

$$f(x) = \max\left(\alpha_1 \cdot \exp(-\alpha_2 \cdot d(x)) \cdot \left| \frac{\alpha_3 - a(x)}{\alpha_3} \right| \cdot \exp\left(-\left| \frac{a(x)}{\alpha_3} \right| \cdot \alpha_4 \right), 0 \right) \quad (1)$$

The parameters $\alpha_1, \ldots, \alpha_4 \in [0, \infty)$, $\alpha_5 \in [0, 2\pi)$ were fitted to each plume individually with the simplex optimisation of Nelder and Mead by *optim* (*stats*: R development core team *et al*. 2011). The direction needed prior adjustment: we started where the values were highest compared to the global average maximum at this distance. For the other parameters optimisation started from the median parameters of the training data. Their optimisation had started with the parameters α_1 and α_2 of a log-linear model of maximum values given the distance to the source, and α_3 was the average width of the non-zero sector more than 5 km from the source.

The methods above do however not use any weather or source term information, which usually is available. This is done by **simulations** from RIMPUFF. To imitate the limited knowledge available in advance, the source term was assumed to be known apart from scaling, and mock weather forecast was derived from the WRF data by "running the time faster" (hours 2, 6, 10, etc. were skipped).

For each plume and mapping method, "measurements" were taken by random, regular, and star sampling designs of different size. For assessment we averaged the mapping errors from all sampling designs of same size.

2.2 Error measures

The samples S from a plume simulation p were turned into a map $\hat{p}_{S,M}$ by each of the methods M. These maps were compared to the original simulation. The first error measure we used, was the sum of the absolute residuals times the related area $A(x)$, integrated over the total area X [in Sv/24h km²] (2).

$$c_1(S, M, p) = \sum_{x \in X} |p(x) - \hat{p}_{S,M}(x)| \cdot A(x) \quad (2)$$

Second, we focused on delineation of the area where dose exceeds a threshold, following Beekhuizen (2008). The threshold was defined as $p_{th} = 5\,\mathrm{mSv/24h}$; there is some international agreement to use it for sheltering zones, however, it is not a legal constraint. We measured the area, where maps would give the wrong classification (3), weighting false negative with a factor 5 for conservative estimation

$$c_2(S, M, p) = \sum_{x \in X} K\left(p(x), \hat{p}_{S,M}(x)\right) \cdot A(x) \quad (3)$$

with an indicator function $K\left(p(x), \hat{p}_{S,M}(x)\right) e \{0,1,5\}$.

3 Results

The values of the plumes are not well suited for interpolation: they vary very fast close to the source, also they are neither stationarity nor isotropic as usually assumed for kriging. The distributions of doses are skewed: on average, 80% of the values are zero, only 6.9% of them exceed the sheltering threshold of 5 mSv/24h, and the highest values of single plumes are between 0.003 Sv/24h and about 8700 Sv/24h, dependent on the scaling. The highest values are usually found close to the source. The maximal values decrease almost exponentially with distance. In many cases, the plume lies within a small sector of on average 18.6°. However, there are also irregular plumes.

3.1 Optimal neighbourhood size for IDW

For all training plumes, maps were produced by IDW using values from regular sampling designs of different size and using only the k nearest neighbours, for $k \in \{1,2,3,4,5,10,20\}$. In most cases, best results are obtained for two or three neighbours (Table 1). Based on this result, we decided to always use $k = 3$ for IDW and kriging.

Table 1. Errors of IDW maps, dependent on neighbourhood size

	k = 1	**k = 2**	**k = 3**	**k = 4**	**k = 5**	**k = 10**	**k = 20**
abs. res.	1887	*1859*	1940	1959	1974	2030	2088
wrong cl.	11969	6456	*2491*	2569	2734	3377	4309

3.2 Variograms for the subregions

The derived parameters support the assumption that values change faster with angle than with distance from the source. The main anisotropy axis –the direction of slowest change – correlates almost perfectly with the direction from the source. The anisotropy ratio is between 0.23 and 0.51 (for isotropic data it is 1), increasing with distance, but not monotonously. This anisotropy was taken into account for further variogram fitting and kriging.

Based on the empirical variograms, the Gaussian model was chosen. We used the average of the parameters fitted by the seven of the ten provided methods, excluding "autostart", "ml", and "reml" by visual inspection. As expected, the total sill decreases with distance because values far from the source are all small and therefore vary little. The proportion of the nugget declines from 60% to 2.6%: close to the source the radioactivity is very concentrated and can change a lot within small distances whereas farther away the plumes are smoother. The range, i.e. the distance up to which values are similar, is about 2.7 km close to the source, 20 km - 30 km elsewhere, and about 70 km at the margins. Nugget and partial sill change significantly with distance from source, but not with direction, whereas the range varies with both. If there was no significant difference we used the average. For seven of the 64 bins variogram fitting failed, for these the range was derived from the ranges for other bins by a linear model on distance and angle.

3.3 Global properties of the trend

Most of the median parameters from fitting the trend to the training data differ little from the input parameters described above. Only the radial decay α_2 and the sector width α_3 increase about 40% - 45%, all other parameters change less than 5%. We also

tested, if parameters could be replaced by global averages. But this would increase absolute residuals: if sector decay α_4 is not fitted individually, 73% of the maps are worse, for radial decay α_2 this happens in 97.3% of the cases. Therefore we kept individual fitting of all parameters.

To test the influence of initial parameters, we used the median parameters from the training data to turn 196 "measurements" from each plume into maps. For the training data, the absolute residuals are on average 89% of the integral of absolute original values, for the test data this fraction is 91%. For fitting to the full set of 7561 values from the training data, it is 84%.

3.4 Comparison of mapping techniques

For four different sample sizes and the three sampling types, samples were taken from all plumes of the testing data and turned into maps by IDW, kriging, and trend fitting. Besides, a simulation with perturbed input parameters was run (Figure 2). For the resulting maps, absolute residuals and wrong classification were computed. Fitting the trend was fastest (below 5 sec for 146 maps); IDW took about 5 times longer and kriging even 450 times.

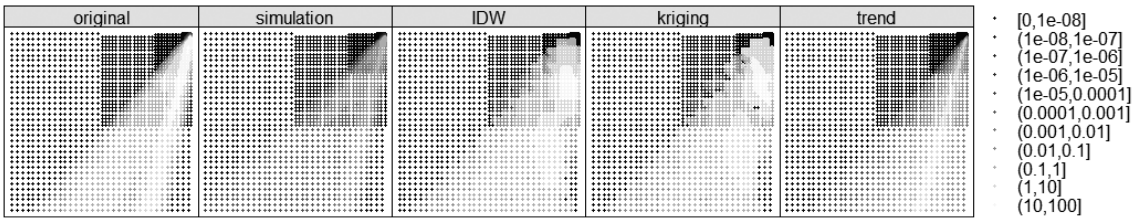


Figure 2. Example of dose [Sv/24h] from a plume, mapped by the different methods using a regular sample of 400 "measurements".

Wrong classification differs highly significant between the methods, whereas the difference for the absolute residuals is only weakly significant ($p = 0.075$). For wrong classification, IDW is best, kriging is only 9% worse, whereas the error increases 120% for simulation and 198% for trend fitting. For absolute residuals, the best method is trend fitting, followed by the 9.3% worse simulation, IDW with 20% impairment, and kriging with 29% (Table 2).

Table 2. Average errors for different methods

	IDW	**kriging**	**trend**	**sim.**
abs. res.	913	979	762	833
wrong cl.	2696	2943	8031	5941

Sample size has a significant effect on both error measures. The improvement for bigger sample sizes is much stronger for IDW than for trend and the effect is stronger for wrong classification than for absolute residuals (Table 3).

Table 3. Average errors for different methods and sample sizes.

		49	**100**	**196**	**400**
absolute residuals	**IDW**	1144	980	813	714
	trend	804	775	751	718
wrong classification	**IDW**	4313	2900	2060	1512
	trend	8410	8019	7939	7756

4 Discussion

This study shows that maps from inverse distance weighted interpolation (IDW), ordinary kriging, and trend fitting can compete with those from simulations, if source term and weather are not known well. Kriging, even with variograms for subregions, was not superior to IDW. For interpolation, only small neighbourhoods should be used. Fitting parametric trends can capture the plumes quite well in terms of absolute residuals, especially for small sample sizes.

All mapping methods need further improvement: other interpolation methods like splines could be tested, interpolation could focus on the zero inflation of the data, or combining measurements and simulations may improve the result.

Acknowledgements

This work was funded by the European Commission under the Seventh Framework Program, by the Contract N. 232662. The views expressed herein are those of the authors and not necessarily those of the European Commission.

References

Beekhuizen, J. (2008). Dealing with uncertainty in determining the optimal locations of mobile measuring devices. Master thesis, Wageningen University, Netherlands.

Ehrhardt J., Weis A. (2000). "RODOS: decision support system for off-site nuclear emergency management in Europe". European Commission, Brussels, Report EUR 19144.

Kalnay, E,. Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., Joseph, D. (1996). "The NCEP/NCAR 40-year reanalysis project". Bulletin of the American Meteorological Society, Vol. 77 (3): 437-471.

Lund, U., Agostinelli, C. (2011). "circular: circular statistics", http://CRAN.R-project.org/package=circular [assessed 2011.03.26]

Mikkelsen, T., Thykier-Nielsen, S., Hoe, S. (2007), "Medium-range puff growth". Developments in Environmental Science, Vol. 6: 243-252.

Mikkelsen, T., Larsen, S., Thykier-Nielsen S. (1984), "Description of the Risø puff diffusion model". Nuclear Technology, Vol. 67: 56-65.

NOAA. http://www.esrl.noaa.gov/psd/ [accessed 2011.03.25]

Pebesma, E. (2004). "Multivariable geostatistics in S: the gstat package", Computers & Geosciences, Vol. 30: 683-691.

Pebesma, E., Skoien, J. *et al*. (2010). "intamap: procedures for automated interpolation", http://www.intamap.org/ [assessed 2011.03.26]

R Development Core Team (2010). "R: a language and environment for statistical computing", http://www.R-project.org/ [assessed 2011.03.25]

R Development Core Team and contributors worldwide (2011). "The R Stats Package", http://www.R-project.org/ [assessed 2011.07.11]

Schlather, M. (2010). "RandomFields: simulation and analysis of random fields", http://CRAN.R-project.org/package=RandomFields [assessed 2011.03.26]

WRF. http://www.wrf-model.org/index.php [accessed 2011.03.25]

Multivariate spatial outlier detection using geographically weighted principal component analysis

Paul Harris[1], Chris Brunsdon[2] & Martin Charlton[1]

[1] National Centre for Geocomputation, National University of Ireland Maynooth, Ireland
paul.harris@nuim.ie, martin.charlton@nuim.ie
[2] School of Environmental Sciences, University of Liverpool, Liverpool, UK
Christopher.Brunsdon@liverpool.ac.uk

Abstract

In this study, we apply a new methodology to detect outliers in multivariate spatial data sets, through the use of a spatial adaptation of principal components analysis (PCA). This non-stationary adaptation of PCA accounts for local spatial effects via the use of geographically-weighted data to form a geographically weighted PCA (GWPCA). Basic and robust GWPCA-based detection methods are calibrated, where the latter are preferable, as outliers can compromise any basic calibration prior to its use as a method of detection. Detection performance is investigated using a geochemical soils data set for countries bordering the Baltic Sea in Northern Europe. Here it is observed that a (local) GWPCA-based approach to outlier detection complements a (global) PCA-based approach, where the nature of a potential outlier, with respect to its unusual spatial and/or multivariate relationship to other observations can be more readily understood.

Keywords: robust, novelty detection, kernel weighting, PCA, non-stationarity

1 Introduction

Principal components analysis (PCA) is a widely used method in the physical and social sciences. It is commonly used to explain the covariance structure of a given data set using only a few components. The components are linear combinations of the original variables and can allow for a better understanding of differing sources of variation. In spatial settings, PCA is frequently applied without consideration for important spatial effects. Such naive applications can be problematic as spatial effects often provide a more complete understanding of the process. In this respect, a PCA can be replaced with a geographically weighted (GW) PCA (GWPCA) (Fotheringham *et al*., 2002, p196-202), when we want to account for a degree of spatial heterogeneity in the structure of the data set. In GWPCA, a different localised PCA is computed at every target location, and as such, the results vary continuously across space allowing them to be mapped.

There are many potential uses and extensions of GWPCA; where for this study we investigate its use as a means to detect outliers in multivariate spatial data sets. PCA-based methods are routinely used to detect multivariate outliers (e.g. Rousseeuw *et al*., 2006), but for spatial applications such global forms can only detect outliers in an aspatial manner. This oversight can result in a false positive identification, when an outlier's spatial neighbours are similar (in a multivariate sense), or a false negative when its spatial

neighbours are dissimilar (in a multivariate sense). Here a GWPCA-based detection method should minimise this particular form of misclassification.

We do not necessarily envisage that a GWPCA-based detection method should replace a PCA-based one, but instead, the two approaches can complement each other. PCA provides a broad, general sweep for outliers, whereas GWPCA provides a deep, more focused identification. Outlier detection can not only be used as a data cleaning or screening exercise, but can also be used to uncover interesting or unusual structures in the data that has not been considered before. In this respect, when data structures are known to vary across space, an application of a GWPCA-based detection method (which is implicitly designed to account for this spatial heterogeneity) is well chosen. Furthermore, the calibration of a PCA (and in turn a GWPCA) detection method is itself compromised by the existence of outliers (Rousseeuw *et al*., 2006), and as such, robust PCA/GWPCA forms are also calibrated to negate this problem. First insights into the detection performance for both basic/robust PCA/GWPCA forms are investigated using a soils geochemical data set for countries bordering the Baltic Sea.

2 Methodology

2.1 GWPCA

GWPCA suits situations when data are not described well by a universal set of components, but where there are spatial regions where a suitably localised set of components provide a better description. The technique uses a moving window weighting approach, where localised components are found at target locations. For an individual GWPCA at a target location, we weight all neighbouring observations according to some distance-decay kernel function and then locally apply standard PCA to this weighted data. The size of the window over which this localised PCA might apply is controlled by the bandwidth. Small bandwidth values lead to more rapid spatial variation in the results while large bandwidths yield results increasingly close to the universal (global) PCA solution.

Formally, a vector of observed variables $\mathbf{x}_i$ at spatial location i is assumed to have a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Further, if location i has coordinates (u,v), then GWPCA involves regarding $\mathbf{x}_i$ as conditional on u and v, and making $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ functions of u and v. Thus $\mathbf{x}_i \big| (u,v) \sim N(\boldsymbol{\mu}(u,v), \boldsymbol{\Sigma}(u,v))$, where $\boldsymbol{\mu}(u,v)$ and $\boldsymbol{\Sigma}(u,v)$ are the GW mean vector and the GW variance-covariance matrix, respectively. To find the GW principal components, the decomposition of the GW variance-covariance matrix provides the GW eigenvalues and GW eigenvectors. The GW variance-covariance matrix is $\boldsymbol{\Sigma}(u,v) = \mathbf{X}^{\mathrm{T}}\mathbf{W}(u,v)\mathbf{X}$ where $\mathbf{X}$ is the data matrix and $\mathbf{W}(u,v)$ is a diagonal matrix of geographic weights. We generate these weights using the bi-square kernel function:

$$w_{ij} = \left(1 - \left(d_{ij}/r\right)^2\right)^2 \text{ if } d_{ij} \le r \quad w_{ij} = 0 \quad \text{otherwise,} \tag{1}$$

where the bandwidth is the geographic distance r; and d_{ij} is the distance between spatial locations of the i^{th} and j^{th} rows in the data matrix. The GW components at location (u_i, v_i) can be written as $\mathbf{LVL}^{\mathrm{T}} \big| (u_i, v_i) = \boldsymbol{\Sigma}(u_i, v_i)$ where $\mathbf{L}$ is a matrix of GW eigenvectors; $\mathbf{V}$ is a diagonal matrix of GW eigenvalues; and $\boldsymbol{\Sigma}(u_i, v_i)$ is the GW variance-covariance matrix. Thus for a GWPCA with m variables, there are m components, m eigenvalues, m sets of component loadings, and m sets of component scores at

each data location. We can also obtain eigenvalues and their associated eigenvectors at unobserved locations, although as no data exists for these locations, we cannot obtain scores.

2.2 Bandwidth selection in GWPCA

To describe how the bandwidth r is selected in GWPCA, we first discuss properties of PCA. Thus if there are m variables in the data matrix, so that each observation is a vector in *m*-dimensional space, the scores corresponding to components $q+1$ to m, represent the Euclidean distances along the axes of the corresponding orthogonal vectors to a *q*-dimensional linear sub-space. Here, the *q*-dimensional sub-space is spanned by the first q loadings (viewed as *m*-dimensional vectors), and is the sub-space that maximised the variance of the data points projected on to that sub-space. Here, q is commonly chosen so that this sub-space contains a reasonably high proportion of the total variance, and thus components $q+1$ to m represent the deviation from this sub-space.

Suppose that $\mathbf{M}_q$ denotes the matrix $\mathbf{M}$ with all but the first q columns removed and $\mathbf{M}_{(-q)}$ denotes the matrix $\mathbf{M}$ with the first q columns removed. For PCA, the first q components are described by $\mathbf{XL}_q$ and the remaining components by $\mathbf{XL}_{(-q)}$ (where $\mathbf{X}$ and $\mathbf{L}$ are the data and eigenvector matrices, respectively). It is possible to show that the best (least squares) rank q approximation to $\mathbf{X}$ is $\mathbf{XL}_q\mathbf{L}_q^{\mathrm{T}}$ and that the residual matrix from this $\mathbf{S}$, given by $\mathbf{S}=\mathbf{X}-\mathbf{XL}_q\mathbf{L}_q^{\mathrm{T}}$ can also be written as $\mathbf{S}=\mathbf{XL}_{(-q)}\mathbf{L}_{(-q)}^{\mathrm{T}}$ (Jollife 2002). In effect, via principal components, we find the minimum of the expression $\sum_{ij}\left([\mathbf{X}]_{ij}-[\mathbf{S}]_{ij}\right)^2$ with respect to $\mathbf{S}$ where $\mathbf{S}$ is a rank q matrix; and the problem is solved with the given expression. The variance levels of the components of the matrix $\mathbf{S}$ therefore measure the 'goodness of fit' (GOF) of the projected sub-planes and as such $\mathrm{GOF}_{\mathrm{i}}=\sum_{j=q+1}^{j=m}s_{ij}^2$ is the GOF for the i^{th} observation and s_{ij} is the j^{th} component score for observation i; that is, the ij^{th} element of $\mathbf{S}$. The total GOF for the entire data set is $\mathrm{GOF}=\sum_{i=1}^{i=n}\mathrm{GOF}_i$.

For GWPCA, the GW components for the i^{th} location represent a similar projection, but with the corresponding loadings defined locally. That is, we find $\mathbf{S}$ to minimise $\sum_{ij}w_i\left([\mathbf{X}]_{ij}-[\mathbf{S}]_{ij}\right)^2$ where w_i is a locally defined weight for location i. GOF statistics for GWPCA can be defined in an analogous fashion as for PCA; except that in each locality, $\mathbf{S}$ is defined using local weights, as above. In turn, a total GOF statistic provides a means of finding an optimal bandwidth for GWPCA, where we use a 'leave-one-out' method to compute the terms of the statistic. Here, a 'leave-one-out' total GOF statistic is computed for all possible bandwidths and an optimal bandwidth relates the smallest GOF value found. Bandwidths can be in a fixed or adaptive form, where for this study only the latter are specified.

2.3 Outlier detection with basic and robust GWPCA

For outlier detection with GWPCA, we adopt a simple approach where we flag data locations that contribute the most to the minimised 'leave-one-out' total GOF statistic,

above. Here for a given GWPCA calibration, we find (standard) boxplot statistics of this 'residual' data $\sqrt{\sum_{ij} w_i \left([\mathbf{X}]_{ij} - [\mathbf{S}]_{ij}\right)^2}$, which is an output data set of sample size. A data location i is deemed outlying if its 'residual' lies beyond the extremes of the boxplot's whiskers. This 'residual' data approach is analogous to the identification of outliers via large residuals from a regression fit.

As for PCA, GWPCA is sensitive to outliers and as such compromises its ability to detect them. In this respect, we also calibrate a robust GWPCA form by replacing the GW mean calculations with GW medians and the decomposition of the GW variance-covariance matrix is done robustly using the algorithm of Hawkins *et al*., (2001). As a robust GWPCA is extremely computer intensive, we calibrate it using optimal bandwidths found from its basic GWPCA counterpart.

For GWPCA, a difficulty lies in that we need to decide *a priori* upon the number of components to retain, i.e. the value of q. Furthermore, we cannot find an optimal bandwidth and in turn identify outliers, if we wish to retain all m components. Therefore in experimentation, we choose to find optimal bandwidths and identify outliers, for all values of q (except when $q = m$), and flag outlying data locations for each of the q calibrations. Thus for our case study data set of $m = 10$ variables, a GWPCA is calibrated $q = 9$ times. From each calibration, a weight of evidence is then built up, which determines the potential of an outlying data location. Evidence is strongest for an outlier if a data location is flagged all nine times.

Observe that with a GWPCA-based detection method, a vector of observed variables at a location *i* is deemed outlying and not one particular observation at this location. This is consistent with standard PCA-based outlier detection methods. Observe also that our detection method is not a direct GW adaptation of some standard (commonly, robust) PCA-based method, where all *m* components are retained. Such an adaptation is eminently viable and is left for future research.

3 Case study data

The study data stems from the sampling of agricultural soils over a large region surrounding the Baltic Sea, at 768 sites (Reimann *et al*., 2000). We concentrate on topsoil samples; and in particular on ten major trace elements, reflected in the compounds: SiO_2, TiO_2, Al_2O_3, Fe_2O_3, MnO, MgO, CaO, Na_2O, K_2O and P_2O_5. We standardize this data and specify all PCA/GWPCAs with covariance matrices. A PCA output reveals that the first principal component contrasts SiO_2 with the other trace elements and accounts for 52.6% of the total variance. The first three components collectively account for 78.4% of the total variance. It is reasonable to ask whether this analysis, should be applied over the whole of the study area. To answer this, a series of GWPCAs for each value of q were undertaken and their outputs mapped. Significant spatial non-stationarities were observed in the structure of the data, suggesting: (i) value in GWPCA itself, as an exploratory tool; and (ii) a GWPCA-based method to outlier identification is likely to be worthwhile.

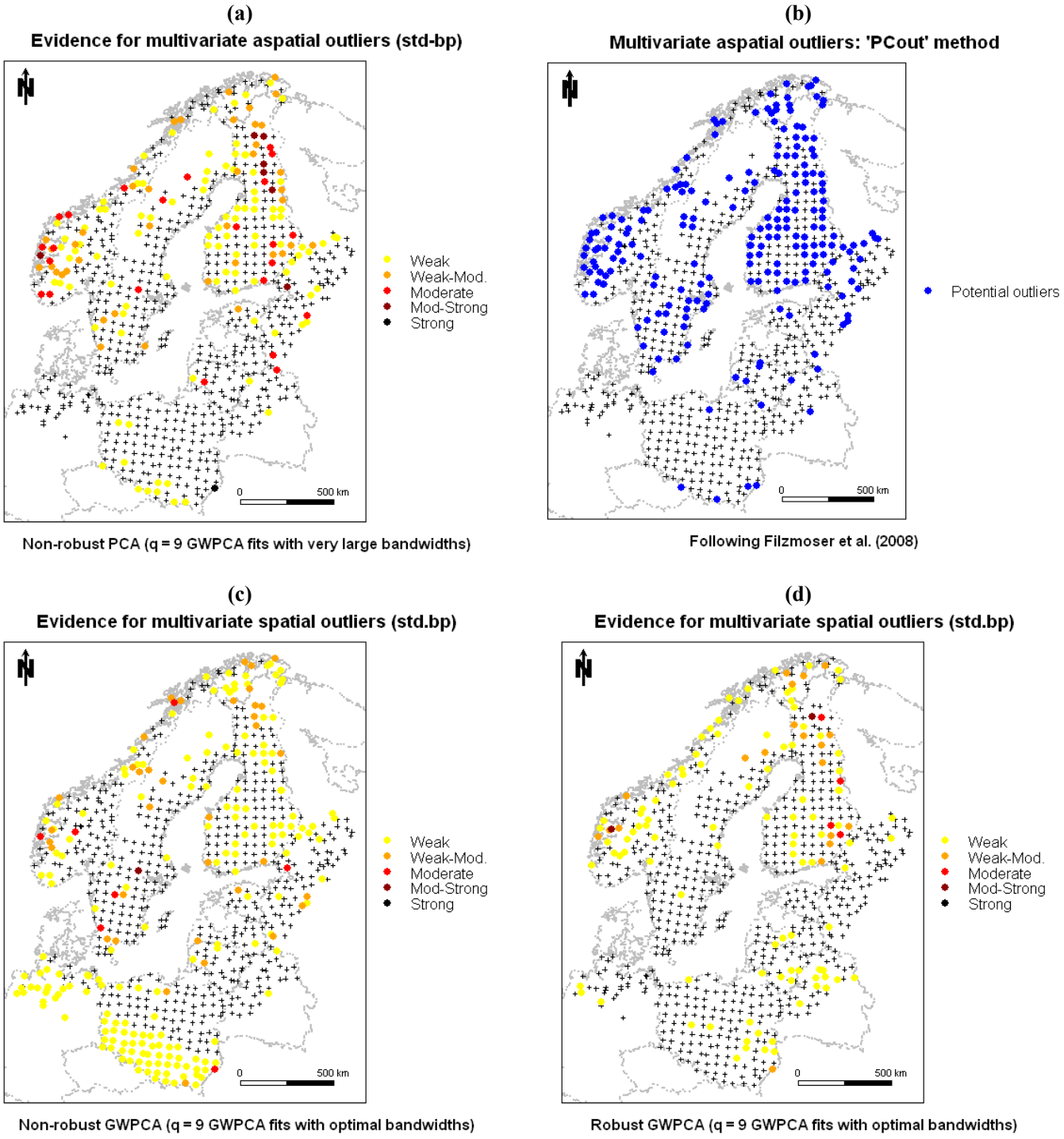


Figure 1. Detection comparisons: (a) non-robust GWPCA fits with very large bandwidths; (b) the robust 'PCout' method of Filzmoser *et al*. (2008); (c) non-robust GWPCA fits with optimal bandwidths; and (d) robust GWPCA fits with optimal bandwidths. For (a), (c) and (d), a GWPCA is calibrated nine times and outlying data locations noted. If a data location is flagged as outlying on 1-2 occasions it is assigned a weak evidence label and so forth.

4 Preliminary analyses: multivariate spatial outlier detection

To our knowledge, there are no direct competitors to our GWPCA-based method for multivariate spatial outlier detection. However, we can go some way in assessing its relative worth by calibrating the individual GWPCAs with very large bandwidths (so the detection method is effectively PCA-based) and comparing its output to that found with an existing PCA-based method. In this respect, we do this simply by a visual comparison

of outlier maps. Further work will look at differences and similarities between methods more deeply. In Figs. 1a-b, an example comparison is given, where our (now aspatial) GWPCA-based method (in basic form) performs in a broadly similar manner to a more recognised method, thus providing some assurance when we come to use GWPCA correctly and locally. It appears that multivariate outliers are most prominent in Southern Norway and much of Finland, when we detect them in this purely aspatial manner.

In Figs. 1c-d, we compare basic and robust GWPCA methods for multivariate spatial outlier detection, our study goal. These maps not only need to be compared to each other, but also to those in Figs. 1a-b. Many differences can be observed between the aspatial and the spatial detection methods, many of which relate to known non-stationarities in data structure (section 3). For example, it is hypothesized that the numerous outliers identified in Finland using the aspatial methods are not in fact unusual in a spatial sense and as such, a (spatial) GWPCA method does not flag as much of this data as outlying. Furthermore, the basic GWPCA method detects a swathe of outliers in southern Poland (Fig. 1c), but this detection is largely absent using the robust version, suggesting that basic GWPCA is compromised by a few outliers in this region (or possibly just one in the south-east, on the border).

5 Conclusion

In this study we have introduced a method of detecting outliers in multivariate spatial data sets, based on GWPCA. The method of detection holds promise and can fill a gap that is missing in the literature. Further clarity on this promise is work in progress. Here we aim to demonstrate objectively the value of this method using contaminated sample data, where (known) outliers are introduced and the method's detection rate (of false negatives) is assessed accordingly.

Acknowledgements

This research was funded by a Strategic Research Cluster grant (07/SRC/I1168) by the Science Foundation Ireland under the National Development Plan.

References

Filzmoser, P., Maronna, R., Werner, M. (2008) Outlier identification in high dimensions. *Computational Statistics and Data Analysis* 52:1694-1711

Fotheringham, A.S., Brunsdon, C, Charlton, M.E. (2002) *Geographically Weighted Regression - the analysis of spatially varying relationships.* Wiley, Chichester

Hawkins, D., Li L., Young, S. (2001) *Robust Singular Value Decomposition. National Institute of Statistical Sciences*, Technical Report Number 122

Joliffe, I.T. (2002) *Principal Components Analysis*, 2nd edition, New York: Springer-Verlag.

Reimann, C., *et al*. (2000) Baltic soil survey: Total concentrations of major and selected trace elements in arable soils from 10 countries around the Baltic Sea. *The Science of the Total Environment* 257:155-170

Rousseeuw PJ, Debruyne M, Engelen S, Hubert M (2006) Robust and outlier detection in chemometrics. *Critical Reviews in Analytical Chemistry* 36:221-242

Assessing the debris around glaciers using remote sensing and random sets

Mus Bandishoev [1], Arta Dilo[2] & Alfred Stein[1]

[1] University of Twente, Faculty of Geo-Information Science and Earth Observation (ITC), Enschede, The Netherlands
[2] University of Twente, Faculty of Electrical Engineering, Mathematics and Computer Science, Enschede, The Netherlands

Abstract

Glacier mapping from satellite multispectral image data is hampered by debris cover on glacier surfaces. Information on the spatial distribution and spatial-temporal dynamics of debris, however, bears various kinds of uncertainties. Debris exhibits the same spectral properties as lateral and terminal moraines and as bedrock outside the glacier margin. Multispectral classification alone is thus not suitable to properly assess its extent. Additional information has to be included, like the low slope angles and curvature characteristics. In this research we propose a random set method for uncertainty modelling of debris-covered glaciers extracted from remote sensed data. Here, we analyse the Fedchenko glacier situated in the Pamir mountains in Central Asia. Clean glacier ice and debris area are represented by random sets. Their statistical mean and median are estimated. The paper combines the advantages of an automated multispectral classification for clean glacier ice and snow with slope information derived from the digital elevation model (DEM). We use an SRTM3 DEM that is resampled to 30m. From a 1999 Landsat ETM+ image the results show that the mean area of clean glacier ice equals 841.87 km^2, and 94.39 km^2 for debris-covered area. Temporal analysis shows that the mean area of clean ice increased from 1992 to 1999 and is decreasing since 1999, in opposite to the debris covered area. We conclude that this method based on random set theory has the potential to serve as a general framework in uncertainty modelling of debris-covered glaciers and is applicable for mountainous glaciers.

Keywords: Random set theory, Uncertainty modelling, Glacier mapping, Debris cover, DEM analysis

1. Introduction

Due to the remoteness and inaccessible nature of mountain glaciers, remotely sensed data are an efficient tool for regular mapping of glaciers in a comprehensive and effective manner. A number of remote sensing techniques for automated mapping of clean glacier ice by means of multispectral classification are available. Commonly used techniques such as single band ratios and Normalized Difference Snow Index (NDSI) take advantage of the high brightness of snow and ice in the visible wavelength to separate them from darker areas such as rock, soil or vegetation. The greatest difficulty in glacier mapping from remote sensed data, however, is the presence of debris on glaciers. A debris-covered

glacier area has a similar visible and near-infrared spectral signature to the surrounding terrain and thus complicates the mapping of glaciers (Bolch and Buchroithner 2007). Here, the traditional multi-spectral classification techniques are of limited value. A number of methods have been proposed to address this problem. These methods use additional information provided by topography (Bishop, Bonk *et al*. 2001), neighbourhood analysis (Paul, Huggel *et al*. 2004) and thermal radiation (Taschner and Ranzi 2002).

Information on spatial distribution of debris-covered glaciers from remote sensed data bears various kinds of uncertainties. Existing techniques for mapping debris-covered glaciers are crisp-based and have limitations in delineation of glaciers boundary where the transition between the debris-covered glacier and the adjacent terrain is gradual. A thresholding segmentation technique can be used in principle for identification of clean glacier ice from band ratio images and the debris-covered areas from glacier slope image (Paul, Huggel *et al*. 2004). Selection of a threshold value, however, is a critical task as a slight change can lead to overestimation or underestimation of the areal extent. The threshold value may be different for different satellite sensors and for different seasons (Dozier 1989; Hall, Riggs *et al*. 1995). Since it is arbitrary to choose a single-valued threshold, uncertainties exist in any segmentation results and can have a large effect on the subsequent spatial analysis (Lucieer and Stein 2002).

To investigate inherent uncertainties in observations of glaciers from satellite imageries, this study proposes a random set method for uncertainty modelling of debris-covered glaciers. A glacier with uncertainties can be treated as a randomly varying set, i.e. a random set. In this research we show that random sets can serve as a framework to model debris-covered glaciers with inherent uncertainties. The Fedchenko glacier situated in the north-western part of the Pamir mountains, Tajikistan, was chosen as a study case for this research. The southern end of the glacier basin is located at 38º30´16´´N, 72º17´00´´E; the northern end at 39º05´10´´N, 72º18´52´´E.

In the following section we elaborate on the data used and the method proposed by this research. Section 3 presents results of the data processing for the Fedchenko glacier. Section 4 concludes the paper.

2. Data and method

2.1 Data used

Landsat images from the TM and the ETM+ sensors are used in this research due to their availability, and their moderate spatial and spectral resolution. For observations on glaciers it is important to work with images that have no cloud over the glaciated area and that have been taken at the end of the ablation season when the temporary snow cover is at its minimum and all the glacier zones can be clearly demarcated. These factors restrict the use of most available imagery. In total two orthorectified Landsat TM images (September 1992 and 2009) and one orthorectified ETM+ image (September 1999) were used.

A digital elevation model (DEM) is used to derive the slope information needed for identification of debris-covered area of a glacier. For the study area the only available DEM were a Shuttle Radar Topography Mission (SRTM3) DEM and an ASTER GDEM. The SRTM3 DEM is used in the research as it has a higher accuracy. A cubic convolution method of interpolation is used to resample the SRTM3 DEM to a 30 m resolution. Absolute vertical accuracies of DEMs were measured by comparison with ground control

points (GCP) – elevation points extracted from Russian Topographic Maps (published data: 1979) distributed throughout the study area.

2.2 Method

The Normalized Difference Snow Index (NDSI) is employed for the identification of glacier snow and ice (GSI). Glacier surfaces that are covered by debris have a gentle slope, whereas at the contact edge of the glacier with the surroundings or bedrock, a distinct change in slope can be observed. This edge can be utilized for a delineation of debris-covered glaciers (Paul, Huggel *et al*. 2004). In this research we used this idea for the identification of plain areas, where potentially debris-covered areas of glacier might be situated. Plain areas are further processed for extracting the debris covered area (DCA).

For uncertainty modelling of debris-covered glaciers a random set model was applied. Thresholding approach of image segmentation was used to generate a random set. The idea of random set generation is that the extents of the two classes, GSI and plain areas, extracted from NDSI and slope images, respectively, are sensitive to the different thresholds. Therefore by slightly changing a threshold, a set of objects is generated. These form the focal elements of a random set.

To map a debris-covered glacier we use the NDSI and slope images. An NDSI image, which has values from -1 to 1, was segmented using a threshold value to obtain a binary image. A range of thresholds was selected combining values proposed in the literature and inspection of the images, as the human eye can estimate the correct values by using textural features. The range of threshold values was divided into *n* equal intervals, resulting into *n+1* thresholds to produce the binary images. Slope information is used to delineate the plain areas where potentially DCA is situated. The minimum threshold value is defined from the mean value of slope calculated from the cross-profiles to glacier body, whereas the maximum threshold value is set equal to 24°, being an upper limit for the steepness that a glacier might have (Paul, Huggel *et al*. 2004).

The covering functions of the generated random sets give the probability of an image pixel to be GSI or to be situated on plain areas. Suppose, ξ is an image pixel in Euclidian space $\mathbb{R}^2$: $\xi \in I \subset \mathbb{R}^2$ with pixel size *r* and a slope value *d*, and $\mathcal{A}_i$, i ={1,2, …,n} and $\mathcal{B}_j$, j ={1,2, …,m} are the focal elements of random sets $\mathcal{A}$ and $\mathcal{B}$ generated from the NDSI and the slope image, respectively. The covering function Γ of the random set $\mathcal{A}$ gives the probability for every pixel to be covered by the set $\mathcal{A}$. The probability of pixel ξ to be in the random set $\mathcal{A}$ is calculated as (Molchanov 1993),

$$\Pr_{\Gamma}(\xi)= \frac{1}{n}\sum_{i=1}^{n} I_{\mathcal{A}_i}(\xi)$$

where $I_{\mathcal{A}_i}$ is the indicator function of $\mathcal{A}_i$ defined as

$$I_{\mathcal{A}_i} = \begin{cases} 1, & \xi \in \mathcal{A}_i \\ 0, & \xi \notin \mathcal{A}_i \end{cases}$$

Due to the rough mountainous terrain the identification of debris-covered areas requires additional analysis to slope bounding. Because plain areas ($0°<d<24°$) occur everywhere in the study area, it is necessary first to eliminate plain areas that are not a part of the glacier. Only plain areas connected to GSI can be part of the debris cover. The covering function Π of plain areas is calculated from the focal elements of the random set $\mathcal{B}$ in the same way as Γ. The support set $\Pi_s=\{\xi \in \mathbb{R}^2: \Pr_{\Pi}(\xi)> 0\}$ describes the possible part of the debris-covered area. It consists of N detached components, $\Pi_s=\bigcup_{j=1}^{N} \Pi_j$. We exclude those areas that are not connected to the possible part of clean glacier ice: $\Gamma_s \cap \Pi_j= \emptyset$, and recalculate Π accordingly.

The second step is calculation of the covering function Δ of DCA. The GSI area of glaciers occurs in plain areas along with DCA, there is no DCA on a plain area but GSI. In other words, for a pixel that is possibly covered by plain areas, its probability to be covered by DCA depends whether the pixel has been classified as GSI or not. If the probability of an image pixel being covered by plain areas is greater than 0 and its probability to be GSI is positive: $Pr_{\Gamma}(\xi)>0$ and $Pr_{\Pi}(\xi)>0$, then the probability to be in DCA is $Pr_{\Delta}(\xi) = \min\{Pr_{\Pi}(\xi), 1 - Pr_{\Gamma}(\xi)\}$, otherwise $Pr_{\Delta}(\xi)= Pr_{\Pi}(\xi)$.

The level sets ($\Gamma_p = \{\xi \in \mathbb{R}^2: Pr_{\Gamma}(\xi) \geq p\}$) are used to reflect the spatial distribution of the varying sizes of the random sets to quantify the extensional uncertainty of segmented objects. The mean area EA of the random sets Γ (GSI) is determined by

$$EA(\Gamma) = r \times \sum_{\xi \in I} Pr_{\Gamma}(\xi).$$

The Vorob'ev expectation as an estimation of the mean is different from the median set of Γ, defined as the 0.5-level set. The mean of the random set Γ is estimated by first determining the mean area EA(Γ), and then finding a p-level set for Γ which has the area equal to EA(Γ) (Zhao, Stein *et al*. 2010).

The set-theoretic variance of a random set Γ is defined as:

$$\Gamma_{var}(\xi) = \sum_{i=1}^{N} (I_{\mathcal{A}_i}(\xi) - Pr_{\Gamma}(\xi))^2,$$

whereas the sum of the Γ_{var}, denoted as SD, as:

$$SD = r \times \sum_{\xi \in I} \Gamma_{var}(\xi)$$

and the coefficient of variation (CV) as CV=SD/EA, being a normalized and dimensionless measure. The CV summarizes the dispersion of the distribution of a random set. A high CV indicates a larger proportion of objects with a high Γ_{var} or Δ_{var} and thus points to a large extensional uncertainty (Zhao, Stein *et al*. 2010). Extensional uncertainty for GSI and DCA is identified by

$$\frac{\sum_{\xi \in I} \Gamma_{var}(\xi)}{\sum_{\xi \in I} Pr_{\Gamma}(\xi)} \text{ and } \frac{\sum_{\xi \in I} \Delta_{var}(\xi)}{\sum_{\xi \in I} Pr_{\Delta}(\xi)}.$$

3. Results

The mean area of GSI and DCA in 1999 equals 841.87 km^2 and 94.39 km^2, respectively. Debris cover amounts to around 10% of the total glaciated area. The differences between the areas of support set and core set (Γ_1, Δ_1) indicate the extensional uncertainty. These are 50.78 km^2 and 33 km^2 for GSI and DCA, respectively, constituting 6% of GSI and 35% of DCA areas respectively. The higher uncertainty corresponds to a higher variance of random sets (Bandishoev 2011).

Due to the rough mountainous terrain, each terrain aspect has specific properties. For example, there is more snow accumulation in the north aspect due to less solar illumination in comparison with the south. We use the aspect information derived from DEM to quantify the uncertainty of GSI and DCA. The results are given in Figure 1.

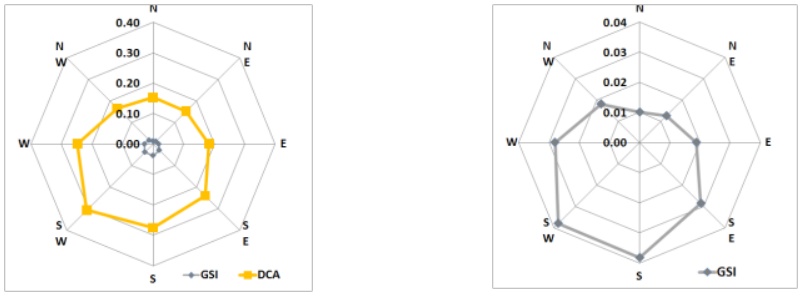


Figure 1. Coefficient of variation of GSI and DCA areas versus aspect.

A large extensional uncertainty of GSI occurs on the south-western and southern aspects of the terrain (CV=0.038) and to a lesser degree on the northern aspect (CV=0.01). With the sun azimuth (143.55) and sun elevation (51.93) of ETM+ image used in this research, the reason of relatively large extensional uncertainty might be saturated pixels on southward aspects. The total CV for GSI equals 0.021, being an indicator of a smaller extensional uncertainty. For DCA a rather large extensional uncertainty occurs (CV=0.2), and most of it occur on the south-westrn and southern aspects. The reason of such a large uncertainty is the rough mountainous terrain.

For the temporal analysis we use three images (date of acquisition: 1992, 1999, 2009) from Landsat TM and ETM+ sensors. The idea is to show the change in the debris-covered glacier extents and to quantify the uncertainties. The temporal DEM for the study area is only available for February 2000. Thus, to perform temporal analysis we assume that the DEM did not changed from 1992 to 2009. The support set, mean and core set areas of GSI and DCA, together with the coefficient of variance (CV) are calculated, and results are shown in Figure 2.

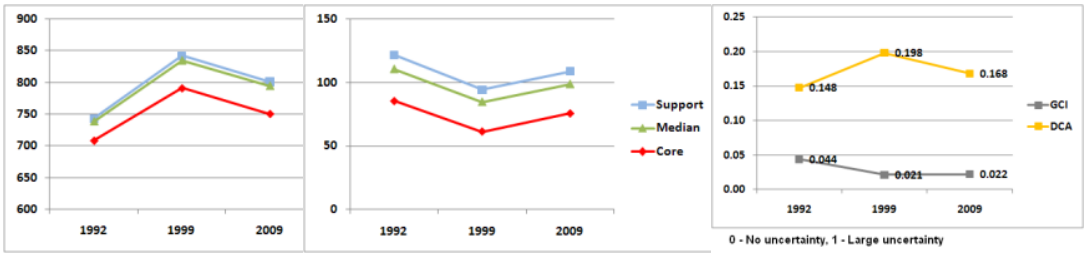


Figure 2. Mean, median and support areas for GSI (left) and DCA (middle) per year. The variation of uncertainty in term of CV per year (right).

To validate the method we use digitized glacier boundaries as a reference. As the study area covers a wide glaciated area we use two different areas in order to account for the location variation of the terrain and roughness where we focus on the core set of debris-covered areas. The overall accuracy equals 87.58% for the smooth area and 73.5% for the rough area.

4. Discussion and conclusion

This study applies a random set model for uncertainty modelling of a debris-covered glacier. Uncertainties are modelled for glacier snow and ice and for debris-covered areas of a glacier separately. This partitioning allows us to quantify the uncertainty for both constituent parts of a glacier. This is important as in some cases glacier areas covered by debris compose a large part of the glacier surface, and ignoring them will lead to misclas-

sification. For example, in this research we found that about 10% of Fedchenko glacier is covered by debris.

By using the statistical parameters of random sets (support, mean, median, variance) the study demonstrates that the randomness of segmentation thresholding parameters has different effects on extracted snow and ice and debris-covered areas. Taking into account the rough mountainous terrain of the glacier these parameters quantify the uncertainty versus aspects. We find that, for both components of the glacier, the extensional uncertainty into southward direction of the terrain is twice as large as in the northward direction.

The temporal uncertainty modelling shows that the mean area of snow and ice increased from 1992 to 1999 and it is decreasing since 1999, as opposed to the pattern for the debris-covered area. The correlation between ice and debris-covered glacier area can be interpreted as the occurrence of debris where ice or snow melts.

The result of the uncertainty modelling for debris-covered glaciers proves that a random set approach is an effective tool for modelling and quantification of uncertainties. This method can thus be used for the assessment of debris-covered glaciers and their changes in time, being of a vital importance for planning and management of water resources as in the IPCC studies.

References

Bandishoev, M. M. (2011). The quality of glacier observation: the debris areas and their role in size estimation. Master thesis, University of Twente, ITC, Enschede.

Bishop, M., R. Bonk, *et al*. (2001). "Terrain analysis and data modelling for alpine glacier mapping." *Polar Geography* 25: 182-201.

Bolch, T. and M. Buchroithner (2007). "An Automated Method to Delineate the Ice Extension of the Debris-Covered Glaciers at Mt. Everest Based on ASTER Imagery." *Grazer Schriften der Geographie und Raumforschung* 43: 71-78.

Dozier, J. (1989). "Spectral Signature of Alpine Snow Cover from the Landsat Thematic Mapper " Remore Sensing of Environment 28: 9 - 22.

Hall, D. K., A. Riggs, *et al*. (1995). "Development of methods for mapping global snow cover using moderate resolution imaging spectroradiometer data." *Remote Sensing of Environment* 54(2): 127-140.

Lucieer, A. and A. Stein (2002). "Existential uncertainty of spatial objects segmented from satellite sensor imagery." *Geoscience and Remote Sensing* 40(11): 2518-2521.

Molchanov, I. (1993). "Intersections and shift functions of strong Markov random closed sets." *Probability and mathematical statistics* 14: 265-279.

Paul, F., C. Huggel, *et al*. (2004). "Combining satellite multispectral image data and a digital elevation model for mapping debris-covered glaciers." *Remote Sensing of Environment* 89(4): 510-518.

Taschner, S. and R. Ranzi (2002). Comparing the opportunities of LANDSAT-TM and ASTER data for monitoring a debris covered glacier in the Italian alps within the GLIMS project. *Proceedings IGARSS*: 1044-1046.

Zhao, X., A. Stein, *et al*. (2010). "Application of random sets to model uncertainties of natural entities extracted from remote sensing images." *Stochastic Environmental Research and Risk Assessment* 24(5): 713-723.

Assessing the spatial variability of the accuracy of multispectral images classification using the uncertainty information provided by soft classifiers

Cidália C. Fonte[1,2] & Luísa M. S. Gonçalves[1,3]

[1] Institute for Systems and Computers Engineering at Coimbra, Portugal
[2] Department of Mathematics, University of Coimbra, Portugal
cfonte@mat.uc.pt
[3] Polytechnic Institute of Leiria, Civil Engineering Department, Portugal
luisag@estg.ipleiria.pt

Abstract

In this paper a methodology is proposed that enables the estimation of the spatial variation of the accuracy of the classification of multispectral images performed with soft classifiers. The use of soft classifiers enables the computation of an uncertainty index for each pixel, which reflects the difficulty found by the classifier to assign a class to the pixel. Since the uncertainty is computed for all pixels, an image of the spread of uncertainty may be built and used to identify regions where the classification is more likely to present different degrees of accuracy. To identify these regions, an image representing the pixel uncertainty is segmented and the mean uncertainty is computed within each region. The classification accuracy may now be computed for each of the regions, providing information about the spatial variation of accuracy. The proposed methodology is applied to a case study.

Keywords: accuracy, uncertainty, soft classifiers, multispectral images.

1 Introduction

The accuracy assessment of multispectral image classification is fundamental to assess the quality of the land cover maps resulting from the classification process. The accuracy assessment of land cover maps is usually performed with confusion matrixes. Confusion matrixes are built considering a sample of points, which may be obtained considering several sampling techniques. A reference database is built for these points, with the classification information and the ground truth assigned to each point. With these matrixes class and overall accuracy measures may be computed, which apply to the whole map. However, the map accuracy may not be homogeneous along the map, that is, regions with higher and lower accuracy values may exist.

Foody (2005) proposed the computation of geographically constrained confusion matrixes, derived for parts of the image, instead of only one confusion matrix to the whole image. To generate these confusion matrixes only sample points located in the vicinity of points of interest were used.

Woodcock *et al.* (2001), computed separate accuracy matrixes for two strata, one corresponding to error least likely stratum and error most likely strata, but no spatial location was assigned to these two levels of error.

Gonçalves *et al.* (2010b) and Gonçalves *et al.* (2009) showed that uncertainty measures may be used as indicators of the classification accuracy. Therefore, the identification of regions with different degrees of uncertainty may be an indicator of the regions where different degrees of accuracy are more likely to occur. In the study herein presented, to identify regions with potentially different levels of accuracy a segmentation of the pixel uncertainty information is made, and the mean uncertainty within those regions is computed. A set of sample points is then chosen inside the regions, and a confusion matrix computed. The proposed methodology is applied to a case study.

2 Methodology

The approach proposed in this paper includes the following steps: 1) classification of the multispectral images with a soft classifier; 2) compute the classification uncertainty; 3) segment the image representing the spatial variation of the classification uncertainty; 4) computed the mean uncertainty within the segmented objects 5) identify regions apparently problematic or regions of interest; 6) consider a random sample of points within the regions identified in the previous step; 7) build a reference database for the points identified in the previous step; 8) build a confusion matrix with the reference database.

3 Case Study

3.1 Data

The study was conducted in a region of the south of Portugal. The area is occupied mainly by agriculture, pastures, forest and agro-forestry areas, where the dominant forest species are eucalyptus, coniferous and cork trees. An image obtained by the IKONOS sensor was used, with a spatial resolution of respectively 1m in the panchromatic mode and 4m in the multi-spectral mode (XS) and a dimension of 4 900 m by 3 708 m. The geometric correction of the multi-spectral image consisted of its orthorectification. The average quadratic error obtained for the geometric correction was 1.39 m, inferior to half the pixel size, which guarantees an accurate geo-referencing.

3.2 Classification

The classification method used is a pixel-based supervised fuzzy classifier based on the Minimum-Distance-to-Means classifier, available in the commercial software IDRISI. With this method, the image is classified based on the information contained in a series of signature files and a standard deviation unit (Z-score distance) chosen by the user. The fuzzy set membership is calculated based on a standardized Euclidean distance from each pixel reflectance, on each band, to the mean reflectance for each class signature, using a sigmoid membership function (Burrough and McDonnell, 1998; Kuncheva, 2000). The underlying logic is that the mean of a given signature represents the ideal point for the class, where fuzzy set membership is one. When distance increases, fuzzy set membership decreases, until it reaches the user-defined Z-score distance where fuzzy set membership decreases to zero. To determine the value to use for the standard deviation unit,

the information of the training data set was used to study the spectral separability of the classes and to determine the average separability measure of the classes.

Unlike traditional hard classifiers, the output obtained with this classifier is a set of images (one per class) that expresses the possibility that each pixel belongs to the class in question. If the class corresponding to the higher possibility value for each pixel is considered, a hard classification is obtained. Figure 1a) shows a composite image obtained with the bands red, green and blue (RGB 321); and b) the hard version of the classification.

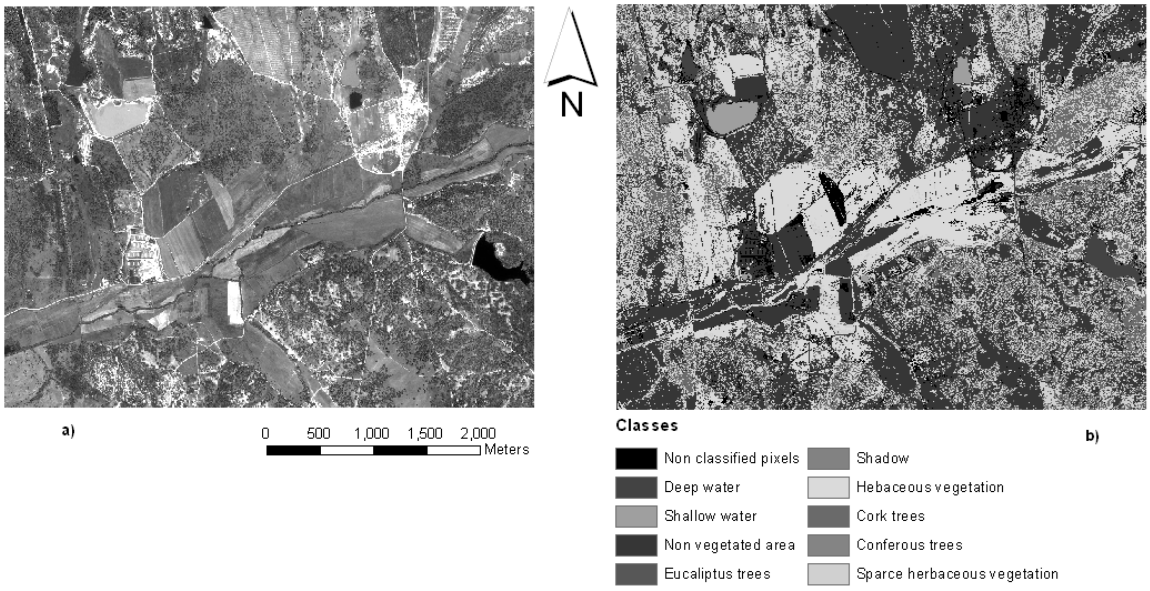


Figure 1. a) Composite image of the bands red, green and blue (RGB 321). b) hardened classification results.

3.3 Uncertainty

Several uncertainty measures may be used to evaluate the classifiers difficulty to assign only one class to each pixel (Gonçalves *et al.*, 2010a; Gonçalves *et al.*, 2010b). In this paper, the uncertainty associated with this assignment is evaluated with the Relative Maximum Deviation Measure, available in IDRISI software. The uncertainty is computed using (1), where $\pi_i(x)$ is the degree of possibility associated to class i and n is the number of classes.

$$RI(x) = 1 - \frac{\max_{i=1,..n}\left(\pi_i(x)\right) - \frac{\sum_{i=1}^{n}\pi_i(x)}{n}}{1-\frac{1}{n}} \quad (1)$$

This uncertainty measure assumes values in the interval [0,1] and evaluates the degree of compatibility with the most possible class and until which point the classification is dispersed over more than one class. Figure 2 shows the spatial distribution the uncertainty.

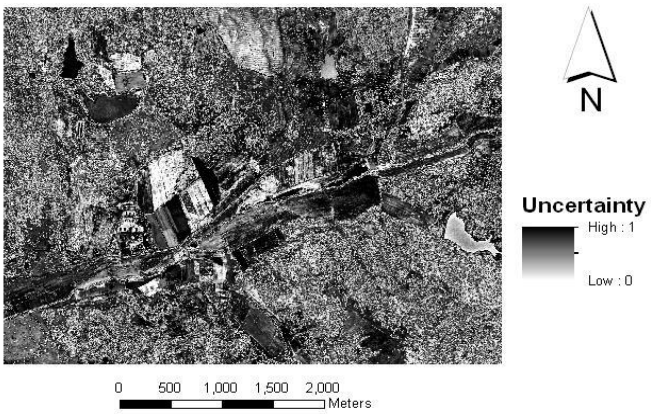


Figure 2. Spatial distribution of uncertainty.

3.4 Segmentation of uncertainty

The segmentation of uncertainty was made with the algorithm available in software IDRISI. This algorithm groups adjacent pixels into image segments according to their similarity. It employs a watershed delineation approach to partition input imagery based on their variance. A derived variance image, obtained using a moving window, is treated as a surface image allocating pixels to particular segments based on variance similarity. The values adopted were: 3 for moving the window size, 0.5 for the weight of mean and variance; and for the similarity tolerance 200 and 300. Figure 3 shows the results obtained with these two parameters, where the shades of grey represent the mean value of the pixels uncertainty within each region. For the segmentation with the similarity tolerance of 200 the mean value of uncertainty ranges between 0.22 and 0.91, while for the segmentation with the similarity tolerance of 300 ranges between 0.46 and 0.80. This shows that there are effectively regions on the map with very different levels of uncertainty.

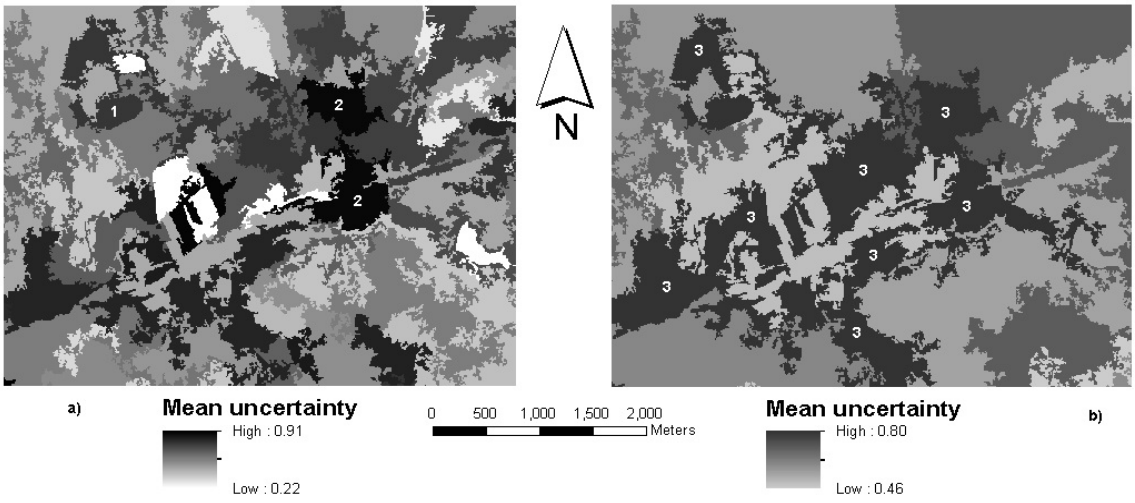


Figure 3. Segmentation of uncertainty with a similarity tolerance of 200 in a) and 300 in b)

3.5 Accuracy assessment

To assess the accuracy of the hardened classification, a stratified random sample containing between 70 and 100 points per class was considered, and a confusion matrix built. These results apply to the whole image. The results obtained for the user's and producer's accuracy can be seen in Table 1. The overall accuracy of the classification is 69%.

Table 1. User's and producer's accuracy of the classification.

Classes	User's accuracy (%)	Producer's accuracy (%)
Deep water	99	91
Shallow water	92	99
Non-vegetated areas	80	75
Eucalyptus Trees	31	62
Shadows	81	90
Herbaceous Vegetation	93	65
Cork Trees	50	59
Coniferous Trees	42	44
Sparse Herbaceous Vegetation	49	33

The information provided by the segmentation of the uncertainty may now be used in two different ways: 1) to evaluate the map accuracy in all the regions, enabling the evaluation of the spatial distribution of accuracy; 2) evaluate the accuracy of regions identified as problematic or which are important for a particular application. Both approaches require the use of a sample of reference points located inside each of the considered regions. The first approach will require the use of a considerable number of points, especially if the regions are relatively small, and may therefore involve a large amount of work. For the second approach only selected regions are analysed and require therefore less work.

For this paper the accuracy was only evaluated for a few regions, namely regions identified in Figure 3a) with number 2, which form only one region, and region 3 in Figure 3b), which also corresponds to only one region. Analysing the composite image and the classification results it can also be seen that, for example, the zone indicated with number 1 in Figure 3a) is clearly a lake in the image, and all pixels in the zone were classified as Shallow Water. Therefore, it is not necessary to create a confusion matrix to conclude that the classification is correct, even though a considerable amount of uncertainty is present in that region. To determine the classification accuracy within regions 2 and 3 a random sample of 50 points was used in each, a reference data base was created and confusion matrixes built. Since only some classes were found within these regions, incomplete confusion matrixes result from the process. For region 2 the user's and producer's accuracy are shown in Table 2 and a zonal overall accuracy of 32% was obtained. For region 3 the user's and producer's accuracy are shown in Table 3 and the zonal overall accuracy is 59%.

Table 2. User's and producer's accuracy obtained for region 2 shown in Figure 3a).

Classes	User's accuracy (%)	Producer's accuracy (%)
Deep Water	100	100
Non-vegetated Areas	28	100
Herbaceous Vegetation	24	100
Sparse Herbaceous Vegetation	100	6

Table 3. User's and producer's accuracy obtained for region 3 shown in Figure 3b).

Classes	User's accuracy (%)	Producer's accuracy (%)
Non-vegetated areas	45	86
Herbaceous Vegetation	68	69
Sparse Herbaceous Vegetation	50	17

4 Discussion and conclusions

The proposed methodology enables the evaluation of the spatial distribution of the classification accuracy, using the information provided by the classification uncertainty. The classification uncertainty may be used as an indicator of classifiers difficulty in assigning one class to each pixel, and can therefore provide valuable information. As seen in the case study presented, regions with high levels of uncertainty may present also high levels of accuracy, as for the lake indicated in Figure 3a) with number 1. That is, even though the classifier had some difficulty in assigning only one class to the pixels, the correct class was identified. However, regions with high levels of uncertainty, such as regions 2 and 3, indicated respectively in Figures 3 a) and b), present levels of accuracy much lower than the overall accuracy of the whole classification. A closer analysis to the results obtained for the user's and producer's accuracy of the several classes shows that within these regions the class Sparse Herbaceous Vegetation has very low levels of producer's accuracy, which means that in these regions a large percentage of reference points assigned to this class were in the reference database assigned to another class, in this case Non-Vegetated Areas and Herbaceous Vegetation.

The obtained results show that the proposed methodology may provide useful information to the user, giving more accurate information about the classification accuracy and problems in different regions of the image.

In this paper only some of the polygons obtained with the segmentation were analyzed, but zonal confusion matrixes could have been built for all polygons, enabling the identification of the spatial variation of accuracy.

References

Burrough P., McDonnell, R. (1998), "*Principles of geographical information systems*". Oxford University Press.

Foody, G.M. (2005), "Local characterization of thematic classification accuracy through spatially constrained confusion matrices". *International Journal of Remote Sensing*, Vol. 26(6): 1217-1228.

Gonçalves, L.M.S., Fonte, C.C., Júlio, E., Caetano, M. (2009), "On the information provided by uncertainty measures in the classification of Remote Sensing images", In: *Proceedings of the "International Fuzzy Systems Association World Congress 2009 " (IFSA 2009) / "European Society for Fuzzy Logic and Technology Conference" (EUSFLAT 2009)*, Lisboa, Portugal, pp.1551-1556.

Gonçalves, L.M.S., Fonte, C.C., Júlio, E., Caetano, M. (2010a), "Evaluation of Remote Sensing Image Classifiers with Uncertainty Measures", *In*: Devillers, R., Goodchild, H. (eds.), *Spatial Data Quality From Process to Decisions*: *Proceedings of the 6th International Symposium on Spatial Data Quality,* St: John's, Canada, pp.163-177.

Gonçalves, L.M.S., Fonte, C.C., Júlio E., Caetano, M. (2010b), "Evaluation of soft possibilistic classifications with non-specificity uncertainty measures". *International Journal of Remote Sensing*, Vol. 31(19): 5199–5219.

Kuncheva, L. I. (2000), "*Fuzzy Classifier Design*". Physica-Verlag, Springer-Verlag.

Woodcock, C.E., Macomber, S.A., Pax-Lenney, M., Cohen, W.B. (2001), "Monitoring large areas for forest change using Landsat: Generalization across space, time and Landsat sensors". *Remote Sensing of the Environment*, Vol. 78: 194–203.

New Horizons for Spatial Data Quality Research

Suzie Larrivée, Yvan Bédard, Marc Gervais & Tania Roy

Centre for Research in Geomatics, Dept of Geomatics Sciences, Laval University, Quebec City, Canada
Suzie.Larrivee@scg.ulaval.ca, Yvan.Bedard@scg.ulaval.ca,
Marc.Gervais@scg.ulaval.ca, Tania.Roy.2@ulaval.ca

Abstract

Producing and using spatial data as well as web services has reached the level of mass market, leading to new research challenges. A number of concepts are appearing regularly in our community: certification, accreditation, inspection, audit, warranty, quality assessment, quality control, quality assurance which are natural ways for a society to organize itself when a mass of citizens is facing increasing risks of misusing given products or services. Such concepts help defining who is responsible or liable for these risks, who absorbs the remaining uncertainty inherent to the use of data once technical means have contributed to reduce this risk. This paper provides an overview of recent issues surrounding formal data quality endorsements in a scenario where spatial data have become a commodity.

Keywords: Spatial Data quality, Audit, Certification, Guarantee, Accreditation

1 Introduction

Research about spatial data quality has taken place for over 30 years (Devillers *et al.*, 2010). However, producing and using spatial data as well as using and offering spatial services over the web has now reached the level of mass market, leading to new research challenges.

Back in the days of paper maps, data producers had a good control over the final output and usages. The integration of data from different maps required technical skills and specialized equipment. Metadata and usage warnings were part of the map legend and easier to understand. With the arrival of digital data in the 1970s, it became easier to exchange and overlay data from different sources. The need to embed quality information within the datasets appeared as soon as the early 1980. For example, Chrisman (1983) clearly stated that "new data structures will have to evolve to encode the quality component, particularly for long-term, routinely maintained projects" and that such evolution was necessary "to assess the fitness of the spatial data to a given purpose". However, the industry focus of the 1980s was to develop more efficient systems to produce and manage spatial data.

Then, the demand to exchange and reuse these expensive data led to international standards in the mid-1990s (cf. ISO/TC-211, OGC). The arrival of the Internet and the need for metadata played key roles into this evolution. As technology and data exchange improved, data quality became a more important issue for industry. The concept of "fitness for use" identified by Chrisman a decade earlier became a more common topic as the

industry acknowledged the problem described by Goodchild (1995): "GIS is its own worst enemy: by inviting people to find new uses for data, it also invites them to be irresponsible in their use".

It is in the early 2000s, along the development of National Spatial Data Infrastructures (NSDI), that the ISO/TC-211 standards about *Quality principles* and about *Quality evaluation procedures* were proposed. This epoch also witnessed the arrival of GPS devices for the masses and of Google Map. Nowadays, we find web-based collaborative systems where map users also produce spatial data. We also see easy-to-deploy spatial data mashups mixing data from several sources thanks to powerful web services, free open-source software and mature interoperability standards. Spatial data and services have become mass-market !

As for every new mass market products and services, our society is adapting itself to protect the public against the increased risks of data misuses. Such misuses are increasingly taking place on a regular basis and appearing in popular literature (ex. daily newspapers), court decisions and specialized literature (scientific journals, conferences, workshops proceedings, web sites and blogs). We are witnessing a growing number of workshops about Law and Spatial Data. Meanwhile, the ISO standards about spatial data quality are evolving (cf. ISO-19157) and the scope of interests with regards to spatial data quality is enlarging. Accordingly, a number of concepts are appearing in our community, some of them more regularly than in the past: certification, accreditation, inspection, audit, warranty, quality assessment, quality control, quality assurance, etc.

These concepts are natural ways for a society to organize itself when a mass of citizens is facing increasing risks of misusing given products or services. Such concepts help to define who is responsible or liable for these risks, who absorbs the remaining uncertainty inherent to the use of data once technical means have contributed to reduce this risk. As stated by Bedard (1988), "most of the ways to reduce uncertainty are technical, while most of the ways to absorb the remaining uncertainty are institutional". Such means characterize a market that is maturing. The goal of this paper is to propose a coherent synthesis of these institutional concepts recently surrounding spatial data quality and to indicate their impact on the data quality research agenda.

2 Professional Services and Products to Assure Quality Processes in the Production of Spatial Data

The first quality concern typically introduced by data producers in their daily operations is a proactive process-driven step aiming at preventing defects. Such an approach is called **Quality Assurance** (QA) and is defined by ISO-9000 as "all the planned and systematic activities implemented within the quality system, and demonstrated as needed, to provide adequate confidence that an entity will fulfill requirements for quality". Several spatial data producers are certified ISO-9001; they normally have an obligation of mean but no obligation of result. Consequently, QA is not sufficient since it cannot *guarantee* the production of quality data. However, in our emerging era of spatial data consumerism, such a *guarantee* is likely to become mandatory for spatial products and services. According to Morris (1981), to **guarantee** means "to assume responsibility for the debt, default, or miscarriage of".

To provide adequate confidence into new products and services, it is common usage to perform an ***Audit*** which is defined as "systematic, independent and documented process

for obtaining audit evidence (records, statements of fact or other information) and evaluating it objectively to determine the extent to which the audit criteria (set of policies, procedures or requirements) are fulfilled" (ISO-19011, 2003). Thus, an *audit* in context of QA, evaluates if the processes and production methods (PPM) are suitable and effective to comply with users' requirements, with standards or with product or service *specifications*. **Data product specifications** are typically defined by the producer and presented as a "detailed description of a dataset or dataset series together with additional information that will enable it to be created, supplied to and used by another party" (ISO-19131, 2007). Spatial data specifications have existed for several decades, especially in photogrammetry, geodesy and topographic mapping and have had a special emphasis on precision and completeness.

A more advanced step towards quality data is the **Certification** process. The most popular *certification* in the QA context is ISO-9001. It refers to "the issuing of written assurance (the certificate) by an independent external body that it has audited a management system and verified that it conforms to the requirements specified in the standard" (ISO Management standards). An organisation having an ISO-9001 certificate has demonstrated "its ability to consistently provide product that meets customer and applicable statutory and regulatory requirements" (ISO-9001). To deliver such a certificate, the *certification* process typically involves a quality *audit*. An *audit* made in a *certification* context can also result in an authorization of using a *certification* mark on the product (ex.: biologic certified). *Certification* is frequently used when meeting a standard is mandatory or when nonconforming products cause high risks of loss and damage.

While *certification* evaluates compliance to requirements or standards, **accreditation** usually evaluates the competency to certify. In the ISO-9001 context: "*accreditation* refers to the formal recognition (a *certification*) by a specialized body (an *accreditation* body) that a *certification* body is competent to carry out ISO *certification* in specified business sectors". However, *accreditation* is also the mechanism used by a customer to evaluate a supplier's competency, processes and production methods, thus requiring a quality *audit*. ISO-19158 defines *accreditation* in such a context, i.e. as a "procedure by which a customer assures that its suppliers are capable of consistently delivering the product to the required quality". This type of customer *accreditation* has been used by government agencies contracting private companies for the production of highly technical documents such as topographic maps and cadastral maps.

When it relates to individuals competency with spatial data, such QA methods can lead to *certifications*, *accreditations* or licenses. For example, the surveyor or the engineer ***license*** may be required to have the right to produce certain categories of spatial data and such a license typically requires a bachelor degree from an *accredited* university program. Another example to indicate a level of competency with spatial data is the *certification* delivered on a voluntary basis by professional associations like the Canadian Institute of Geomatics, the ASPRS and the URISA. Such *certification* processes can follow ISO 19122 *Qualification and certification of personnel*. Finally, short-term diplomas entitled "certificate" can be delivered by education institutions after completion of a small number of courses or by private companies after completion of equipment training (ex. ESRI, Microsoft, CISCO, Trimble *certifications*).

3 Quality Control Professional Services and Products to Ensure Spatial Data Quality

While QA is a proactive process-driven approach aiming at preventing defects and focusing on the obligation of means, Quality Control (QC) is a reactive product-based approach aiming at finding defects and focusing on the obligation of results. QC directly addresses the quality of spatial datasets, i.e. of the final products. **Quality Control** (QC) is defined by ISO-9000 as "the operational techniques and activities that are used to fulfill requirements for quality". To verify each spatial dataset, QC activities such as tests and inspections are performed by skilled people accordingly to quality requirements and specifications. When defects are found, they are reported to managers who can adjust the process chain and QA. Normally, QC answers an obligation of results and can be used as support to produce a **commercial guarantee.** Such a *guarantee* is any promise given voluntarily by the producer in writing to the customer, "to reimburse the price paid or to replace, repair or handle goods in any way if they do not meet the specifications set out in the guarantee statement or in the relevant advertising" (Consumer Affairs Act, art.72). A *guarantee* assumes responsibility for a default but does not *guarantee* the absence of default. To ensure safety and no risk, there exist many mandatory standards for which products must conform. In this circumstance, *certification* is practically a necessity. With voluntary standard, *certification* also certifies conformance to standard and allow products to use *certification* mark, such as Certified OGC-compliant logo.

4 Levels of Trustability into Professional Activities and Documents with regards to QA and QC

The value of previous QA and QC activities depends on who examines the quality. According to Parker (2005) and Raynolds *et al.* (2006), there is 3rd, 2nd and 1st party assessments. Each of them has a different level of trust.

A ***3rd-party QC or QA assessment*** is an evaluation made by an independent and neutral outside body. Third-party examination typically entail: (1) an outside *audit* of an organization's documentation of requirements compliance if we are in a context of QA or (2) a product quality control inspection in a context of QC. Third-party bodies are typically *accredited* to be able to assess, inspect or certify. For example in the context of QA, ISO-9001 *certification* is a 3rd-party *certification* done by an ISO-9001 accredited auditor. In a QC context, many products are 3rd party certified by laboratories or accredited bodies. Third party assessment is generally viewed as the most comprehensive and accurate method to ensure quality.

A ***2nd-party QA or QC assessment*** is performed by a user or customer to evaluate for themselves the fitness of products or processes. In a QA context, 2nd-party examinations entail *audit* by them and are often utilized to ensure quality of an organization's supply chain (ex. ISO-19158 *accreditation)*. In a QC context, one can also assess if a product meets requirements different than those originally intended; it is the case when GIS professionals have contracts from their customers to validate if certain commercial spatial datasets meet his requirements. The latter requirements are usually expressed using spatial data quality elements and techniques described in standards such as ISO-19113, ISO-19114, National Standard for Spatial Data Accuracy (NSSDA) and ASPRS.

*A **1st-party QA or QC assessment*** is a producer's self-examination of compliance of processes or products with self-defined criteria. Self-inspection and self-*audit* are regularly made by producers to ensure that their products and processes reach expecting quality. Such *self-certifications* can be used by the producer to deliver *guarantee certificates* that spatial data meet quality measures written in the specifications. To increase the trustability of self-*certification*, organizations can obtain an ISO-9001 *certification*, meaning that the organization gives oneself the means to produce quality products.

5 Ethics and Legal aspects of Professional Activities

The current state of the law raises many legal aspects related to data quality delivered by spatial data producers, including liability, *legal guarantee*, citizens' privacy and copyright to name a few. Insofar, many spatial data producers use liability exclusions and *guarantee* exclusions about their data such as: (1) no warranty of any kind to the use or appropriateness of the use, (2) no *guarantee* of completeness or currentness, (3) no obligation to correct defects, errors and to update. When spatial data are distributed in a controlled environment where a contract exists between the producer and the customer (typically two organizations), the liability and *guarantee* clauses are clearly defined. However, when spatial data are distributed to the general public, there may be a voluntary *Commercial guarantee* offered by the producer, but there is always a **legal guarantee** which is the obligation to deliver goods in conformity with the descriptions and specifications in the contract of sale to consumers. According to art.73 of Consumer Affair Act and if we assume that spatial data are goods like other goods, then, they are in conformity with the contract if : "(a) they comply with the description given by the trader and possess the qualities of the goods which the trader has presented to the consumer as a sample or model; (b) they are fit for any particular purpose for which the consumer requires them and which he made known to the trader at the time of the conclusion of the contract and which the trader has accepted; (c) they are fit for the purposes for which goods of the same type are normally used or (d) they show the quality and performance which are normal in goods of the same type and which the consumer can reasonably expect". Such obligation may extend to a given period.

Besides laws to protect consumers, there also exist obligations for professionals which are dictated by contracts, codes of ethics and laws. According to Gervais (2003), contractual obligations of GIS Professionals include: (1) to consider user requirements and wishes, (2) effectiveness, (3) to verify and control spatial and descriptive dimensions, (4) to ensure technology compatibility, (5) to ensure evolution capabilities, and (6) to ensure database monitoring. Injuries sustained due to breach of contract, negligence or misfeasance are awarded by monetary compensation. Most codes of ethics have similar guidelines based on concept of morality. In the GISCI code of ethics, we can find guidelines such as: "encouragement to make data and findings widely available, to document data and products, to be actively involved in data retention and security, to show respect for copyright and other intellectual property rights, and to display concern for the sensitive data about individuals discovered through spatial or database manipulations" (Craig *et al.*, 2003). If a professional violates a code of ethics, his licence or certificate can be revoked.

6 Conclusion

As spatial data are entering the consumer world, a new era of rights and obligations has begun. Nowadays, many users *de facto* perceive spatial data as reliable for their usages. Their perception of quality is different than that of experts. The increasing number of incidents and accidents involving spatial data is driving Society to protect these users against the risks of data misuses. Accordingly, we see a growing number of workshops, conferences, blogs and publications involving Law and Spatial Data. Such trend suggests Society is maturing regarding digital spatial products. Nevertheless, if someone complains about damages and wants to know who is liable for the quality of the data involved, do we know immediately what to look for? If the answer is "no", then research is still needed to offer Society the services required and to stand in court as experts in front of judges. To provide professional answers, we must expand our R&D horizons towards the concepts of QA and QC. Accordingly, we started by investigating *which type of services* (audit, inspection, certification, etc.) can be performed *by who* (1st, 2nd and 3rd party) *regarding what* (system, product, individual). Analysing the pertinence of consumer affair acts for spatial data and a cross national comparison should be next.

Acknowledgments:

GEOIDE Project #PIV-23 and Canada NSERC.

References

Bédard, Y. (1988), "Uncertainties in Land Information Systems Databases". *In: Proceedings of AutoCarto 8*, Baltimore, USA, pp. 175-184.

Chrisman, N. (1983), "The role of quality in the long-term functioning of a Geographic Information System". *In: Proceedings of AutoCarto 6*, Hull, Canada. pp.303-321

Consumer affairs Act (1996), chapter 378, Laws of Malta.

Devillers, R., Stein, A., Bédard, Y. *et al.* (2010), "30 years of research on Spatial Data Quality-Achievements, failures and opportunities", *Transactions in GIS*, Vol. 14(4):387-400.

Gervais, M. (2003), *Élaboration d'une stratégie de gestion du risque juridique découlant de la fourniture de données géographiques numériques*. PhD thesis, Laval Univ., Canada

Craig, W. J., Fetzer, J. H., Onsrud, H., Somers, R. and Olson J. M. (2003), *A GIS Code of Ethics*. Approved by URISA Board of Directors.

Goodchild, M.F.(1995), "Sharing Imperfect Data". *In:* Onsrud, H. J. and Rushton, G. (eds.). *Sharing Geographic Information*, New Brunswick NJ, USA, pp.413-425.

ISO-19011 (2003), *Guidelines for quality and environmental management systems auditing*.

ISO-19131 (2007), *Geographic information – Data product specifications*, 48p.

ISO/DTS-19158 (2010), *Geographic information – Quality assurance of data supply*, 33p.

ISO-9001 (2008), *Quality management systems –Requirements,* 50p.

Morris, W. (1981), *The American Heritage Dictionary of the English Language.* Publisher Houghton Mifflin Company, Boston, USA, 1550p.

Parker, B. (2005), *Introduction to globalization & business*. Sage, London, U.K., 536p.

Raynolds, L.T., Murray, D. L., Wilkinson, J. (2007), *Fair trade: the challenges of transforming globalization,* Routledge, Taylor & Francis Group: London, UK., 240p.

Predicting spatial uncertainties in stereo photogrammetry: achievements and intrinsic limitations

André Jalobeanu

Centro de Geofísica de Évora - Universidade de Évora
Rua Romão Ramalho, 59, 7002-554 Évora, Portugal
jalobeanu@uevora.pt

Abstract

We present a new probabilistic method for digital surface model generation from optical stereo pairs, with an expected ability to propagate errors from the data to the final result, providing spatial uncertainty estimates to be used for quantitative analysis in planetary or Earth sciences. Existing stereo-derived surfaces lack rigorous, quantitative error estimates, and we propose to address this issue by deriving a method of error prediction, rather than error assessment as usually done in the area through the use of reference data. We use only the information present in the available data and perform the prediction using Bayesian inference. We start by defining a forward model, using an adaptive radiometric change map to achieve robustness to noise and reflectance effects. A priori smoothness constraints are introduced to stabilize the solution. Solving the inverse problem to recover a surface from noisy data involves fast deterministic optimization techniques. Though the reconstruction results look satisfactory, we conclude that uncertainty estimates computed from two images only are unreliable, which is due to major limitations of stereo, such as non-Lambertian reflectance and incorrect spatial sampling, which violate our underlying assumptions and cause biases that cannot be accounted for in the predicted error budget.

Keywords: Bayesian inference, probabilistic modeling, digital photogrammetry, stereo, DSM generation, image processing.

1 Introduction

Stereo optical images are still widely used to generate digital surface models (DSM). Commercial cameras exhibit increasingly higher resolution, signal-to-noise ratio and dynamic range, so that the uncertainty in topographic measurements is expected to shrink accordingly. In this study, we show that not only this is not happening, but there are also fundamental limitations that make error prediction unreliable in practice, despite a rigorous probabilistic treatment of the problem.

We wish to predict the DSM accuracy, rather than assess it using reference data sets as done usually. Moreover, we want to capture the spatial variability of this error and its spatial correlation which has an impact in most applications as noticed by (Wechsler and Kroll, 2006). Many 3D reconstruction methods have been developed in the computer

vision community (Brown *et al.,* 2003), however they do not provide quantitative error estimates. Some attempts have been made to predict the accuracy (Davis *et al.*, 2001) based on local terrain characteristics and a qualitative matching quality measure, however these approaches do not directly take into account the data and the spatial variability of stereo cues. Bayesian approaches (Bernardo and Smith, 1994) to stereo disparity estimation have been developed (Cheng and Caelli, 2007) but do not explicitly produce error maps, and the data terms are not really adapted to photogrammetry as the method was mainly designed to handle computer vision problems with indoors imagery. We propose to use the image information content to compute the uncertainty, as the presence of edges, texture and noise can have a dramatic impact on it. A rigorous probabilistic modeling of data formation followed by Bayesian inference enables us to build a probability density function (pdf) of the disparity map, that helps provide a DSM with an associated error map via a geometric transform (known or estimated via calibration).

2 Bayesian stereo disparity inference

We extend here the approach first presented in (Jalobeanu *et al.*, 2010). Within a probabilistic framework, all the parameters are random variables, which helps to account for the randomness of phenomena affecting observations (noise, radiometry) and the underlying variability of the object of interest (DSM or disparity map). Bayesian inference makes use of available knowledge expressed by specifying a priori pdfs, and combines it with a data formation model to derive the a posteriori pdf of the object given the data. The idea is to use Bayesian networks (Jordan, 1998) to model all variables and causal relations between them, in order to form a joint pdf by simply multiplying all the prior and conditional pdfs. Fig. 1 shows the proposed network, where nodes represent variables, converging arrows conditional pdfs and terminal nodes prior pdfs. Observed variables or data are in gray, fixed variables (whose estimation is beyond the scope of this paper) are in blue. One typically has to integrate out unwanted variables, thus performing a marginalization, ending up with a marginal pdf proportional to the sought posterior.

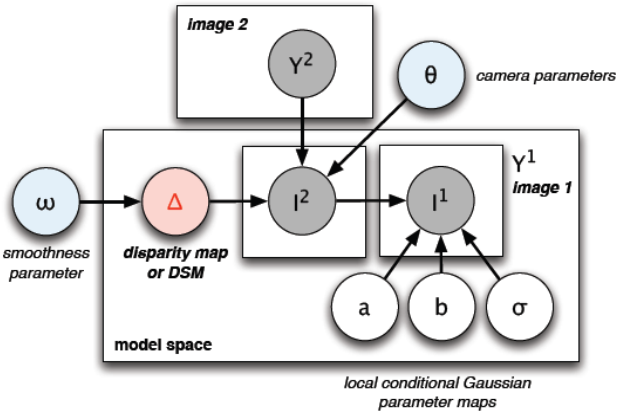


Figure 1. Directed graphical model or Bayesian network that helps to build the joint pdf.

We define a local area matching (Brown *et al.*, 2003) method as follows, considering a fixed patch I^1 on image Y^1 and a moving patch I^2 extracted from Y^2 by assuming a uniform shift (determined by the local disparity), the resampling being done via Spline interpolation (Thévenaz *et al.*, 2000) in order to minimize sampling artifacts or aliasing (Jain, 1989). To account for additive and multiplicative radiometric changes between windows I^1 and I^2, we assume a linear transformation of the pixel values with local parameters a

and b. We also assume the noise to be Gaussian of mean 0 and variance σ^2. These parameters are assumed constant over the window support. The conditional Gaussian pdf writes:

$$P(I^1 \,|\, I^2,\, a, b, \sigma) = \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(I_i^1 - aI_i^2 - b\right)^2 / 2\sigma^2} \tag{1}$$

Integrating with respect to the three local change parameters gives the following expression of the likelihood, where c is the normalized correlation coefficient:

$$P(I^1 \,|\, I^2) = \iiint P(I^1 \,|\, I^2,\, a, b, \sigma)\, da\, db\, d\sigma \propto \left(1 - c^2\right)^{-\alpha} F_\alpha(c) \tag{2}$$

where $\alpha = (n-3)/2$, n being the number of pixels of the patch. The attenuation function F_α arises from the constraint $\alpha > 0$ and is involves the error function erf:

$$F_\alpha(x) = 1 + \mathrm{erf}\left(x\sqrt{\frac{\alpha}{1-x^2}}\right) \tag{3}$$

This mapping from correlation to likelihood is the main originality of the proposed method. It is illustrated in Fig. 2 which shows how a cost function (the -log likelihood) based on this approach behaves, for a 5x5 window, penalizing negative and small values of the correlation.

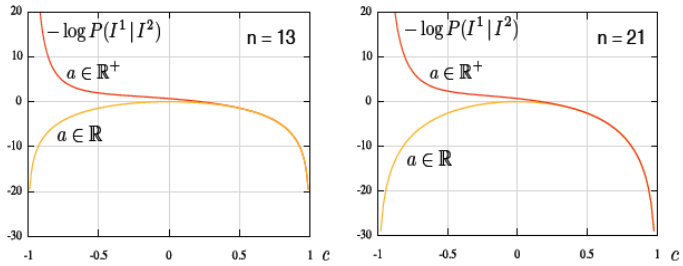


Figure 2. - log likelihood (2) as a function of the correlation coefficient for two different values of n, with and without the constraint $\alpha > 0$.

Now we form the joint likelihood as the product of local likelihoods, assuming independence, where the k–th patches explicitly depend on the data and the local disparity parameter Δ_k:

$$P(Y^1 \,|\, Y^2, \Delta, \theta) \simeq \prod_k P\left(I^{1k}(Y^1) \,|\, I^{2k}(Y^2,\, k,\, \Delta_k,\, \theta)\right) \tag{4}$$

The sought posterior is proportional to the joint likelihood times the prior:

$$\mathrm{P}(\Delta \,|\, Y^1, Y^2, \theta, \omega) \propto P(Y^1 \,|\, Y^2, \Delta, \theta)\, P(\Delta \,|\, \omega) \tag{5}$$

The prior model for the disparity (or DSM) is defined by a first order Markov Random Field (Li, 1995) which helps constrain the smoothness of the solution, with a global regularization parameter ω (assumed fixed in this work):

$$P(\Delta \,|\, \omega) = \frac{1}{Z_\omega} e^{-\omega\, \Phi(\Delta)} \quad \text{with} \quad \Phi(\Delta) = \sum_{i,\, k \sim i} (\Delta_i - \Delta_k)^2 \tag{6}$$

where Z_ω is a normalizing constant.

We refer to (Jalobeanu *et al.*, 2010) for a description of the inference algorithm. A fast, deterministic optimization algorithm based on Loopy Belief Propagation (Sun *et al.*, 2003), applied to the posterior (5), is used to generate the DSM. The input is a set of lowpass-filtered, sampled likelihood terms, and the processing is done in a multiscale framework using a pyramid decomposition (Jain, 1989). The uncertainties are essentially derived from the shape of the likelihood functions at the optimum, and in the following we examine the issues related to these functions.

3 Intrinsic limitations: unpredictable biases

3.1 Fundamental assumptions and their consequences

The ability to predict local uncertainties depends on the consistency of the likelihood terms. These terms embed the local information carried by the data as a pdf of the local

disparity without any prior information. A bias is judged significant when the true disparity lies outside a predefined confidence interval. Unfortunately we observed enough significant biases (outside 95% and even 99% confidence intervals) to raise concerns about the reliability of the method. Indeed, a number of spatial locations suffers from error underestimation as the bias can not be predicted. After analysis it appears that the limitations are mainly due to the intrinsic quality of the data and not to algorithmic simplifications. For this analysis we went back to the basic assumptions and investigated their effects (see Table 1 for a summary).

Figure 3 illustrates the two major types of artifacts that affects stereo photogrammetry that we want to point out. The likelihood pdfs $P(I^1|I^2)$ computed using (2) are shown and the bias is clearly visible in both cases, as the bulk of the pdf is far from the true elevation, checked during field work. Artifacts are due to undersampling (Jain, 1989) or image aliasing are strongest on high spatial frequency objects such as truck tracks in the sand, and the pdfs are so narrow that the final optimization stage using the smoothness prior (6) does not help to remove them. Aliasing is an intrinsic property of data and reflects a deliberate choice in the optical design (people like sharp images!). It can be reduced if a frequency space lowpass filter is employed (Jain, 1989) but with a loss in spatial resolution (factor 2 or higher). A drawback of such a filtering is the amplification of the radiometric errors described below, which was confirmed on simulations.

Radiometric changes were accounted for in a simple manner with a reasonable number of parameters (3, for about 20 data points) as defined in (1). Obviously this is insufficient in some cases, as on the edge of the tennis court (the reflectance depends on the viewing angle and changes within the patch area, despite the small size of the patch). Such cases are not uncommon in nature, as reflectance properties are spatially variable and rarely Lambertian. It is difficult to address this issue without significantly increasing the number of parameters of the radiometric model and ending up overfitting the data. The uniform parameter assumption is unavoidable and yields yet another intrinsic limitation of stereo.

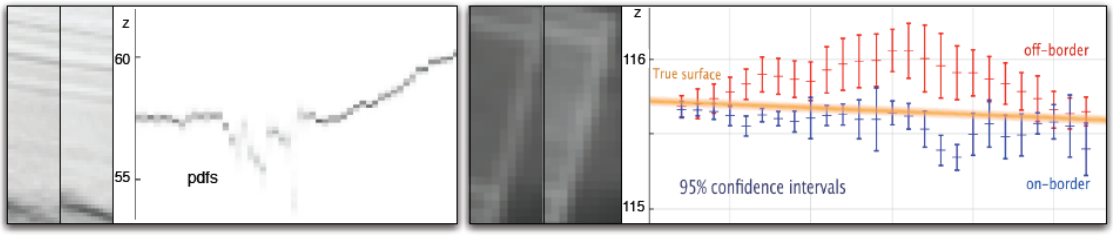


Figure 3. Problematic likelihood pdfs computed along a segment. Left: aliasing artifacts, right: radiometric artifacts, on two segments, on the border, and 3 pixels to the right.

Compression noise is neither white nor Gaussian; its spatial structure can induce outliers, and is commonly found in satellite or planetary data due to communication bandwidth constraints, however aerial cameras do not have such concerns. On the other hand, the uniform motion assumption is independent of data quality, and causes well-known fattening effects when the patch covers multiple objects or complex terrain. Such biases are related to slope, curvature and roughness, but not in a simple way that would help cancel or identify them. Finally the independence assumed between area-based data terms in (4) is questionable, as some overlap might be tolerated in order to achieve a satisfactory spatial resolution of the final DSM. With 5x5 patches one typically estimates a disparity every 2x2 pixels. While the dependence is obvious in this case, it is not clear how to evaluate it, or if joint likelihoods of two neighboring areas could be integrated at all in this methodology. In any case the related biases are negligible compared to radiometric or aliasing biases, as simulations have shown.

Table 1. Main assumptions made, actual real data properties and related unpredictable bias.

Assumption	Real data properties	Observed bias
Uniform radiometric change parameters	Spatially-varying non-Lambertian brdf	Artifacts
Correct sampling	Undersampling (aliasing)	Aliasing artifacts
Uniform motion	Strong topography variations, multiple objects	Fattening effect
White Gaussian noise	Correlated noise (compression)	Isolated outliers
Independent data terms	Overlapping windows	Uncertainty errors

3.2 Bayesian DSM generation from digital aerial images: an illustration

A series of aerial images were acquired in 2009 over the Portuguese coastline between Troia and Sines, at 20 cm ground sampling distance (GSD) and using an Intergraph DMC camera, with 60% overlap to allow for a stereo processing. One of the images is shown in Fig. 4, with the corresponding DSM reconstruction at 40 cm GSD in shaded relief and color. The direct georeferencing was insufficient, so control points had to be used to correct the orientation. The DSM was generated in the image space with the proposed method, and then resampled in UTM coordinates to be used within a cliff erosion monitoring project (Jalobeanu *et al.*, 2010). Computing the topography variation with an error map is easily done by subtracting two multidate DSMs and adding the error variances. This enables one to check the statistical significance of the computed elevation changes, provided that the uncertainties are consistent. The DSM looks satisfactory after a preliminary inspection; a validation procedure using RTK GPS tracks as ground truth is in progress.

The trench on top left of the area is an aliasing artifact, but could easily be interpreted as an actual change if we were to trust the estimated errors (see Fig. 3). We notice that the vegetation appears in the DSM, as expected, but raises an important question about the relevance of surface-related error maps (even when they are consistent with the true surface) when most users are interested in topography.

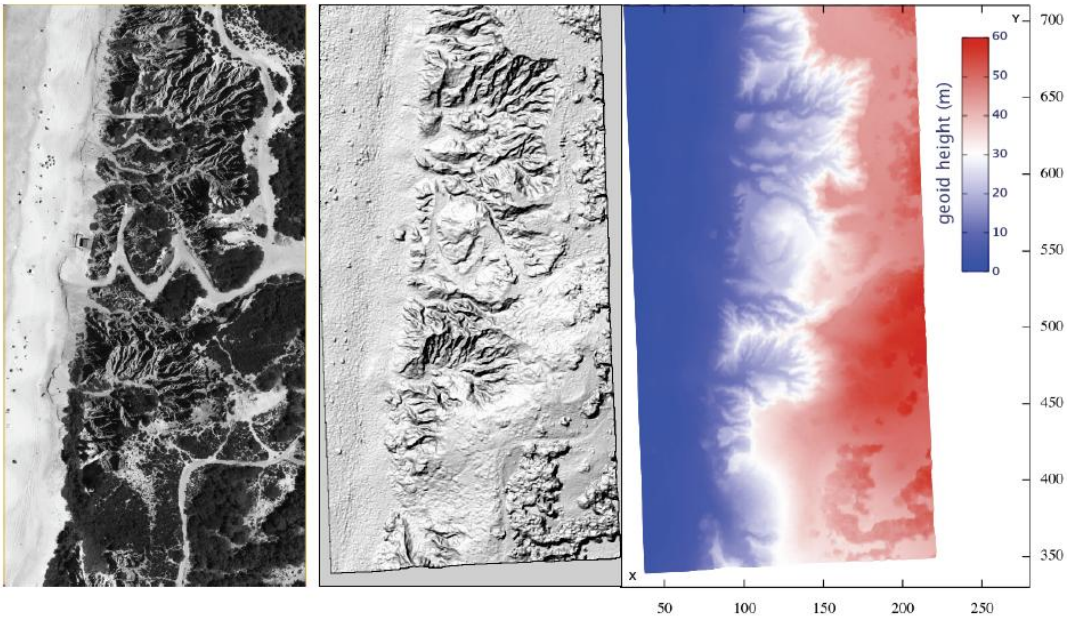


Figure 4. Results: from left to right, image 1, shaded-relief DSM and color-coded DSM(UTM/WGS84 projection, coordinates in meters, arbitrary origin).

3 Conclusion

Fundamental limitations impede the robust estimation of the spatially variable uncertainty related to surface models. They are due to departures from assumptions, some of which are related to intrinsic data quality issues (sampling, noise), while the others reflect the insufficiency of stereo images to constrain all the physical parameters involved in image formation (radiometry and geometry). Filtering the images, multiplying the number of observations and reducing the parallax might help reduce some of the biases but could also increase some others. In any case, if camera manufacturers avoided undersampling and compressing, two important sources of bias would be eliminated and the error prediction would be improved.

Nowadays, to obtain reliable and topographically consistent error bars, one may consider using LiDAR data (when available), for which the error propagation is more straightforward to perform, as range measurements are made directly.

We have shown examples of failures in error estimation, however the frequency of these failures in real data is still unknown; experiments are in progress to assess the robustness of error estimation using dense reference data sets and determine in what cases the proposed uncertainty map can be used, especially when only stereo data are available for practical reasons.

Acknowledgments

This work was partially funded by the French Research Funding Agency (ANR) (SpaceFusion project, "Jeunes Chercheurs 2005" JC05_41500) and by the Portuguese Funding Agency (FCT) (AutoProbaDTM project PTDC/EIA-CCO/102669/2008, FCOMP-01-0124-FEDER-010039).

References

Bernardo, J., Smith, A. (1994), *Bayesian Theory*, John Wiley and Sons

Brown, M, Burschka, D., Hager, G. (2003), "Advances in computational stereo", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 28(8)

Cheng, L., Caelli, T. (2007), "Bayesian stereo matching", *Computer Vision and Image Understanding*, Vol. 106

Davis, C.H., Jiang, H., Wang, X. (2001), "Modeling and Estimation of the Spatial Variation of Elevation Error in High Resolution DEMs From Stereo-Image Processing", *IEEE Trans. on geoscience and remote sensing*, Vol. 39(11)

Jain, A.K. (1989), *Fundamentals of digital image processing*, Prentice Hall

Jalobeanu, A., Gama, C., Gonçalves, J.A. (2010), "Probabilistic surface change detection and measurement from digital aerial stereo images", *IEEE International Geoscience & Remote Sensing Symposium*, Honolulu, Hawaii, USA

Jordan, M.I., editor (1998), *Learning in graphical models*, MIT Press

Li, S.Z. (1995), *Markov Random Field Modeling in Computer Vision*, Springer-Verlag

Sun, J., Zheng, N.N., Shum, H.Y. (2003), "Stereo matching using belief propagation", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 25(7)

Thévenaz, P. *et al*. (2000), "Interpolation revisited", *IEEE Trans. on Medical Imaging*, Vol. 19(7)

Wechsler, S.P, Kroll, C.N. (2006), "Quantifying DEM Uncertainty and its Effect on Topographic Parameters", *Photogrammetric Engineering & Remote Sensing*, Vol. 72(9)

Indicators of spatial autocorrelation for identification of calibration targets for remote sensing

Nicholas A.S. Hamm[1], E.J. Milton[2] & V.O. Odongo[1]

[1] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, PO Box 217, 7514 AE Enschede, The Netherlands. email: hamm@itc.nl
[2] School of Geography, University of Southampton, Southampton SO17 1BJ, United Kingdom. email: ejm@soton.ac.uk

Abstract

Vicarious calibration of remote sensors requires the identification of suitable ground targets (GTs). One criterion for suitable ground targets is that they should be spatially uniform. Traditionally, this assessment has been performed by calculating the coefficient of variation within a small window (e.g., 3×3, 5×5 pixels). More recently the Getis statistic, Gi, calculated within a similar window has been used to help identify appropriate targets. This paper considers the use of Gi* in more detail and uses it alongside the variogram to obtain a more detailed understanding of the spatial structure of the site. This knowledge was used to identify a wider range of possible targets. It is further suggested that this knowledge could assist in the understanding of the spatial-temporal dynamics of the site and will be helpful for planning site visits for in situ data acquisition.*

Keywords: Vicarious calibration, ground targets, LISA, Getis, variogram.

1 Introduction

Calibration and validation (Cal/Val) is emerging as a key component of data quality research in remote sensing. Post-launch calibration of remote sensors is an area an active area of research. Some onboard systems have been devised but an alternative, vicarious calibration, is often used.

Vicarious calibration uses ground targets (GTs) with known reflectance properties. This knowledge is combined with an atmospheric correction model. The remotely sensed measurements can then be validated against the known properties of the target (Thome, 2001).

Various criteria are imposed for the selection of appropriate GTs, as outlined by Thome (2001). These are: (1) high reflectance (> 30%), (2) high elevation (>1000 m), (3) high spatial uniformity, (4) temporal stability, (5) lambertian surfaces, (6) high spectral uniformity and (7) accessibility. High altitude arid areas and dry lake beds are often considered to meet these criteria. Example calibration sites include Railroad Play, Nevada and Tuz Gölü, Turkey. The GT is a subsection of the site.

This paper focuses on criterion (3), spatial uniformity. There remain open questions regarding the (i) criteria for spatial uniformity, (ii) how it should be assessed and (iii) what the implications are for sampling in the field. Typically the uniformity of a site is

determined first from remotely sensed data (Bannari *et al*., 2005). When a suitable site has been found a field visit is undertaken in order to characterize the spectral properties of the GT.

The uniformity of a site is assessed typically by calculation of the coefficient of variation (CV = *s* / *m*, where *s* is the standard deviation and *m* is the mean) within a small fixed window (e.g., 3×3, 5×5, 9×9 pixels). Contiguous areas with a CV < 3% are considered to be spatially uniform (Bannari *et al*., 2005). Recently some authors have questioned the use of the CV as the only measure of uniformity (Bannari *et al*., 2005; Gurol *et al*., 2008; Odongo, 2010). They noted that the CV is not a spatial statistic and, since it is standardized, adjacent windows can have similar CVs, but quite different means. This has led to the same authors proposing that the Getis statistic, Gi*, be used in conjunction with the CV to identify spatially uniform areas. Gi* is a local indicators of autocorrelaion (LISAs). The Gi* has also been calculated within fixed small windows. This has been useful for identifying appropriate areas for vicarious calibration, where appropriate is judged to mean bright areas with low variance.

The objective of this paper is to elaborate further on how both local and global measures of autocorrelation might be used to help identify suitable GTs. In particular, the use and meaning of the Gi* is considered in greater detail than in the above-listed papers. It is also calculated within much larger windows.

2 Study site and data

The site used was Tuz Gölü, which is candidate Cal/Val site, as endorsed by the Committee on Earth Observation Satellites (CEOS). It is approximately 150 km south-east of Ankara, Turkey (latitude: 38.83°, longitude: 33.33°). It is an ephemeral saline lake, approximately 50 km wide and 80 km long. During the summer the lake dries out exposing a salt flat. It is this exposed salt flat that provides the possible calibration target (Gurol *et al*., 2008).

The data used for this research was a Landsat 5 image from 31 August 2009. This was atmospherically corrected using the ATCOR-2 program and the MODTRAN mid-latitude summer atmospheric profile. This yielded the hemispherical directional distribution function (HDRF) value of reflectance. For analysis the NIR band 4 was chosen (see Figure 1). This has a pixel size of 30 m.

3 Methods

Two methods were used to provide information about the global and local autocorrelation. These were the variogram and Gi* statistic respectively.

3.1 The variogram

The variogram is defined as half the expected squared difference for a given geographical separation (Webster and Oliver, 2001). The sample variogram is calculated from the data, *x*, as:

$$\widehat{\boldsymbol{\gamma}}(\boldsymbol{h}) = \frac{\mathbf{1}}{\mathbf{2}\boldsymbol{N}(\boldsymbol{h})}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}(\boldsymbol{h})}\left(\boldsymbol{X}(\boldsymbol{u_i}+\boldsymbol{h}) - \boldsymbol{X}(\boldsymbol{u_i})\right)^{\mathbf{2}} \quad (1)$$

Where *u* indicates location and *h* is the separation or lag. The variogram rises to the sill at the *range*. Two points separated by a distance greater than the range are uncorrelat-

ed. The *nugget* has a duel interpretation as measurement error or spatial autocorrelation at lags shorter than the minimum sample separation. In this research the sample variogram was used as an exploratory tool to help understand the nature of the spatial autocorrelation.

It is important to realize that the value of γ is dependent only on the lag and not the actual location. This is the stationarity assumption. The variogram does not provide information about a value at a particular location, only how it is correlated with another value at a specific lag.

3.2 The Getis statistic

The Getis statistic, Gi*, is a local indicator of autocorrelation (LISA) (Getis and Ord, 1996). It is defined as follows:

$$G_i^*(d) = \frac{\sum w_{ij}(d)x_j - W_i^*\bar{x}}{s\left[W_i^*(n-W_i^*)/(n-1)\right]^{1/2}} \quad \textbf{(2)}$$

where $w_{ij}(d)$ is a symmetric binary weights matrix where $w_{ij} = 1$ for all pixels found within distance d of pixel i and $w_{ij} = 0$ for all pixels found outside d and x is the pixel value. W_i^* is the number of pixels within the distance d (i included). $\bar{x}$ and s are the mean and standard deviation for the entire image. For this research d was defined in terms of a radius, although previous authors used a window (e.g., 3×3 for d=1 pixel, 9×9 for d=4 pixels). The Gi* was investigated for a wide range of d, from d = 30 m through to d = 990 m (67×67).

Positive values of Gi* indicate a cluster of values that are larger than the mean (bright clusters), whilst a negative value indicates relatively dark clusters. The d for which Gi* is a maximum can be interpreted as the maximum extent of the local autocorrelation (Getis & Ord, 1996).

The Gi* values may be further interpreted in terms of the properties of the normal distribution. For example, values greater than 1.96 or less than -1.96 are indicative of significant clustering (0.05). The intermediate values may arise due to a lack of clustering or to clustering around the mean. The Gi* is unable to distinguish between these.

4 Results

At the first stage the whole lake was considered (Gurol *et al*., 2008; Odongo, 2010). The summary statistics are shown in Table 1. The histogram of HDRF values and the Gi* statistics, computed within for d = 30 m are shown in Figure 1. The lake was not completely dry and the large area of negative Gi* values is associated with a water body. There was a large area of positive Gi* values towards the north of the lake, but much of this area does not exceed the +1.96 threshold.

This simple analysis raises the question of whether the lake should be further analyzed as a whole. It is known that the water body will not be appropriate as a vicarious calibration target. Furthermore, the histogram of HDRF values is highly skewed and the surface is clearly non-stationary from the perspective of geostatistical analysis. From this preliminary analysis the decision was made to work further on a subset of the lake to the north of the water body.

Table 1. Global statistics for the lake and lake subset as well as for the identified bright and dark GT. HDRF is expressed as a percentage.

Image	Min	Mean	Median	Max	SD	CV (%)
Whole	0.3	53.3	58.0	81.7	13.2	24.7
Subset	31.3	59.2	59.0	73.0	2.9	4.9
Bright GT	32.0	61.6	61.7	70.3	2.0	3.3
Dark GT	34	57.0	57.1	67.3	1.4	2.5

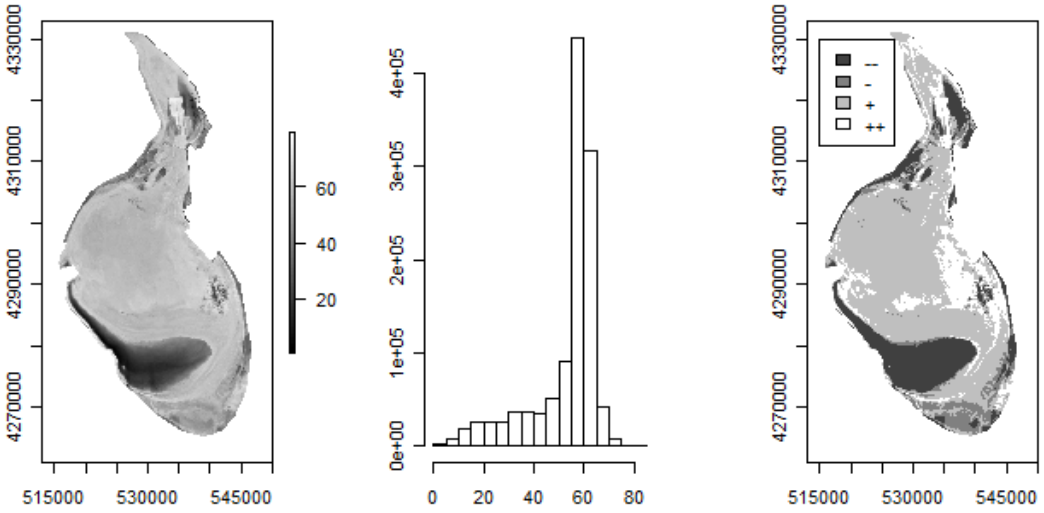


Figure 1. Whole lake NIR. Left: HDRF (%). Middle: Histogram HDRF values. Right: Gi* for *d*=30 m. ++(--) indicates Gi* greater (less) than, +(-) indicates Gi* between 0 and 2(-2).

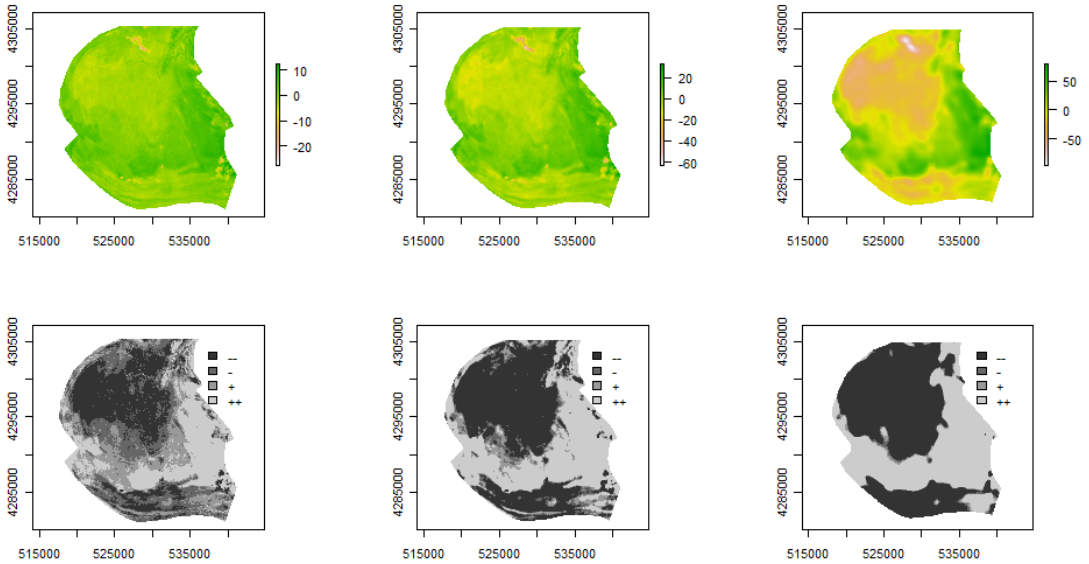


Figure 2. Lake subset. Upper shows Gi* values for *d*=30 m, *d*=90 m and *d*=480 m. Lower shows the bright and dark areas for the same *d*. For key see Figure 1.

The summary statistics for the lake subset are shown in Table 1. From the histogram it was clear that the distribution of HDRF values was symmetric. The HDRF for the subset was high, with a minimum value above the required threshold. The CV for the whole target was low (4.9%). The Gi* computed for *d*= 30 m, *d*= 90 m (9×9) and *d* = 480 m (33×33) are shown in Figure 2. There are three important related points to note. First, the Gi* values tended to increase in magnitude as *d* increased. This was also the case for

d = 990 m (67×67). This suggested that the range of local autocorrelation has not yet been reached. Second, the possibility to identify large contiguous patches of relatively bright or relatively dark areas increased as the window size increased. Finally, both relatively bright and relatively dark areas may be suitable since both will fall above the 30% reflectance threshold.

Analysis of the directional variograms revealed clear directional anisotropy. There was also evidence of a trend on location. The variograms were recalculated after first removing a polynomial trend on location. This removed the anisotropy. The omnidirectional variogram is shown in Figure 3. It is clear that the sill is not reached until after 8000 m. This is large by comparison to the maximum window size for which Gi* was calculated.

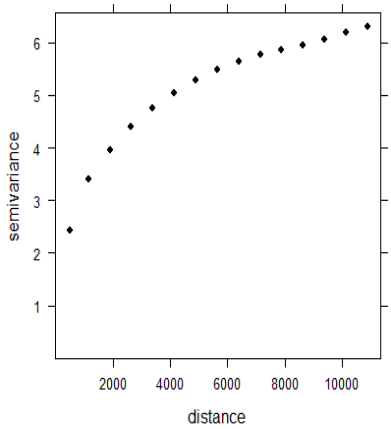


Figure 3. Omnidirectional variogram (polynomial trend on location) for the lake subset.

The largest contiguous relatively-bright and relatively-dark areas were identified using the Gi* values for d = 480 m window size. For the dark area, there was a small patch with much lower Gi* values. This was masked from further analysis. Summary statistics for these two areas are shown in Table 1. Both have an average reflectance above 30% and a CV of 3% or lower, they are both suitable GTs.

5 Discussion

Previous research has analyzed the whole of the candidate site by calculating Gi* and CV within a small window. This has allowed the identification of suitable GTs (Bannari *et al.*, 2005; Odongo, 2010); however, it is proposed that a richer understanding of the spatial structure could be obtained.

The first step was to mask out areas that were clearly unsuitable. The remaining area was above the required 30% HDRF and had a low CV of 4.9%. Variogram analysis revealed that global spatial autocorrelation was present until after 8000 m and there was clear evidence of a trend on location. This is important information, since both would need to be accounted for when undertaking fieldwork to characterize the at-surface reflectance of the target.

The Gi* continued to increase as d increased. Hence significant local autocorrelation may be present where it was not identified in previous studies, owing to the small window sizes used (e.g., 3×3, 5×5). This is to be expected, given that the variogram range was substantially larger than the window size.

Previous studies only considered relatively bright areas. In this study, both relatively bright and relatively dark areas could be considered, since both were above the 30%

HDRF threshold. Following the argumentation of Getis (1994), it was expected that the CV of these areas would be below the global CV and that was confirmed in this study (Table 1). This represents a step forward from previous studies. Studies based only on CV could not identify these patches whereas those that used the Gi* were restricted to identifying bright areas within small windows.

Despite the progress made, unanswered questions remain. First, this analysis needs to be translated into a methodology for identifying suitable GTs for multiple images and multiple bands. Clear rules and automation are required to achieve this. Second, this research used quantile values based on the normal distribution (1.96 at 95%). Getis & Ord (1996) indicated that this may be inappropriate because of the overlap between windows, but did not provide guidance for such large images. This matter requires further investigation. Finally, it is unclear whether true local autocorrelation has been identified or whether these are simply areas of generally low or high values (Ord and Getis, 2001). In one sense this is unimportant, since the overall goal was simply to identify sufficiently bright targets with low variability. Nevertheless, a fuller understanding of this issue would be useful, not least because this study actually chose to mask out a particularly intense hot spot.

6 Conclusions

The analysis conducted in this paper led to greater insight into the spatial structure of a proposed Cal/Val site. This is useful for four reasons. First, it may be possible to identify larger areas than were obtained previously. Second, it may help to understand the spatial dynamic of change over time. Third, it may be helpful for field sampling. Finally, the information may be applied to different sensors with different spatial resolutions.

References

Odongo, V.O.. (2009) Uncertainty in reflectance factors measured in the field: implications for the use of ground targets in remote sensing. MSc thesis, Univeristy of Twente.

Bannari, A., Omari, K., Teillet, P.M., Fedosejevs, G. (2010), "Potential of Getis statistics to characterize the radiometric uniformity and stability of test sites used for the calibration of Earth observation sensors". IEEE Transactions on Geoscience & Remote Sensing, Vol. 43:2918-2926.

Getis, A. (1994), "Spatial dependence and heterogeneity and proximal databases". In: Fortheringham, S., Rogerson, P. (eds.). Spatial Analysis and GIS, Oxford, Taylor and Francis, pp. 105-120.

Getis, A., Ord, J.K., (1996), "Local spatial statistics: an overview". In: Longley, P., Batty, M. (eds.). Spatial Analysis: Modelling in a GIS Environment, Oxford, Taylor and Francis, pp. 261-277.

Gürol, S., Ozen, H., Leloğlu, U.M., Tunali, E.. (2008), "Tuz Gölü: new absolute radiometric calibration test site". In: ISPRS Congress XXXVii, International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Beijing, China, pp. 35-39.

Thome, K.J. (2007), "Absolute radiometric calibration of Landsat 7 ETM+ using the reflectance-based method". Remote Sensing of Environment, Vol.78: 27-38.

Webster, R., Oliver, M. (2001) Geostatistics for Environmental Scientists, Wiley, New York.

UNCERTAINTY MODELING AND PROPAGATION

Area Measurement Error Caused by Rasterization

Qianxiang Xu & Wenzhong Shi

Department of Land Surveying and Geo-informatics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong.
xuqx1983@gmail.com

Abstract

With the development of Remote Sensing technology, raster data is widely accepted in fields, such as geo-science and landscape ecology, for its convenience for storage, analysis, sharing, and displaying. There is a large need to convert other sources of data, such as vector data, to raster format by the process of rasterization. However, rasterization is accompanied by information loss, which results in rasterized stair-shaped maps cannot precisely represent the vector data. Errors always exist, no matter how fine the raster cell is. This makes area measurement value calculated from the rasterized data can be unpredictable. Although former researchers have made some efforts to reveal and model the variation of area measurement errors, this old but important issue has not been resolved yet. This paper makes a comprehensive review of the existing contributions. We mainly focus on issues such as the factors influencing the errors. In the conclusion part, the authors also discuss some urgent relevant issues which need to be resolved in the future.

Keywords: Area measurement error, Rasterization, Rasterizing method.

1 Introduction

Rasterization is a process accompanied by information loss as a result of inconsistency of vector edges with raster cell boundaries (Piwowar *et al*., 1990; Liao and Bai, 2009). Discrepancies always exist along the rasterized vector edges, which lead to a difference of measured area between rasterized data and original vector data. However, this kind of error has not been seriously considered owing to the fact that rasterized data do look like the same as the vector data (Dunn *et al*., 1990). Even though it risks causing the users to make a totally wrong decision, nearly all users directly, or, only by visual assessments, utilize the area measurement from rasterized data, without concerning the magnitude of errors. In fact, there may be large differences in number and location among the rasterized data, which are not acceptable, especially when huge amounts of data are exchanged between people and systems (Van Der Knaap, 1992).

As a risk for misuse of area measurement from rasterized data and lack of concern for area measurement errors by users and researchers, motivated by a desire to make an alert for a careful use of this kind of area measurement, this paper reviews the existing contributions of former researchers and gives an overview on area measurement errors caused by rasterization. The remainder of this paper is structured as follows. In section 2, area measurement errors are formulated and discussed from two different standpoints, absolute error versus relative error and numerical error versus spatial error respectively. Sec-

tion 3 analyzes the factors which may influence the magnitude and distribution of area measurement errors. Finally, section 4 draws conclusions and proposes some potential issues which need to be resolved in the future.

2 Area measurement error

Representing random direction and length vector polygon boundaries with fix-sized square cells will produce stair-shaped raster data (Figure 1). No matter how small the cell size is, errors always exist, because raster data cannot be perfectly consistent with vector data. Hence, some error indicators should be employed to assess the discrepancies before and after rasterization.

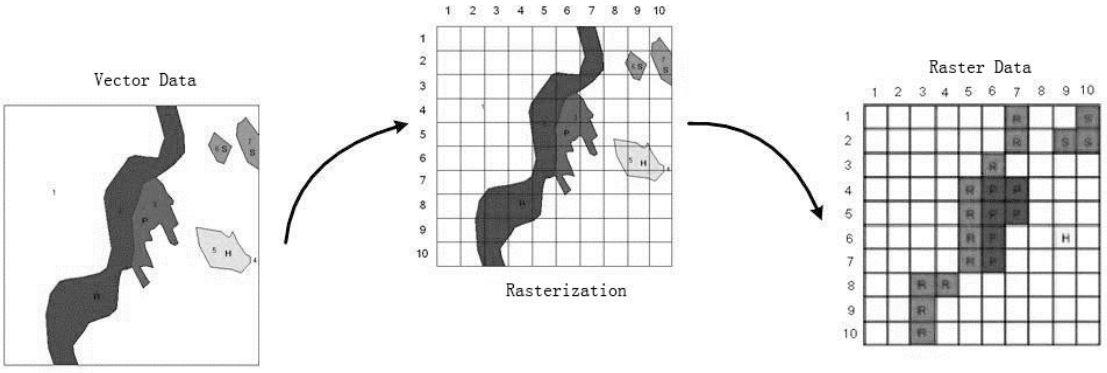


Figure 1. The error caused by rasterization

2.1 Absolute error versus relative error

The discrepancy between raster data and vector data is called Absolute Area Error (AAE), and its units often are m^2, hectare, and km^2. Dividing the AAE by the area before rasterization, we get Relative Area Error (RAE). RAE is normally represented in percentage format (%).

$$AAE = \text{Area after rasterization - Area before rasterization} \quad (1)$$

$$RAE = \frac{\text{Area after rasterization - Area before rasterization}}{\text{Area before rasterization}} \quad (2)$$

AAE is mostly used for operational applications, while RAE is more significant for designers and managers. For example, RAE of agricultural land is of concern to governmental officials for food security in a country, whereas AAE is relevant to a family who aim to determine how many seeds should be bought for planting. In geo-scientific circle, it is generally accepted that RAE is more meaningful and reasonable than AAE (Frolov and Maling, 1969; Liao and Bai, 2009).

2.2 Numerical error versus spatial error

AAE and RAE are easy for calculation; therefore they are largely employed for evaluation of rasterization results. However, the compared area value is only a numerical value, and the positional changes have not been considered. AAE or RAE, which has a value of zero, cannot imply that the position of a polygon remains unchanged. In other words, although there is a great spread of each category and the location of the assigned cells, and the land is poorly classified, the area of each category can still be accurately estimated, which may confuse the user and mislead the manager to make a wrong decision (Van

Der Knaap, 1992; Hollister *et al*., 2004). So here, considering whether the area measurement error concerns positional displacements, we define AAE and RAE as Numerical Error (NE), and those errors with concerns of spatial changes as Spatial Error (SE).

Numerical error stabilizes at a large extent (Hollister *et al*., 2004). In a GIS environment, it is especially important to retain the same spatial coverage of each polygon after rasterization, which has motivated us to pay more attention to spatial error (Piwowar *et al*., 1990).

Relative Commission Area Error (RCAE) and Relative Omission Area Error (ROAE) are defined as:

$$RCAE = \frac{\text{Commission area}}{\text{Area before rasterization}} \tag{3}$$

$$ROAE = \frac{\text{Omission area}}{\text{Area before rasterization}} \tag{4}$$

AAE and RAE can also be calculated from RCAE and ROAE by:

$$AAE = (\mathrm{R}CAE - ROAE) \times (Area\,before\,Rasterization) \tag{5}$$

$$RAE = \mathrm{R}CAE - ROAE \tag{6}$$

3 Influencing factors for area measurement errors

Based on the process of rasterization, we classify the influencing factors into three categories: (1) characteristics of grid mesh or vector data: grid cell size, polygon size and polygon shape; (2) relative overlay relationships: relative position and relative orientation; and (3) rasterization methods: classification assignment algorithm.

3.1 Grid cell size, polygon size and cell/polygon size ratio

Cell size and polygon size play a key role in affecting the magnitude and distribution of area measurement errors. Increasing polygon size will decrease area errors (Dunn *et al*., 1990). Actually, cell size and polygon size is a pair of relative values. Researchers always study the error changes affected by one value after fixing the other one.

Cell/ polygon size ratio is more meaningful and significant, such that a reduction of error is observed as cell/polygon size ratio decreases (cell size becomes much smaller than polygon size) (Frolov and Maling, 1969; Piwowar *et al*., 1990; Carver and Brunsdon, 1994). Theoretically, the area measurement error is proportional to the square of cell/polygon size ratio, which means the error should reduce by half when the ratio decreases to ¼ (Switzer, 1975). In general, the total area error will remain quite consistent and drop below 3% after the cell/polygon size ratio decreases to 1/10 (Congalton, 1997; Shortridge, 2004). However, this relationship is indistinct when cell/polygon size ratio is large (larger than 0.5) (Brunsdon and Carver, 1991). There are no significant errors when the polygon is large enough relative to the cell size (Wade *et al*.,2003).

Practically, it is more attractive to find out how to choose an appropriate cell size. Basically, four rules can be employed. (1) set the cell size as the smallest polygon (or ½ , or ¼ of this area) to maintain the accuracy of the data (Switzer, 1975; Congalton, 1997). If a cell size is larger than ½ of the smallest polygon, 100 percent errors should be expected for the smallest polygons (Wehde, 1982). (2) set the cell length equivalent to the smallest vector distance between two points (Seong and Usery, 2001). (3) consider the national

map accuracy standard relating to the nominal scale of the vector data. (4) consider the cell size of raster products of the relative nominal scale (Shortridge, 2004).

3.2 Polygon shape

The concept of polygon shape is much richer than polygon size, which makes it impossible to differentiate all shapes only by a single index (Forman, 1995). Some shape indices, such as boundary index, angular complexity, perimeter area ratio, fractal dimension, shape index and so on, have been used to reveal and model the relationship between shape variation and rasterization errors (Carver and Brunsdon, 1994; Congalton, 1997; Longley *et al*., 2001). Shapes of the features can be dramatically changed after rasterization: elongated river polygons may be omitted while circles will never be split into multiple polygons. Generally, to a map of constant extent, area measurement errors are affected by boundary length and boundary complexity. An increase of the total boundary length of each category will enlarge area errors (Shortridge, 2004). The more complex the region is, the larger the error becomes (Carver and Brunsdon, 1994; Liu *et al*., 2001).

3.3 Relative position

Relative position between a map and a grid mesh affects the accuracy of rasterization. A small polygon may be lost if the grid mesh is on one position, while it may be included if the grid mesh is shifted (Congalton, 1997). In statistics, this error can be compensated by repeating the rasterization process several times on different random positions (Frolov and Maling, 1969). In practice, it is time-consuming and costly, so the grid origin is often located at one corner of the source data, typically the lower-left corner (Shortridge, 2004). Although relative position is a significant source of area measurement variation for individual polygons, the numerical error of a whole map will only change slightly by position shifts, especially for smaller cell sizes (Wehde, 1982; Shortridge, 2004)

3.4 Relative orientation

Orientation could be a considerable interest for area measurement errors, especially for elongated features, whose area calculations may be largely affected by various orientations (Congalton, 1997; Shortridge, 2004). Area measurement errors caused by orientation are unpredictable (Kam *et al*., 2000). In common sense terms, the raster grid cell will be oriented parallel to the coordinate system of the source data (Carver and Brunsdon, 1994).

3.5 Rasterizing methods

There are mainly three very popular rules for cell attribute assignment during rasterization.

(1) Largest Area Rule: the category of the largest area in a cell zone is assigned to the whole grid cell. If more than one cut of the same category are located in one cell, summation is taken for comparison (Congalton, 1997).

(2) Central Point Rule: the category located at the central point of the cell is assigned to the whole grid cell (Nichols, 1975; Shortridge, 2004).

(3) Highest Weight Rule: the most important category, which owns the highest weight in the cell, is assigned to the whole grid cell (Congalton, 1997).

Most commonly, the largest area rule is employed, but it greatly depends on the shape and pattern of polygons (Zhou *et al*., 2007). As long as polygon size is large enough relative to cell size, the rasterizing method appears to have little effect on area measurement errors (Shortridge, 2004).

4 Discussions and conclusions

Although some researchers have dedicated to reveal the relationships between influencing factors and area measurement errors, no quantified results have been widely accepted yet. There is still an urgent need to model and evaluate these area measurement errors (Dunn *et al*., 1990), especially in the following aspects:

(1) Quantify the influencing factors. Most researchers concentrate on the influence of cell size on area measurement errors, while only few researchers investigate other factors. However, all factors, except for cell size, are not quantified clearly, which make it hard to reveal the influencing mechanism precisely.

(2) Model the variation of spatial errors. In GIS environment, except for the numerical changes of area measurement errors, spatial changes are more attractable to precisely represent vector data by raster format.

(3) Explore algorithms to minimize area measurement errors. Some algorithms have been designed to minimize area measurement errors caused by rasterization (Zhou *et al*., 2007; Liao and Bai, 2009). However, none of them has been used in commercial software.

(4) Label errors in the raster data. Area measurement errors cannot be omitted from the rasterization process. Therefore, mechanism should be designed to label the spatial distribution of errors, such as error maps, which can promote the raster data to be used accurately and efficiently.

Acknowledgments

The authors gratefully acknowledge the support provided by The Hong Kong Polytechnic University (Project No. G-YX0P, G-YF24, G-YG66, 1-ZV4F) and Hong Kong RGC General Research Fund (Project No. 5276/08E).

References

Brunsdon, C., Carver, S. (1991), ”The accuracy of digital representation of 2D and 3D geographical objects: a study by simulation”. In Geographical Information Systems Modelling and Policy Systems, edited by M.M. Fischer and P. Nijkomp (Berlin: Springer-Verlag), 115-130.

Carver, S.J., Brunsdon, C.F. (1994), ”Vector to raster conversion error and feature complexity: an empirical study using simulated data”. International Journal of Geographic Information Systems, 8(3): 261-270.

Congalton, R.G. (1997), ”Exploring and evaluating the consequences of vector-to-raster and raster-to-vector conversion”. Photogrammetric Engineering and Remote Sensing, 63(4): 425-434.

Dunn, R., Harrison, A.R., White, J.C. (1990), ”Positional accuracy and measurement error in digital databases of land use: an empirical study”. International Journal of Geographical Information Systems, 4(4): 385-398.

Forman, R.T.T. (1995), ”Land mosaics: the ecology of landscapes and regions”. Cambridge University Press, Cambridge, United Kingdom.

Frolov, Y.S., Maling, D.H. (1969), ”The accuracy of area measurements by point counting techniques”. Cartographic Journal, 6: 21-35.

Hollister, J.W., Gonzalez, M.L., Paul, J.F., August, P.V., Copeland, J.L. (2004), ”Assessing the accuracy of national land cover dataset area estimates at multiple spatial extents”. Photogrammetric Engineering and Remote Sensing, 70(4): 405-414.

Kam, S.P., Chu, T.H., Alveran, A. (2000), ”Area representation errors associated with Rasterization”. In Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Amsterdam, the Netherlands.

Liao, S.B., Bai, Y. (2009), ”A new grid-cell-based method for error evaluation of vector-to-raster conversion”. Computational Geoscience, DOI: 10.1007/s10596-009-9169-3.

Liu, M.L., Tang, X.M., Liu, J.Y., Zhuang, D. (2001), ”Research on scaling effect based on 1 km grid cell data”. Journal of Remote Sensing, 5(3): 183-190.

Longley, P.A., Goodchild, M.F., Maguire, D., Rhind, D.W. (2001). Geographical Information Systems and Science (Ney York: Wiley).

Nichols, J.D. (1975), ”Characteristics of computerized soil maps”. Soil Science Society of America Journal, 39: 927-931.

Piwowar, J.M., Ellsworth, F.L., Dudycha, D.J. (1990), ”Integration of spatial data in vector and raster formats in a geographic information system environment”. International Journal of Geographical Information Systems, 4(4), 429-444.

Seong, J.C., Usery, E.L. (2001), ”Assessing raster representation accuracy using a scale factor model”. Photogrammetric Engineering and Remote Sensing, 67(10): 1185-1191.

Shortridge, A.M. (2004), ”Geometric variability of raster cell class assignment”. International Journal of Geographical Information Science, 18(6): 539-558.

Switzer, P. (1975), ”Estimation of the accuracy of qualitative maps”. Display and Analysis of spatial data, edited by J. C. Davis and M. J. McCullagh (New York: Wiley), 1-13.

Van Der Knaap, W.G.M. (1992), ”The vector to raster conversion: (mis)use in geographical information systems”. International Journal of Geographical Information Systems, 6(2): 159-170.

Wade, T.G., Wickham, J.D., Nash, M.S., Neale, A.C., Riitters, K.H., Jones, K.B. (2003), ”A comparison of vector and raster GIS methods for calculating landscape metrics used in environmental assessments”. Photogrammetric Engineering and Remote Sensing, 69(12): 1399-1405.

Wehde, M. (1982), ”Grid cell size in relation to errors in maps and inventories produced by computerized map processing”. Photogrammetric Engineering and Remote Sensing, 48(8): 1289-1298.

Zhou, C.H., Ou, Y., Yang, L., Qin, B. (2007), ”An equal area conversion model for Rasterization of vector polygons”. Science in China Series D: Earth Sciences, 50: 169-175.

Multiphase sampling using expected value of information

Sytze de Bruin, Daniela Ballari & Arnold Bregt

Wageningen University, Laboratory of Geo-Information Science and Remote Sensing,
P.O. Box 47, 6700 AA, The Netherlands
Sytze.deBruin@wur.nl, Daniela.Ballari@wur.nl, Arnold.Bregt@ wur.nl.

Abstract

This paper explores multiphase or infill sampling to reduce uncertainty after an initial sample has been taken and analysed to produce a map of the probability of some hazard. New observations are iteratively added by maximising the global expected value of information of the points. This is equivalent to minimisation of global misclassification costs. The method accounts for measurement error and different costs of type I and type II errors. Constraints imposed by a mobile sensor web can be accommodated using cost distances rather than Euclidean distances to decide which sensor moves to the next sample location. Calculations become demanding when multiple sensors move simultaneously. In that case, a genetic algorithm can be used to find sets of suitable new measurement locations. The method was implemented using R software for statistical computing and contributed libraries and it is demonstrated using a synthetic data set.

Keywords: Iterative sampling, adaptive sampling, infill sampling, decision analysis, mobile sensors.

1 Introduction

After a major incident such as the recent fire in a chemical factory in Moerdijk (January 5, 2011), The Netherlands, authorities have to decide whether or not food produced in the vicinity of the accident is suitable for human consumption. Such decision making typically relies on information obtained from a small sample, but it may improve when non-covered regions are "filled in" by additional sampling (Johnson, 1996; Cox *et al.*, 1997) by mobile sensors. The costs of misclassification in cases such as depicted above are often unequal for type I and type II errors, with the costs of false negatives or "safe" decisions being higher than those of false positives. Selection of new sample locations should therefore account for this difference. At the same time, the costs for visiting new sites may differ between mobile sensors located within the area. For example, sensors situated near a new sample location need less travelling.

The method described in this paper involves optimising new sample locations based on information obtained from the previous sample. The phenomenon to be mapped is considered static within the time frame of the analysis (*e.g.* surface contamination after an incident). Expected value of information (EVOI) is used for quantifying the suitability of the sample. EVOI expresses the benefit expected from data collection prior to actually doing the measurements (De Bruin *et al.*, 2001; De Bruin and Hunter, 2003; Back *et al.*,

2007). In contrast to kriging variance (Baume *et al.*, 2011) and entropy based methods (Zidek *et al.*, 2000), EVOI is data dependent and it can incorporate different misclassification costs for false positives and false negatives. Heuvelink *et al.* (2010) used a stochastic model of the environmental phenomenon and also accounted for differences between misclassification costs, but here a direct Bayesian approach is used that is potentially faster when few samples are added per iteration.

Our aim is to demonstrate and discuss some strategies for using EVOI to add observations to a previous sample while accounting for constraints imposed by a sensor network.

2 Methods

2.1 Expected value of information

EVOI is estimated as the difference between expected costs at the present stage of knowledge and expected costs when new information becomes available. Figure 1 shows a tree with square nodes indicating decisions to place a sensor for measuring the phenomenon at some location and decisions about mapping presence or absence of the phenomenon using the information at hand. Chance nodes (circles) indicate the outcome of random events once a decision has been taken. For example, if a sensor is placed, measurement with it may indicate presence (signal) or absence ($\overline{signal}$) of the phenomenon. The probability of obtaining a sensor signal at some location, Pr(signal) can be computed from sensor properties and the prior probability of presence, Pr(present), as follows (1):

$$\begin{aligned}\Pr(signal) = \Pr(signal \mid present) \times \Pr(present) + \\ \Pr(signal \mid absent) \times \Pr(absent)\end{aligned} \quad (1)$$

where $\Pr(signal \mid present)$ is the probability that a warning is issued if the phenomenon is present and $\Pr(\overline{signal} \mid absent) = 1 - \Pr(signal \mid absent)$ is the probability that the sensor correctly gives no signal. These probabilities are given in the sensor specifications (i.e. sensitivity and specificity).

Decision making is assumed to be based on Bayes actions, i.e. minimising expected loss. Accordingly, placing a sensor is sensible if the expected loss of the upper branch of Figure 1 is lower than the expected loss of the lower branch. If only misclassifications involve costs, the latter is calculated as (2):

$$E(cost_{lower}) = \min(cost_{false\_positive} \times \Pr(absent),\quad cost_{false\_negative} \times \Pr(present)) \quad (2)$$

where min(.) is a function returning the minimum of its arguments and $cost_{false\_negative}$ and $cost_{false\_positive}$ are costs of misclassification. The conditional probabilities shown in Figure 1 are calculated with Bayes' rule, e.g (3):

$$\Pr(absent \mid signal) = \frac{\Pr(signal \mid absent) \times \Pr(absent)}{\Pr(signal)} \quad (3)$$

Hence, the expected cost of the upper branch is calculated by (4):

$$
\begin{aligned}
E(cost_{upper}) = \Pr(signal) \times \min(&cost_{false\_positive} \times \Pr(absent \mid signal), \\
&cost_{false\_negative} \times \Pr(present \mid signal)) + \\
\Pr(\overline{signal}) \times \min(&cost_{false\_positive} \times \Pr(absent \mid \overline{signal}), \\
&cost_{false\_negative} \times \Pr(present \mid \overline{signal}))
\end{aligned}
\tag{4}
$$

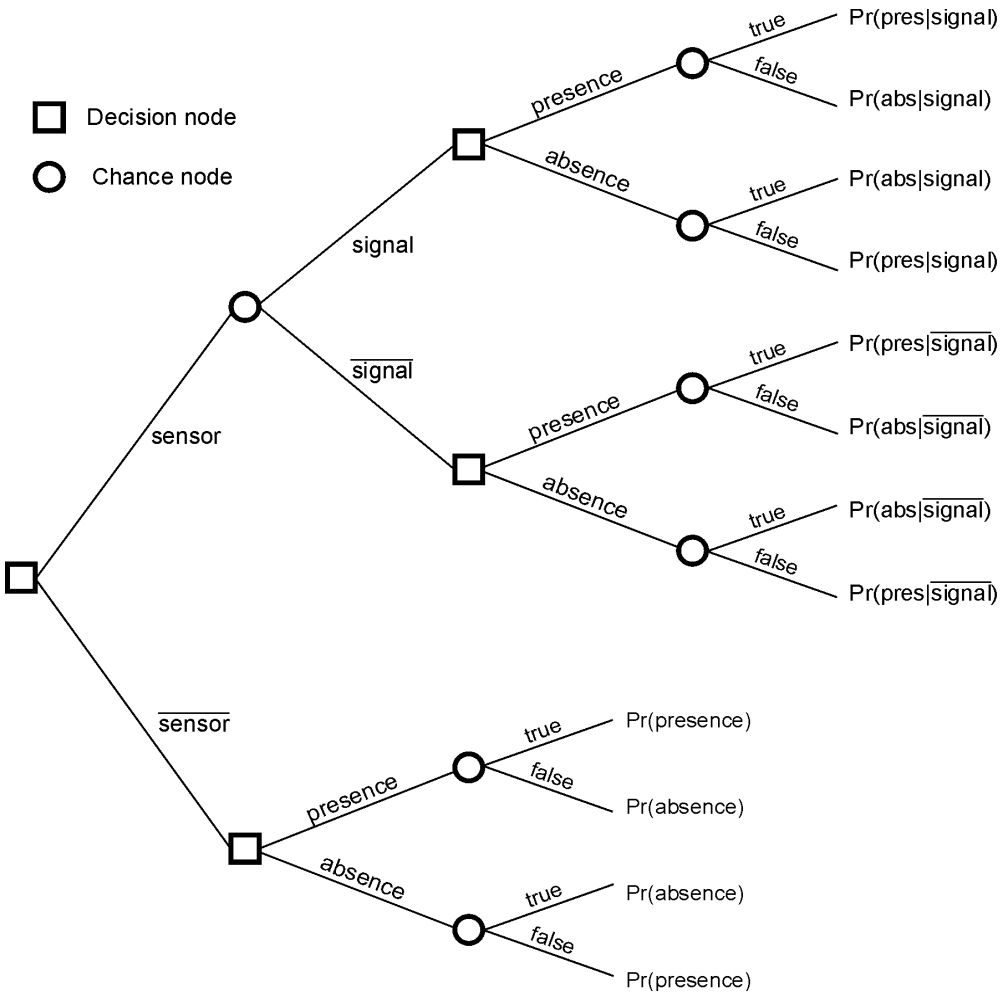


Figure 1. Decision tree showing decisions to place a sensor or not and to map presence or absence of a phenomenon (e.g. hazard).

EVOI is the difference between $E(cost_{lower})$ and $E(cost_{upper})$, where *lower* refers to the lower branch of the decision tree and *upper* to the upper branch. We consider the aggregated expected costs of misclassification over the study area and find a single optimal sample location as the one that maximises EVOI and thus minimises $E(cost_{upper})$. The aggregated costs of misclassification are computed by creating maps for both a signal and no signal obtained at the sensor location and multiplying the expected costs for these situations with the probability of their occurrence. If the locations of two or more observations are to be simultaneously optimised, complexity of the computations increases, since nearby observations are typically conditionally dependent. At the same time the size of the solution space increases substantially. For example, with two simultaneous observations, four expected cost maps and their probabilities need to be computed for each pair of locations while solution space increases by a factor (n-1), with n being the number of potential sample locations. This situation was handled using a genetic algorithm.

2.2 Case study

A case study was conducted using a synthetic data set constructed by applying a threshold at 20 to a Gaussian random field of 100 x 100 grid cells of unit size with mean 20 nugget 1 and a spherical structural spatial correlation component with range 40 and a partial sill (semivariance) 16. Sensor data were obtained by sampling the synthetic data and adding random measurement error. The initial sample consisted of 16 points on a regular grid. Sensor data were interpolated using indicator kriging. Computations were done in R (Venables *et al.*, 2010) using the geostatistical package *gstat* (Pebesma, 2004) and the genetic algorithm implemented in the package *genalg*.

Three approaches were considered for adding new sample locations to the original sample: (1) add single location at a time, move sensor with lowest cost (in this case Euclidean distance); (2) add two locations simultaneously and scan only the area that can be reached by the sensors within one time step; (3) add two sample locations simultaneously, scan the whole area, and move the sensors with lowest cost. The costs of misclassification were arbitrarily set at 2 and 3 (no unit) for false positives and false negative, respectively.

3 Results

Figure 2a shows probabilities of occurrence interpolated from the initial sample of 16 sites. Figure 2b shows the map of global EVOI, i.e., EVOI computed after aggregating expected misclassification costs for observations made at each grid location, separately. The best location thus corresponds to the highest global EVOI. Not surprisingly, this occurs between observations differing in value (indicated by arrow).

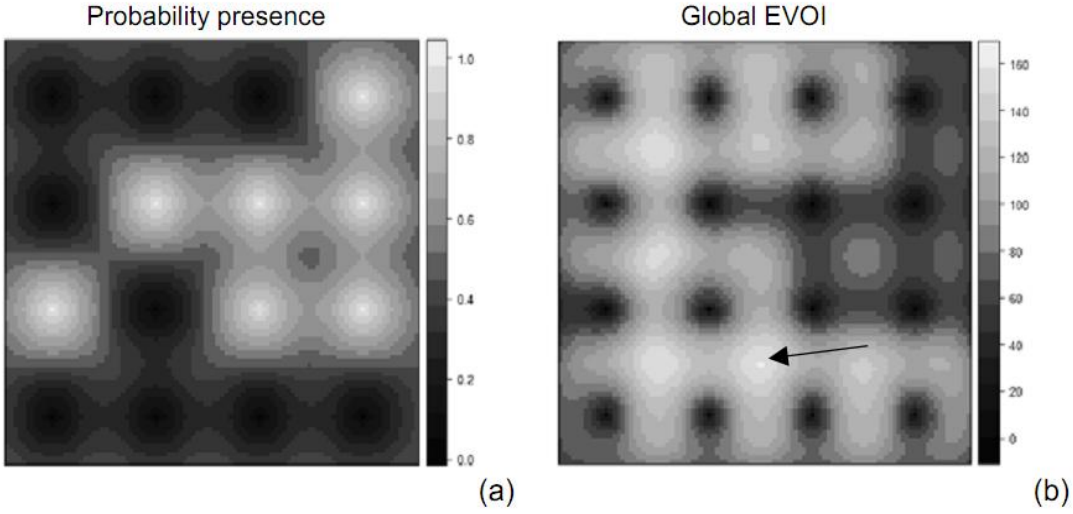


Figure 2. Probability map (a) and global EVOI (b) computed from the initial sample of 16 regularly spaced points. The arrow points to the location having highest global EVOI.

Figure 3 shows an example of an optimised sensor configuration after the 17th observation has been made (16 initial and 1 infill measurements) on a backdrop of the probability of presence of the phenomenon (cf. Figure 2). Euclidean distance was used for deciding which sensor to move to the next location, but another cost criterion could have been used with only minor modification of the algorithm.

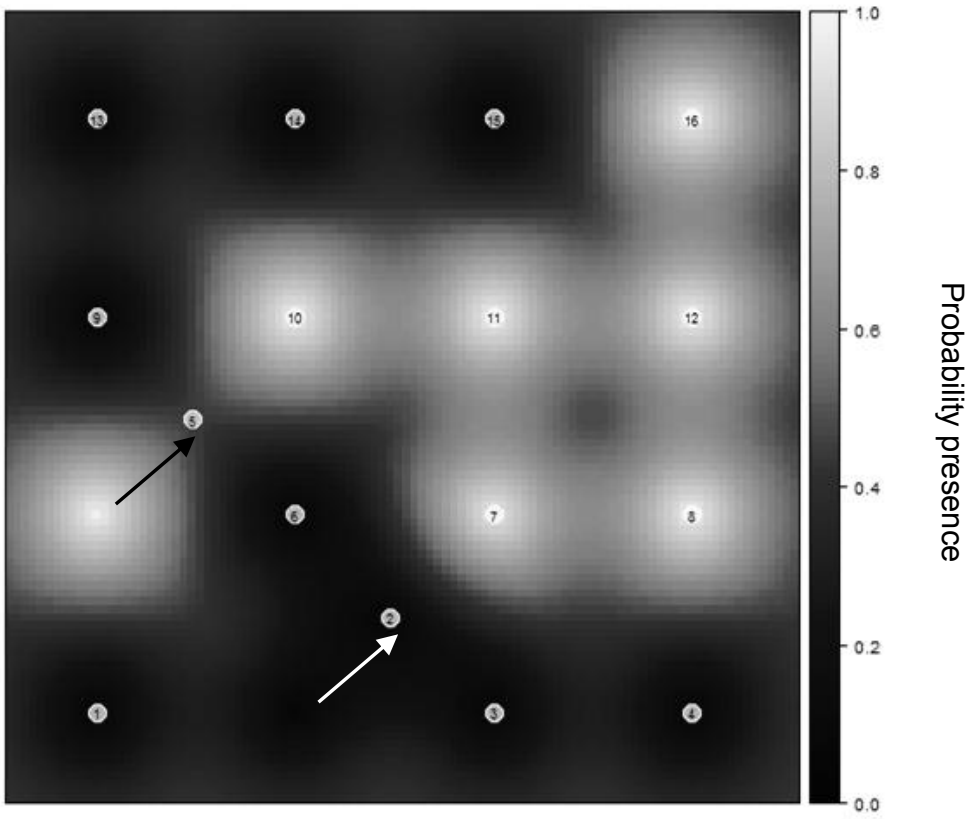


Figure 3. Configuration of initially regularly spaced sensors after two iterations with a single observation per step (approach 1). First sensor 2 moved (white arrow) and a measurement was taken, next sensor 5 moved (black arrow), but the measurement has not yet been taken.

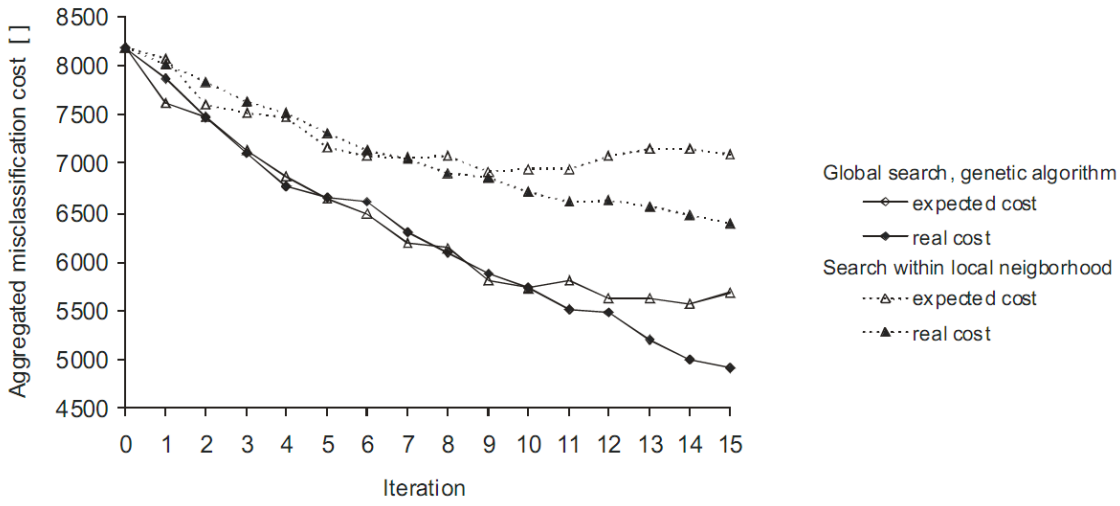


Figure 4. Effect of the way sensor constraints are taken into account on aggregated misclassification costs with two simultaneously moving sensors (approaches 2 and 3).

Figure 4 shows the effect of the two approaches to account for sensor constraints described in section 2.2, with two simultaneously moving sensors. Not surprisingly, both expected and real misclassification costs were much lower when the full study area was scanned in search of the best sample locations. Of course, in the case of local sensor neighbourhood scanning, results depend on the start locations chosen. Large differences between real costs (normally not known) and expected costs are indicative of misspecification of the geostatistical model used for interpolating the probability map.

4 Conclusions

The Expected value of information (EVOI) approach puts new observations at locations that intuitively make sense and it can help deciding when to stop a survey. The method accounts for specified misclassification costs; these can be dissimilar for different kinds of errors (*e.g.* false positives or false negatives). Constraining potential sample locations to the space that can be travelled by a small set of mobile sensors is a bad idea since the sensors may get trapped in some area and may thus fail to visit potentially interesting spots. Rather, cost distances can be used for deciding which sensors to move to next globally optimal locations. Genetic algorithms may be useful for optimising the sample locations for multiple sensors moving simultaneously.

References

Back, P.E., Rosén, L., Norberg, T. (2007), "Value of information analysis in remedial investigations". *Ambio,* Vol.36: 486-493.

Baume, O.P., Gebhardt, A., Gebhardt, C., Heuvelink, G.B.M., Pilz, J. (2011), "Network optimization algorithms and scenarios in the context of automatic mapping", *Computers & Geosciences,* In Press, Corrected Proof.

Cox, D.D., Cox, L.H., Ensor, K.B. (1997), "Spatial sampling and the environment: some issues and directions". *Environmental and Ecological Statistics*, Vol. 4: 219-233.

De Bruin, S., Bregt, A., Van de Ven, M., (2001), "Assessing fitness for use: the expected value of spatial data sets", *International Journal of Geographical Information Science*, Vol.15: 457-471.

De Bruin, S., Hunter, G.J. (2003), "Making the trade-off between decision quality and information cost". *Photogrammetric Engineering and Remote Sensing* Vol. 69: 91-98.

Heuvelink, G.B.M., Jiang, Z., De Bruin, S., Twenhofel, C.J.W. (2010), "Optimization of mobile radioactivity monitoring networks". *International Journal of Geographical Information Science,* Vol. 24: 365-382.

Johnson, R.L. (1996), "A Bayesian/geostatistical approach to the design of adaptive sampling programs". *In:* Rouhani, S., Srivastava, R.M., Desbarats, A.J., Cromer, M.V., Johnson, A.I. (eds.). *Geostatistics for Environmental and Geotechnical Applications*, American Society for Testing and Materials**,** pp. 102-116.

Pebesma, E.J. (2004), "Multivariable geostatistics in S: the gstat package". *Computers & Geosciences*, Vol. 30: 683-691.

Venables, W.N., Smith, D.M., R Development Core Team (2010), *An Introduction to R*, The R Foundation for Statistical Computing, Vienna, Austria, 101p.

Zidek, J.V., Sun, W.M. Le, N.D. (2000), "Designing and integrating composite networks for monitoring multivariate Gaussian pollution fields". *Journal of the Royal Statistical Society Series C-Applied Statistics*, Vol 49: 63-79.

Which Spatial Quality can be Meta-Propagated?

Didier G. Leibovici, Amir Pourabdollah & Mike Jackson

Centre for Geospatial Science, University of Nottingham, U.K
Didier.Leibovici@nottingham.ac.uk, Amir.Pourabdollah@nottingham.ac.uk,
Mike.Jackson@nottingham.ac.uk

Abstract

As a mid-way between the usability criterion proposed by the upcoming ISO19157 and the error-propagation methods, the concept of meta-propagation allows to derive an estimate of the propagated uncertainties using only metadata information about quality of the datasets and processes used in a workflow. The principle of meta-propagation has been illustrated using the thematic accuracy quality with the quantitative attribute accuracy sub-element. The purpose of this paper is to explore the other quality elements, which can be meta-propagated. Using a similar approach as for the thematic accuracy, one needs also to address further the metadata quality for processes as they are dependent in this context. The paper focuses on generic principles and few specific situations where appropriate quality measures can be fully described. Basic meta-propagation, based on separability of the assessments, (i.e., one input to one output) is presented but one also discusses the potentiality of using quality measures for non-separable approaches. Spatiality of the measures, i.e., taking the benefit of a spatial data quality value being a map are also investigated.

Keywords: metadata, spatial data quality, geo-processing, uncertainty, error propagation, scientific workflow.

1 Introduction

The upcoming spatial data quality standard ISO19157 will soon replace the ISO19113, ISO19114 and ISO19138; it is more an integration and a simplification rather than a remould but one new addition in the quality elements is `DQ_Usability`. Usability is described as: "it is the degree of adherence to a specific set of quality requirements", " it shall be used to describe specific quality information about a dataset's adherence to a particular application or requirements" and "it may be used to declare the conformance of the dataset at a particular specification", for example for a particular usage within a specific application.

If one wants to go further than just a green light, one may want to assess the impact of using this particular dataset in terms of quality of the output that this application can provide when using this dataset. The full quantification of the green light provided `DQ_Usability` is in fact dependent on the processing tasks used in the application and on the other datasets used in the workflow.

Recently Leibovici *et al.* (2010ab) have proposed a mid-way between the *usability* described above and the error propagation which is obtained only at runtime: the meta-propagation of uncertainties. The meta-propagation enables to have an estimate of the

error propagated within the workflow used for the application without running the scientific model.

The principle of meta-propagation is defined as using metadata about quality, from data and geo-processes involved in the workflow, to derive an uncertainty assessment of the outputs of a given workflow. A computational assessment ,seen as a meta-workflow dealing with quality measures and their values, acts as a substitute of an error propagation analysis which would involve running, in a Monte Carlo experiment, the whole workflow many times. It has been used and demonstrated with *quantitative attribute accuracy* within the *thematic accuracy* principle (Leibovici *et al.,* 2011; Pourabdollah *et al*., 2011).

In order to extend this principle to further quality aspects, this paper addresses (i) the use of other quality elements for the meta-propagation and investigates (ii) the separable versus non-separable approaches for the quality assessments (iii) the spatial specificity for the quality elements and their propagation. The geo-process quality elements have not been standardised yet and to be able to investigate those three points, the paper also introduces some general principles for geo-process quality and some associated measures.

2 Quality of spatial data and quality of geo-processing

From the current ISO standard 19113 about data quality principles of *completeness, logical consistency, positional accuracy, temporal accuracy and thematic accuracy*, one can derive similar and nearly matching principles for geo-processes (Table 1) (Leibovici *et al.,* 2010b). These principles are focusing on the variability, uncertainties and reliability of the output in term of decision-making and not on the quality of services (QoS) focusing on the operational aspects such as response time (Leibovici *et al.*, 2009). Not all are directly devised for external quality assessments such as *conceptual validity* being more a set of internal quality properties.

The meta-propagation of quality elements depends on the ability of the quality principles for data and processes, to be used in conjunction, according to the measures used for both. For the *quantitative attribute accuracy* data quality measured by a variance, we introduced the *variance transfer functions* for corresponding process quality that can be derived from classical variance-based sensitivity analysis (Saltelli *et al.,* 2008). As in the latter, a separable approach is the simplest approach, but obviously during the geo-processing the quality elements associated to many data inputs will influence the quality of one or more outputs in general. In the mean time spatial properties may play an important role in quality assessment of a workflow. These two aspects can also be considered in meta-propagation (see next section) but are nonetheless challenging as much on the computational side for the propagation as on the metadata production in the first place.

Basic measures were also given in ISO 19138 and in the ISO 19157 such as the *rate of missing items* for the *omission* in *Completeness*. *Completeness* qualities could meta-propagate as is, but may also depend on one hand on the *conflation* principle (and the hereby given measure), and on the other hand on the semantic of the output considered. Simply multiplying the rates, say *omission* by *information loss* will give a new omission, but this will also depend on the scope attributes of these elements.

Table 2. Geo-Processing quality elements and sub-elements.

Process quality element	Process quality sub-element	Definition {example of a *basic measure*}
conflation	information loss	loss in conflating input data sources {*rate*}
	information gain	gain in conflating input data source {*rate*}
conceptual validity	semantic conformance	adherence to the semantic relations within the "disciplinary" domain {}
	domains integration	level of integrated modelling in relation to the "disciplinary" domains involved {*rate or ?*}
logical validity	conceptual conformance	adherence to rules of the conceptual model {*rate or ?*}
	domain conformance	adherence to the output data values of the domain {*rate or pdf*}
	computational format	degree to which the encoding format follows standards{rate or ?}
	topological preservation	preservation of the explicitly encoded topological characteristics of the input data sources {rate or ?}
positional error propagation	absolute error propagation	propagation of the uncertainty in the absolute positions of features in datasets {*inertia transfer function*}
	relative error propagation	propagation of the uncertainty in the relative positions of features in datasets
	gridded error propagation	propagation of the uncertainty in the gridded data position values {*inertia transfer function*}
	scale preservation	preservation of scale(s) of the input datasets {?}
	spatial scale error propagation	propagation due to outranging scale conformance of input datasets {?}
temporal error propagation	time propagation	propagation of the uncertainty in the time measurement {*variance transfer function*}
	time scale propagation	error propagation due to outranging scale conformance for the input datasets {*variance (or inertia) transfer function*}
thematic error propagation	impact of classification correctness	propagation of uncertainty due to departure from accurate classification {misclassification transfer function or ?}
	impact of non-quantitative attribute correctness	propagation of uncertainty due to correctness of non-quantitative attribute {}
	quantitative attribute error propagation	propagation of uncertainty of quantitative attribute {*variance (or inertia) transfer function*}

inertia generalising the variance as trace(Σ) (see text)

Other Boolean and rate based measures with matching principles, such as *logical consistency* and *logical validity*, can be used in the same way for meta-propagation, noticing that here procedures to establish these rates for geo-processing quality needs either expert judgment or appropriate experiments.

The *thematic accuracy* sub-element *classification correctness* directly linked its equivalent for processes about its *impact*, may appears difficult to meta-propagate as this may depend strongly on which misclassification is occurring and may be also very specific to the actual processing involved. Here the role of one to one meta-propagation (separable approach) may help to fully specify the type of impact on uncertainty attributable to misclassification. As for *quantitative attribute accuracy* (and its propagation), using a variance measure, a misclassification can easily impact on another misclassification using a *misclassification transfer function*, which may be a collection of multivariable functions (one for each element of the output misclassification function). Monte-Carlo experiment can also be used to estimate the impact of misclassifications on a *quantitative attribute accuracy* measured by a variance. Some straightforward examples of use case are any model involving the use of land cover data and land use data (two classifications), for example to model CO2 absorption (quantitative attribute).

Nearly all of the *Positional accuracy* and *Temporal accuracy* quality sub-elements can borrow from the demonstration on *quantitative attribute accuracy* (Leibovici *et al*., 2011) to be able to perform a meta-propagation of uncertainty.

When considering more than one variable such as with the CE90 (Circular Error at 90% measure) an *inertia transfer function* instead of a *variance transfer function* can operate the meta-propagation. The inertia measure, *i.e.* the sum of variances (like the CE90 measure) or the trace of the variance-covariance matrix can be also used in a multivariate context as for a non-separable approach as discussed in the next section.

3 Non-separability and spatiality issues

In the previous section, a one input to one output error propagation, using metadata quality to derive uncertainty information, has been described for a single geo-process. Uncertainty information can be propagated through a whole workflow to the final outputs by either recording all the possible paths (set of all propagated values towards one output) and/or making summaries of the derived quality values. A summary can be the mean of the values or the maximum of them, if the sets of values are numerical. Nonetheless the question remains about usefulness of this derived accuracy information in term of decision-making. The *usability* criterion newly introduced gives a short-cut answer but has the merit to raise the problem, to permit its expression and to allow further research on it (Goodchild, 2009; Zargar and Devillers, 2009; Devillers *et al.,* 2010). This becomes a crucial point when sharing interoperable data and geo-processing services such as in the GEOSS for various applications (*e.g.* Giuliani *et al.,* 2011). The metadata quality standard addresses all the single aspects and put a step forward with this *usability* element, which goes into the concept of error propagation and acknowledges by its definition the multivariate and non-separable context.

Multivariate error propagation, multivariate sensitivity and multivariate calibration are difficult to perform because of the sampling schema when dependence is taken into account (Saltelli *et al.,* 2008; Kurowicka and Cooke, 2006). For meta-propagation of uncertainties, the difficulty also comes from the fact it would mean getting a hand on the input multivariate distribution just by knowing the marginals. This adds up an extra metadata

quality parameter, needed as a process quality, which may be nonetheless approximated using a dependence model associated to a copula distribution (Sklar, 1959). A simpler set-up, using a multivariate variance transfer function as process quality element for the meta-propagation may constitute a good compromise multivariate complexity and separable error propagations.

Most of the current/basic measures mentioned for data qualities are in fact often aggregating spatial information quality; rarely pointwise information will be filled in the metadata quality. In the meantime quality measure for geo-process would have to match the support of the measures for data quality. Taking into account he spatial distribution of the uncertainties may play a very important role in the uncertainty of the output (Heuvelink, 2002). As potentially the metadata quality measures for geo-processes can take into account the spatial correlation of uncertainties, this can be also used in meta-propagation. Nonetheless spatial auto-correlation will have to be taken into account when performing experiments to define the geo-processes metadata quality values. For example the *variance transfer function* can be a geo-processing function itself dealing with map input, a map of uncertainties.

4 Conclusion

The *usability* concept for data quality raises the debate about the real purpose and the usefulness of the spatial quality elements? In reviewing the elements and matching or associated processes metadata quality, which together can be used for the meta-propagation of data quality, the paper aims to contribute to this debate. First of all, are these quality element really treated as spatial quality? Can the error propagation practice and the meta-propagation warn on the need of good meta-quality for usefulness? On the meta-meta level, *usability* and meta-quality could be promoted as subjective assessments that users may value first.

Meta-propagation cannot replace an uncertainty analysis of the whole workflow instance, but it gives an approximate answer to it and also at each steps of the workflow, in a fast computational way. The existence of the metadata for data quality, is a limitation to any error propagation method. Meta-propagation needs also process quality metadata which can be seen as beforehand atomic uncertainty analyses and adds up to the burden of providing metadata information. So, in the title, the word “which” plays a double act, it asks also: is there any or enough data having metadata about quality information? Tools such as DUE (Brown and Heuvelink, 2007) linked to metadata encoding tools and spatial quality standards are needed to allow users to annotate appropriate data quality and *idem* for processing services.

When considering error propagation by meta-propagation another question is emerging: “these quality information interact in influencing the outcome quality, don’t they?” This is relevant to any kind of error propagation method: meta-propagation or propagation at run-time (using a full model, a simulator or an emulator), and opens up a serious computational challenge.

Acknowledgments

This work was supported by European Framework Program 7 (FP7) ENV.2008.4.1.1.1: European Environment Earth Observation system supporting

INSPIRE and compatible with GEOSS (Global Earth Observation System of Systems):"EuroGEOSS", 2009-2012, http://www.eurogeoss.eu

References

Brown, J.D., Heuvelink, G.B.M., (2007) "The Data Uncertainty Engine (DUE): A software tool for assessing and simulating uncertain environmental variables". *Computers & Geosciences,* Vol. 33(2): 172–190.

Devillers, R., Stein, A., Bédard, Y., Chrisman, N., Fisher, P., Shi, W. (2010), "Thirty Years of Research on Spatial Data Quality: Achievements, Failures, and Opportunities". *Transactions in GIS,* Vol. 14(4): 387-400.

Giuliani, G., Ray, N., Lehmann, A. (2011), "Grid-enabled Spatial Data Infrastructure for environmental sciences: Challenges and opportunities". *Future Generation Computer Systems,* Vol. 27(3): 292–303.

Goodchild, M. F. (2009), "The quality of geospatial context". *In: Proceedings of the 1st international conference on Quality of context, (QuaCon'09), Springer-Verlag,* pp. 15–24.

Heuvelink, G.B.M. (2002), "Analysing uncertainty propagation in GIS: why is it not that simple?" *In: Foody G.M., Atkinson, P.M, (eds.). Uncertainty in Remote Sensing and GIS,* John Wiley & Sons, pp.155-166.

Kurowicka, D., Cooke, R. (2007), *Uncertainty analysis with high dimensional dependence modelling*. John Wiley & Sons, 284p.

Leibovici, D.G., Hobona, G., Stock, K., Jackson, M. (2009), "Qualifying geospatial workfow models for adaptive controlled validity and accuracy". *In: IEEE proceedings 17th International conference on GeoInformatics, August 2009, USA*, pp. 1-5.

Leibovici, D.G., Pourabdollah, A. (2010a), "Workflow Uncertainty using a Metamodel Framework and Metadata for Data and Processes". *OGC TC/PC Meetings, 20-24 September 2010, Toulouse, France.*

Leibovici, D.G., Pourabdollah, A. (2010b), "Interim Report on Multiscale and Multisource Modelling". FP7 European EuroGEOSS project, deliverable D2.4.1, 41p.

Leibovici, D.G., Pourabdollah, A., Jackson, M. (2011), "Meta-propagation of Uncertainties for Scientific Workflow Management in Interoperable Spatial Data Infrastructures". *In:* Proceedings of the European Geosciences Union (*EGU2011)*, *Austria June 2011.*

Pourabdollah, A., Leibovici, D.G., Jackson, M. (2011), "MetaPunT: an Open Source tool for Meta-Propagation of uncerTainties in Geospatial Processing". *In: Proceedings of OSGIS2011, Nottingham June 2011.*

Saltelli, A., Ratto, M., Terry, A., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S. (2008), *Global Sensitivity Analysis. The Primer.* John Wiley & sons, Chichester, UK, 292p.

Sklar, A. (1959), *"Fonctions de répartition à n dimensions et leurs marges", Publ. Inst. Statist. Univ. Paris, 8*: 229–231

Zargar, A., Devillers, R. (2009), "An Operation-Based Communication of Spatial Data Quality". *In: Advanced Geographic Information Systems & Web Services, International Conference on, IEEE Computer Society, Los Alamitos, CA, USA, pp. 140-145.*

Model Parameter Uncertainty Assessment in the Land Transformation Model

Amir Hossein Tayyebi[1], Saeid Homayouni[1], Jie Shan[2], Mohammad Javad Yazdanpanah[3], Bryan Christopher Pijanowski[4] & Amin Tayyebi[4]

[1] Department of Surveying and Geomatics Engineering, School of Engineering, University of Tehran, Iran
amirhossein.tayyebi@gmail.com, saeid.homayouni@gmail.com
[2] Purdue University, School of Civil Engineering, Department of Geomatics Engineering, West Lafayette, IN 47907
jshan@purdue.edu
[3] Control & Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University of Tehran, P.O. Box 14395/515, Tehran, Iran
yazdan@ut.ac.ir
[4] Purdue University, College of Agriculture, Department of Forestry and Natural Resources, West Lafayette, IN 47907
bpijanow@purdue.edu, amin.tayyebi@gmail.com

Abstract

The Land Transformation Model (LTM) has the ability to simulate, forward and backward, urban land use change patterns in a GIS environment. LTM possesses uncertainty because it approximates reality as do any other model. Model parameter uncertainty is derived from the impacts of model on the outcome of urban land use simulation. This study examines how model parameter uncertainty from the LTM propagates through urban land use change simulation. The simulated urban expansion of 100, 1K, 10K and 100K cycles of the LTM were compared to an observed map of Muskegon River Watershed (MRW) to assess model parameter uncertainties. We proposed a metric here to compare the simulated error maps with each other and evaluate the impact of model uncertainty parameters on the outcome of the LTM. Identification of model uncertainty is crucial for the utilization of the simulation results of the LTM. Our study can also help urban planners to understand the implications of LTM better.

Keywords: Land Transformation Model, Error and Uncertainty, Assessment, Geospatial Information System, Model Parameter Uncertainty

1 Introduction

Errors and uncertainty are common in the land change modeling community (Jantz and Goetz, 2005; Burnickia *et al.*, 2006 and 2010; Pijanowski *et al.* 2006 and 2009; Pontius *et al.* 2008; Pontius and Li, 2010; Pontius and Millones, 2011; Tayyebi *et al.*, 2009 and 2010). Uncertainty has also been considered in other sciences like developing model-based decision support activities for policy analysis (Walker *et al.*, 2001, 2003 and 2010), hydrologic rainfall-runoff predictions and water resource management system

(Ajami *et al.*, 2006 and 2008). Walker *et al.* (2003) identified four classes of uncertainty that may rise from using spatially explicit models in decision making. These include uncertainties in: (1) model inputs due to errors in spatial maps (data uncertainty); (2) assumptions made in the way the model is built (model structure uncertainty); (3) spatial parameter estimates used in a model (model parameter uncertainty); and (4) applying model output to decision making (model outcome uncertainty). In this paper, we used the best available digital maps and assume the data are correct, as is frequently done, in spite of the fact that we suspect that all data possess errors.

The information generated here can serve as a basis for developing possible growth scenarios which are essential for sustainable urban planning and development. From a sustainable spatial planning of view, the present study is particularly important because the spatial characteristics of urban change are useful for understanding various impacts of human activity on the urban environment. The results of the present study can be utilized to develop urban growth scenarios for forecasting possible future changes for sustainable urban land use planning. Results from modeling urban growth can be used by the public, land use planners, and policy makers to anticipate and plan for the future. This paper attempts to evaluate the influences of model parameter uncertainty of LTM in urban land use simulation by conducting several experiments. This study was conducted in the Muskegon River Watershed (MRW) located in the west-central part of the Lower Peninsula of Michigan, USA between 1978-1998.

2 Uncertainty dimension in LTM

2.1 Data uncertainty is associated with data that describe the predictor variables that have an influence on the LTM and its performance (Walker *et al.* 2003). Uncertainty about the predictor variable values that produce change within the LTM are not under the control of the decision makers are very important to decision making analyses, especially if they affect the outcomes of LTM. There is not only great uncertainty in the predictor variable values; there is also uncertainty in the LTM response to these values.

2.2 Model parameter uncertainty is associated with data quality but also is a consequence of the assumptions made about model structure (Walker *et al.* 2003). LTM calibrated parameters are internal weights and biases that must be determined by calibration, which is performed by comparison of model outcomes with historical data series regarding both input and outcome. In order to investigate model parameter uncertainty in LTM, the LTM was run and weights saved for 100, 1K, 10K and100K cycles to create four different urban land change simulations.

2.3 Model outcome uncertainty is the accumulated uncertainty created by the uncertainties in all of the above criteria including model structure, data and model parameter uncertainty (Walker *et al.* 2003). They are propagated through the LTM and are reflected in the resulting estimates of the outcomes. A contingency table of errors and correct predictions are used to quantify location and quantity errors per simulation following Fielding and Bell (1997). Percent Correct Match (PCM) is used as goodness-of-metric to evaluate the agreement between simulated and actual maps in the same time.

3. Implementation of LTM

We considered five independent variables as input in 1978 affect urban growth in MRW: elevation, slope and distance to urban, distance to stream, distance to roads. The cells corresponding to the urban, water and protected legally by government such as federal lands in 1978 are not candidates for transition as new urban areas in 1998 and are thus held out of the training run. The LTM was then trained with 30% of data on the inputs and output for 100K training cycles. Weights were saved after 100, 1K, 10K and 100K cycles and urban cells were then simulated using those network files.

4. Results and discussions

We followed Mean Square Errors (MSEs) across training cycles experiment (Pijanowski *et al.* 2002 and 2009; Washington *et al.* 2010 and Tayyebi *et al.* 2011). Figure 1 illustrates a training run of the LTM in MRW with MSE plotted across training cycles. We halted the training at 100K cycles where the MSE was 0.087.

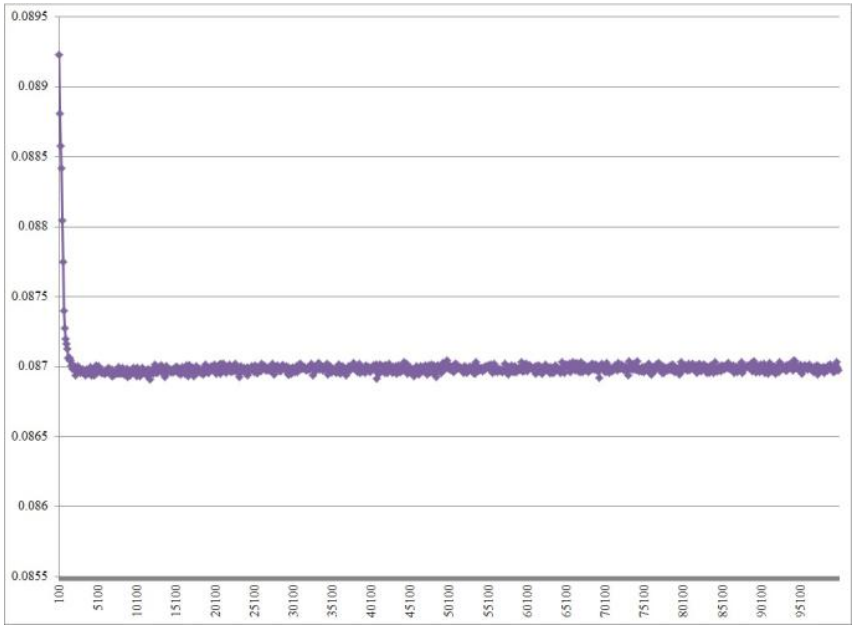


Figure 1. MSEs across 100k training cycles

The pattern file and the 100, 1K, 10K and 100K network files used to generate the probability maps in 1998 based on five independent input variables in 1978. Most of the cells with high probability values in simulated maps follow urban and roads in 1978. ArcGIS10.0 determined 854,418 cells transitioned into urban between 1978-1998. Thus, 854,418 cells were selected from the each probability maps that had the greatest change likelihood values; these cells were then classification as new urban for 1998 (Pijanowski *et al.*, 2009). Cells that were simulated to transition to new urban areas for 1978-1998 were compared with the cells that actually did transition during the time period of study (Figure 2).

The PCM metric indicates that four maps have accuracy over 80% and could accurately simulate urban land use change in MRW. The resulting display (Figure 2) used throughout the model building process to visualize the spatial distribution of the TP, TN, FP and FN (where: no real change and no predicted change (code 1) = True Negative; no real change but change predicted by the model (code 2) = False Negative; real change but

not predicted by the model (code 3) = False Positive and real change and predicted change (code 4) = True Positive).

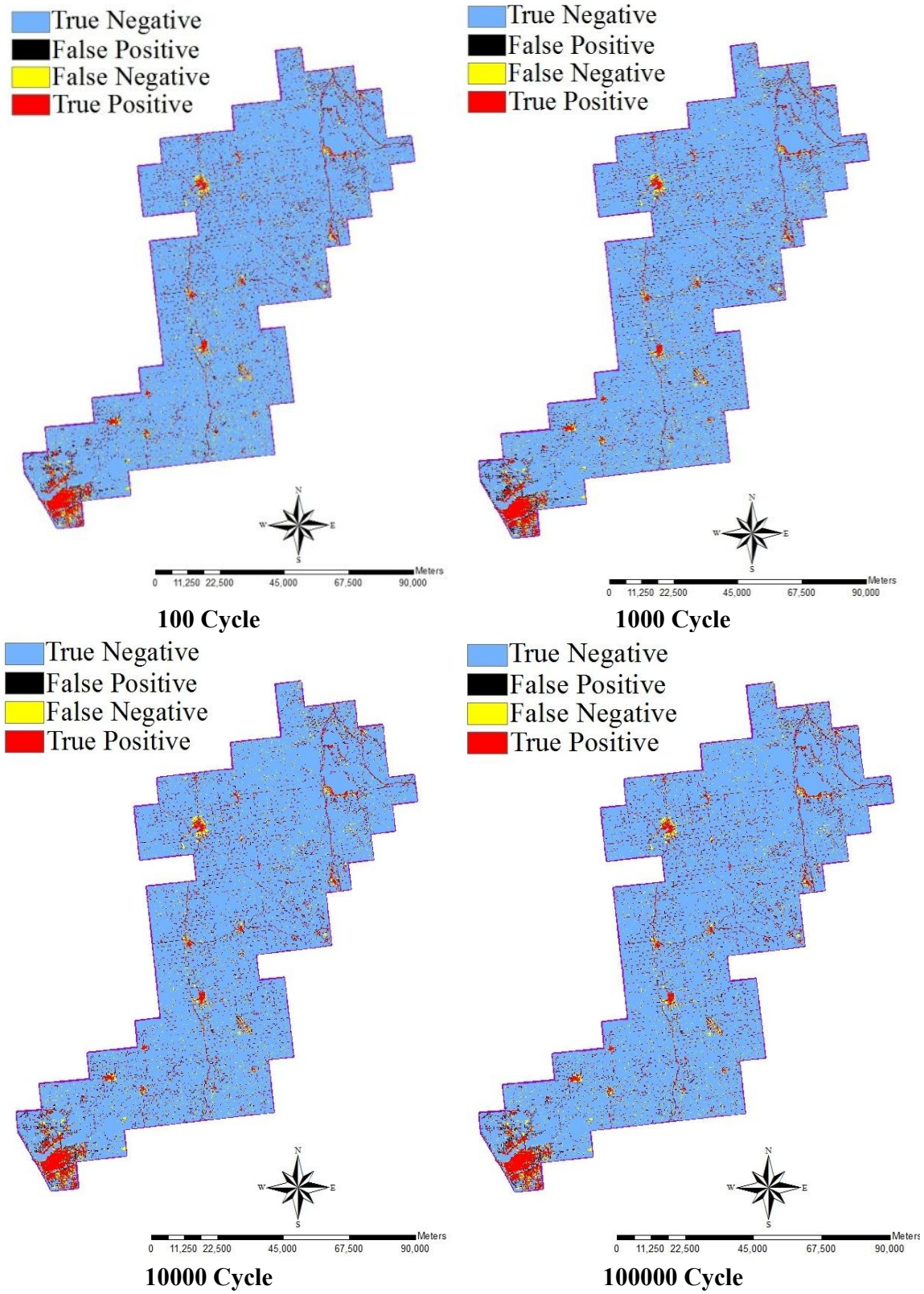


Figure 2. Spatial distribution of FP, FN, TP and TN for 100, 1K, 10K and 100K cycles

We also multiplied error maps (which contains values from 1 to 4) resulted from cycle to create a coding system in the GIS as follows:100,000 cycles by 1000, 10,000 cycles by 100, 1000 cycles by 10 and 100 cycles by 1 and then summed these maps. The new map contained 32 unique values indicating spatial and quantity distribution of agreement and

disagreement of location error across different cycles (Table 1). Then, we divided the total number of cells that are in agreement (Codes 1111, 2222, 3333 and 4444 in Table 1) and disagreement (all other codes in Table 1) by the total number of cells in one of the maps, respectively. With assuming that the data are free from error in this research, the final result of simulations from different cycles suggest that model parameter uncertainty cause less than 2% disagreement in urban simulation of LTM.

Table 1. Spatial and quantity location of errors in error maps from 4 different cycles (1 = TN; 1 = FN; 2 = FP and 3 = TP)

Value	Count	Value	Count	Value	Count	Value	Count
1111	10484058	2111	1794	3333	336781	4333	1423
1112	27266	2112	2717	3334	15554	4334	1782
1121	10314	2121	1844	3343	6799	4343	2131
1122	14007	2122	12951	3344	9418	4344	8826
1211	13758	2211	14973	3433	10015	4433	11249
1212	2585	2212	6484	3434	1998	4434	5228
1221	2083	2221	21928	3443	1360	4443	17155
1222	929	2222	319847	3444	740	4444	620044

5 Conclusion

The simulation of land use change models is affected by a variety of source errors such as uncertainty in data, model structure, model parameter and model outcomes. It is vital to measure whether the maps of various scenarios of future land change are meaningfully different, because differences among such maps serve to inform urban planners. The next step of this research is to define a framework to investigate data uncertainty, model parameter uncertainty and outcome uncertainty in urban land use change simulation simultaneously. This framework will help scientists to decide what the most important source of errors is in the final product. If decision-makers use predicted urban land use maps, researchers need to understand the sources of uncertainty in their models. The results suggest that (1) the neural net learns well early but can't improve over large number of cycles and (3) thus, we are left with having to improve the other aspects of the model, including better data, better structure of the model to represent the complex process.

The LTM creates a lot of network files that it would be confusing to understand which of these files would lead to best urban expansion simulations. Addressing these questions would also be very helpful to know that how we can improve the accuracy of LTM in our next simulations: (1) MSE is a quantitative metric and it does not give any information about model goodness of fit relative to location. MSE appears to be an appropriate criteria to stop training run in the LTM; (2) If MSE is appropriate, how many cycles are necessary to continue training run; and (3) Which of these cycles would lead to best PCM value of LTM for urban growth simulation.

References

Ajami, N. K., Q. Duan, and S. Sorooshian, (2006). "An integrated hydrologic Bayesian multi-model combination framework: confronting input, parameter, and model structural uncertainty in hydrologic prediction", water resources research, 43 (1), 1-19.

Ajami, N. K., G. M. Hornberger, and d. L. Sunding, (2008). "Sustainable water resource management under hydrological uncertainty", water resources research, 44 (11), 1-10.

Burnickia, A. C., Brown, D. G., and Goovaertsc, P. (2006). "Simulating error propagation in land-cover change analysis: The implications of temporal dependence", Computers, Environment and Urban Systems.

Burnickia, A. C., Brown, D. G., and Goovaertsc, P. (2010). "Propagating error in land-cover-change analyses: impact of temporal dependence under increased thematic complexity", International Journal of Geographical Information Science, Vol. 24, No. 7, 1043-1060.

Jantz, C. A., and Goetz, S. J. (2005), "Analysis of scale dependencies in an urban land-use-change model", international journal of geographical information science, vol. 19, no. 2, 217–241.

Kwakkel, J. H., Walker, W. E. and Vincent A. W., J. Marchau, (2010). "Classifying and communicating uncertainties in model-based policy analysis", int. J. Technology, policy and management, Vol. 10, No. 4.

Pijanowski, B. C., D. G. Brown, B. A. Shellito, G. A. Manik, (2002). "Using neural networks and GIS to forecast land use changes: a land transformation model". Computers environment and urban system. 26 (6), 553-575.

Pijanowski, B. C., K. Alexandridis and D. Mueller. (2006). "Modeling urbanization patterns in two diverse regions of the world". Journal of land use science. (1): 83-108.

Pijanowski, B. C., Tayyebi, A., Delavar, M. R., and Yazdanpanah, M. J. (2009). "Urban expansion simulation using geographic information systems and artificial neural networks". International Journal of Environmental Research, 3(4), 493-502.

Pontius, R. G., Jr. and X. Li, (2010), "Land transition estimates from erroneous maps". Journal of land use science 5(1): 31-44.

Pontius R. G. Jr, and Millones, M. (2011). "Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment". International journal of remote sensing, in press.

Tayyebi, A., Delavar, M. R., Pijanowski, B. C., and Yazdanpanah, M. J. (2009). "Accuracy assessment in urban expansion model". In R. Devillers, H. Goodchild, Taylor., & Francis. (Eds.), Spatial data quality, from process to decisions, 107-115. Canada: CRC Press.

Tayyebi, A., Delavar, M. R., Pijanowski, B. C., Yazdanpanah, M. J., Saeedi, S., and Tayyebi, A. H. (2010). "A spatial logistic regression model for simulating land use patterns, a case study of the shiraz metropolitan area of Iran". In E. Chuvieco, J. Li, & X. Yang (Eds.), Advances in earth observation of global change. Springer Press.

Tayyebi, A., Pijanowski, B. C., A. H. Tayyebi, (2011). "An urban growth boundary model using neural networks, GIS and radial parameterization: An application to Tehran, Iran". Landscape and Urban Planning 100, 35-44.

Walker, W. E., P. Harremoes, J. Rotmans, J. P. Van Der Sluijs, M. B. A. Van Asselt, P. Janssen, and M. P. Krayer Von Krauss, (2003), "Defining uncertainty a conceptual basis for uncertainty management in model-based decision support". Integrated assessment, 4 (1), 5-17.

Yeh, A. G., X. Li, (2006). "Errors and uncertainties in urban cellular automata". Computers, environment and urban systems, 30, 10–28.

A model to estimate length measurements uncertainty in vector databases

Jean-François Girres

Laboratoire COGIT, Institut Géographique National, 73 avenue de Paris, 94165 Saint-Mandé, France
jean-francois.girres@ign.fr

Abstract

The evaluation of geometric accuracy on vector databases mainly focuses on the production of positional accuracy indicators. In order to integrate more external quality, indicators of length and area measurement uncertainty could be provided to the final user or decision maker. Indeed, many applications are based on measurements performed on the geometry of geographical objects, and positional indicators not allow to quantify measurement uncertainty. In this context, the paper presents an original approach, integrating production processes and representation rules of vector databases, to quantify their respective impact on length measurement. Methods to compute these impacts are integrated in a general model, elaborated to allow a user to estimate geometric measurements uncertainty on vector objects. Experimentation is realised on a road network sample, and further investigations are evocated, like the computation of a global indicator.

Keywords: geometric measurement uncertainty, evaluation model, vector data, production processes, representation rules.

1 Introduction

Over the past years, the context of geographic databases production and usages has considerably changed. The number of databases producers has increased, involving the development of data exchanges, and geographic information applications have also diversified. In this context, providing spatial data quality information is fundamental to avoid users' misuses. To evaluate the quality of vector databases, several elements are provided (van Oort, 2005). If these elements are relevant, they are generally more oriented on internal quality (i.e. the ability to satisfy the specifications), than on external quality (i.e. fitness for use). More specifically, indicators provided to assess geometric accuracy only focus on positional accuracy, but not on the impact of positional error on measurements (length or area) performed with the geometry of vector objects. In the same time, more and more applications are based on length or area measurements, e.g. a travel time by car or the computation of a density in an administrative entity, which requires the area. In order to integrate external quality in spatial data quality evaluation methods, the production of length and area measurement uncertainty indicators looks particularly relevant for the final user.

Uncertainty in length and area computation has already been studied by several researchers. For instance, (Chrisman and Yandell, 1988) proposed a statistical model to estimate the error in area computation. (Griffith, 1989) modeled the impact of digitization error on distance and area computation, and (Leung *et al.*, 2004) developed a general framework for error-analysis in measurement-based GIS. Besides these contributions, it is still very complex to communicate indicators of uncertainty on length and area computation to the final user. Indeed, the geometry of geographical objects is affected by different causes which impact these computations. We assume that these impacts are not homogeneous in the entire dataset and that the use of positional indicators (the Root Mean Square Error, for instance) does not allow to estimate measurement uncertainty.

Thus, to be able to produce indicators of uncertainty on length or area computation, all the causes of measurement error need to be understood and modeled, in order to compute their respective contribution in the final measurement error. Our assumption is that measurements performed with the geometry of vector objects are affected by two main causes: the production processes and the representation rules of the database. By integrating knowledge on these causes of errors affecting the database, the objective of this research is to build a model to allow a user to estimate measurement uncertainty using the geometries of vector objects.

This paper presents in the next section the origins of length measurement error and the strategy used to model their respective impacts. Then, the elaboration of the model is exposed (section 3), followed by results of experimentation on a sample of road objects (section 4) before concluding and evocating research perspectives.

2 Origins of length measurement error

Length measurements performed using geometries of linear vector objects are affected by two causes: the production processes and the representation rules of the database. This section exposes the strategy used to model their respective impact.

2.1 Production processes and data sources

Digitizing error is the first production process impact to consider in the estimation of length measurement uncertainty. Digitizing error can be defined as the positional error on the vertices of the geometry involved by the operator during the construction of objects. This is a human and accidental error, modeled statistically by a normal law. It is commonly accepted that the digitizing precision is about 0.1 mm according to the production scale. The propagation of this error on the geometry generates uncertainty on the length computed (Figure 1a). It is modeled with the standard deviation *σ(e)* presented in "(1)", as proposed in (Girres and Julien, 2010).

$$\sigma(e) = \sqrt{1 + 2 \sum_{2 \le i \le n-1} \sin^2 \frac{\theta_i - \theta_{i-1}}{2}} * \varepsilon_q \qquad (1)$$

where $\theta_i - \theta_{i-1}$ is the angle between consecutive segments of the polygonal line, and εq is the digitizing precision on the coordinates of the vertices.

Using the properties of the normal law, the error e on the total length of the polygonal line ranges from $-3\sigma(e)$ to $3\sigma(e)$.

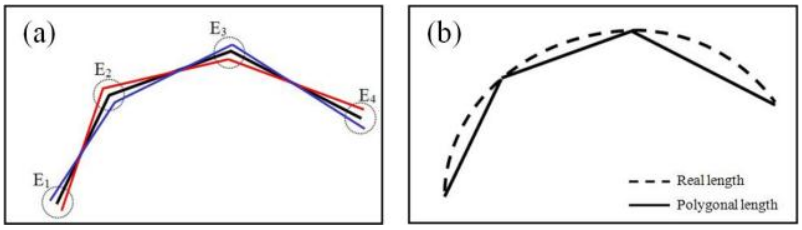


Figure 1. (a) Uncertainty in length computation involved by digitizing error on the vertices of a polyline (b) Polygonal approximation of a curve by the geometry of a polyline

Polygonal approximation of curves is the second production process impacting length measurement to consider. As exposed in Figure 1b, the representation of a curve by a polygonal line generates underestimation on the computed length. If this impact of polygonal approximation is almost never compensated, several algorithms can be applied to estimate it, using spline interpolations. The impact of acquisition imprecision performed with Global Navigation Satellite System (GNSS) can also be responsible of errors in the construction of geographical objects geometries, and in measurements performed subsequently. GPS errors can be explained by several causes (shifts in the satellite orbits, tropospheric effects, environmental disturbances, etc) and its impact can be modeled using the method proposed by (Bogaert *et al.,* 2005) who developed a framework for assessing the error of polygonal area measurements in agriculture applications. The effects of cartographic generalisation need also to be considered, if the database is captured from maps. In this case, geometries of vector objects are impacted by modifications of the shape and/or position. We assume that these effects vary according to three parameters: the type of object (roads, railways, etc), the scale of representation of the source map (which influences the level of generalisation) and the spatial context of objects. Indeed, generalisation effects are not the same if an object is located in urban, mountainous or rural area (Touya *et al.*, 2010). To estimate the impact of cartographic generalisation on length computation, a statistical function is developed, as detailed in (Girres, 2011), using prior length comparisons performed between generalised datasets and a reference one, representing the same road network.

2.2 Representation rules

Map projection is the first representation rule impacting length computation. Indeed, because a map projection is a systematic representation of the surface of a round body on a plane (Snyder, 1987), it cannot be done without distortions which vary according to the projection used. To estimate this impact, the computation of length on the reference ellipsoid is performed using distortion grids. It is exposed in "(2)", assuming that the scale factor *Em* is given by the distortion grid.

$$Lref = Lproj/(1+Em) \quad (2)$$

where *Lref* is the length on the reference ellipsoid, *Lproj* is the projected length and *Em* is the average scale factor of the polyline.

The second representation rule deals with the consideration of the terrain. Indeed, computation of lengths in two dimensions systematically generates underestimations. The use of digital terrain model (DTM) to affect altitudes on vector objects geometries gives the possibility to compute 2D5 lengths, and estimate the impact of the terrain on measurements. If the use of DTM can provide a more realistic approximation of the real lengths, it can be limited by the precision of the DTM and the resolution of geographical objects.

3 A model to estimate length measurement uncertainty

The origins of length measurement uncertainty presented in section 2 (production processes and representation rules) are integrated in a general model elaborated to estimate their respective impact in the final measurement error. This model (Figure 2) is designed as a user-oriented model. To facilitate the estimation of length measurement uncertainty, the first step consists in integrating user's knowledge on the dataset (scale, production processes, etc.) with a user interface or metadata files, in order to activate or not the estimation functions and to parameterize them.

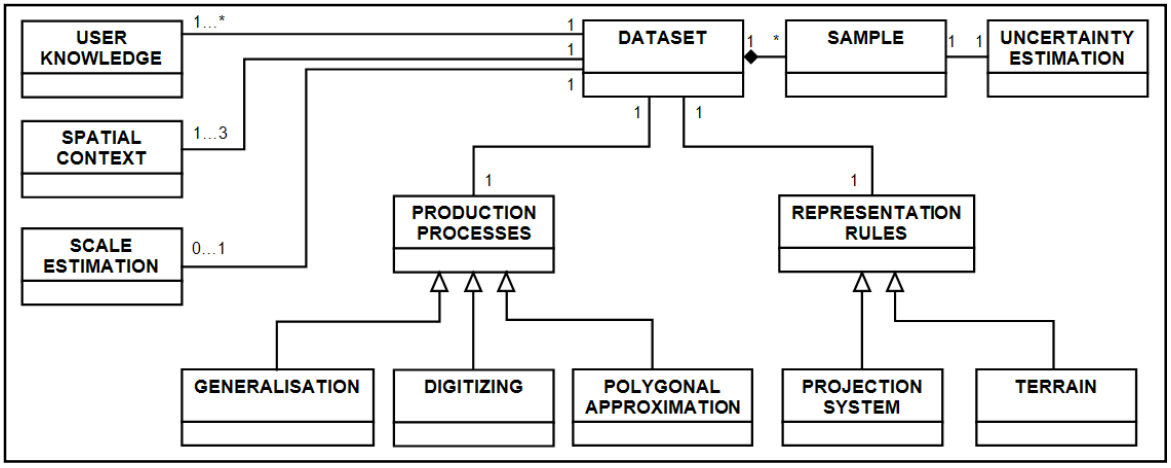


Figure 2. UML Class diagram of the estimation model

The second step consists in quantifying the respective impacts of selected production processes and representation rules on length measurement. When the parameters necessary for a function are unknown (e.g. the capture scale), original methods, presented in (Girres, 2011), are used to estimate them. Each impact is actually computed separately in order to estimate its contribution. The ultimate step consists in computing a global indicator of length measurement uncertainty. The methodology for the elaboration of this global indicator is actually investigated. Indeed, a simple addition of the impacts does not constitute a realistic solution. Thus, different methods need to be tested and validated, in order to define the most realistic way to quantify length measurement uncertainty.

4 Experimentation

To illustrate the functioning of the model by estimating respective impacts of production processes and representation rules on length measurement, experimentation is proposed on a road network located in the French department of *Pyrennées-Atlantiques*. Three road samples are extracted from the TOP100® database produced by IGN, the French National Mapping Agency. Each sample is located in a different spatial context, as exposed in Figure 3.

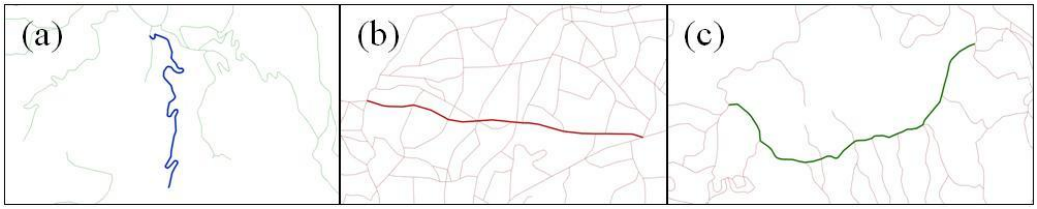


Figure 3. Samples of road network located in mountainous (a), urban (b) and rural area (c)

To assess representation rules impacts on length measurements, estimation of the effect of the projection system is realised with the Lambert 93 distortion grid, in order to compute lengths on the reference ellipsoid (IAG GRS80), using "(2)". To evaluate terrain consideration, the BDALTI® DTM is used to affect altitudes on road samples geometries and perform computations of 2D5 curvilinear abscissa.

To estimate the impacts of production processes, user's knowledge information indicates that the database is digitized from 1:100 000 scale maps, is impacted by cartographic generalisation, and represents curve objects. Estimation of the impact of digitizing error is computed using "(1)". As assumed previously, the digitizing precision is evaluated as 0.1 mm according to the scale map, meaning that ε_q is equal to 10 meters. The impact of cartographic generalisation is evaluated using the statistical function elaborated in (Girres, 2011) based on prior length comparisons. This function allows to compute an average length error on each road sample, according to the spatial context and the scale. External datasets (BDALTI® and Corine LandCover classifications) are used to delineate the three spatial contexts. Finally, the assessment of the impact of polygonal approximation of curves is performed using cubic spline interpolations. Different interpolation methods are tested, and the Kochanek-Bartels curve (Kochanek and Bartels, 1984) proved to be the most realistic by visual validation, using the following parameters: tension = 0.5, bias = 0 and continuity = 0. Experimentation results are exposed in Table 2.

Table 2. Estimated length errors computed for each impact.

Impacts	**Road samples**		
	Mountain	**Urban**	**Rural**
Computed 2D Length	4.47 km	3.50 km	7.66 km
Reference 2D5 Length	4.77 km	3.52 km	7.91 km
Digitizing Error	+/- 95 m	+/- 48 m	+/- 57 m
Polygonal Approximation	17 m	1 m	10 m
Cartographic Generalisation	245 m	105 m	291 m
Projection System	- 3.7 m	- 1.5 m	- 4.1 m
Terrain	77 m	2 m	11 m

Computed 2D lengths are compared with reference 2D5 lengths extracted from the BDTOPO®, a French topographic database (of metric resolution). Results of the experimentation primarily show that the three lengths computed are underestimated in comparison with them homologous references. The sample located in mountainous area is much more affected by terrain inconsideration (77 m) than the two other samples, but it does not constitute the main reason of length error. For the three samples, main impacts of length uncertainty are involved by digitizing error (using a digitizing precision equals to 10 m) and especially cartographic generalisation, as shown by the average length error estimated by application of the statistical function based on prior length comparisons. Impacts of projection system and polygonal approximation of curves are relatively residual in this experimentation report.

These results already give an idea of the main causes of length error affecting a generalised road network in different spatial contexts. The ultimate step of the elaboration of the model deals with the production of a global indicator of length measurement uncertainty, which takes into account the different impacts presented in this experimentation.

5 Conclusion

This paper presents the elaboration of a general model, developed in order to allow a user to estimate uncertainty in length measurement computed with the geometries of vector objects. The originality of this model is that it takes into account the origins of measurement uncertainty affecting geographic objects, i.e. production processes and representation rules of the database. It constitutes an experimental study on the consideration of external quality in the production of indicators of geometric accuracy for vector databases. As shown by the experimentation performed on a sample of three road objects, respective impacts are still computed separately. Perspectives of this research focus on the development of a method to produce a single global measurement uncertainty indicator for length and area computations.

References

Bogaert, P., Delincé, J., Kay, S. (2005) "Assessing the error of polygonal area measurements: a general formulation with applications to agriculture". *Measurement Science and technology*, Vol. 16(5):1170-1178.

Chrisman, N.R., Yandell, B.S. (1988), "Effects of point error on area calculations: a statistical model". *Surveying and Mapping*, Vol. 48:241-246.

Girres, J., Julien, P. (2010), "Estimation of digitizing and polygonal approximation errors in the computation of length in vector databases". *In:* Tate, N.J., Fisher, P.F. (eds.). *Proceedings of the ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences,* Leicester, UK, pp. 373-376.

Girres, J. (2011), "An evaluation of the impact of cartographic generalisation on length measurement computed from linear vector databases". *In:* Ruas, A. (ed.). *Proceeding of the Twenty-fifth International Cartographic Conference,* Paris, France.

Griffith, D.A. (1989), "Distance calculations and errors in geographic databases". *In:* Goodchild, M.F., Gopal, S. (eds.). *The accuracy of spatial databases,* Taylor & Francis, pp. 81-90.

Kochanek, D.H.U., Bartels R.H. (1984), "Interpolating splines with local tension, continuity and bias control". *ACM SIGGRAPH*, Vol. 18(3):33-41

Leung, Y., Ma, J.H., Goodchild, M.F. (2004), "A general framework for error analysis in measurement-based GIS. Part 4: Error analysis in length and area measurements". *Journal of Geographical Systems*, Vol. 6(4):403-428.

Snyder, J.P. (1987), *Map projections – a working manual*, USGS Paper 1935, Washington, USA, 383 p.

Touya, G., Duchêne, C., Ruas, A. (2010), "Collaborative generalisation: Formalisation of generalisation knowledge to orchestrate different cartographic generalisation processes". *In:* Fabrikant, S., Reichenbacher, T., van Kreveld, M., Schlieder C. (eds.), *Geographic Information Science, Lecture Notes in Computer Science.* Springer Berlin / Heidelberg, pp. 264-278.

van Oort, P. (2005), *Spatial data quality: from description to application*. PhD thesis, Wageningen University, The Netherlands.

A test for stationarity of aggregated spatio-temporal point processes

Alfred Stein[1] *& Marie-Colette N.M. van Lieshout*[2]

[1] Faculty of Geo-Information Science and Earth Observation of the University of Twente (ITC), Enschede, The Netherlands
[2] CWI, PO Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract

The goal of this paper is to present a test of spatial homogeneity for a spatial point pattern. This test is based on the well-known KPSS test. The test is applied to a catalogue that contains spatial coordinates of shallow earthquakes of magnitude 4.5 or larger aggregated over calendar years in Pakistan. We test relative temporal stationarity.

Keywords: Stationarity test, space-time point process, earthquakes

1 Introduction

Spatial patterns of points may occur in a multitemporal way when considered in successive periods of time. Such patterns may raise questions of stationarity in space and time, i.e. whether a pattern remains similar throughout the years, maybe except for some smaller deviations throughout the years, or that they change substantially. Typical examples from the natural sciences are the occurrence of disasters like earthquakes or landslides, whereas in socio-economic applications one may think of the distribution pattern of a particular enterprise within a city, or the distribution of settlements within a country. In this study we will focus on the occurrence of earthquakes.

Disasters like earthquakes apparently occur at erratic seismic locations and at unexpected moments. Many applied scientific papers have been published (e.g. Holden *et al*., 2003, Anwar *et al*., 2011). Also, the statistical literature on earthquakes in space and time has a long history. Important contributions were made by Ogata and his co-authors in an outstanding series of papers (e.g. Ogata (1998), Zhuang *et al*. (2002)). Other contributions in the point process literature include (Vere-Jones, 1970; Vere-Jones and Musmeci, 1992). These papers consider space-time data, like earthquakes, but often restricted for island states. In this context, a temporal point process marked by location of occurrence is appropriate and conditional intensity functions given past occurrences can be written down explicitly. These in turn suffice to write down a likelihood function on which inference can be based. Moreover, edge effects are no issue.

The aim of this study is to explore spatial statistical techniques for data aggregated over time for which an explicit likelihood function is not available. Attention focuses on Pakistan, for which country annual patterns of earthquakes have been recorded for more than thirty five years. During this period, two major earthquakes of magnitude larger than seven were recorded: one occurring in 1997 and the major Kashmir earthquake of 2005.

The paper extends previous work in Van Lieshout and Stein (*subm.*) by exploring sensitivity issues for testing of stationarity.

2 Relative spatial rates in time

The first analytical stage is an investigation into the spatial intensity function of events. We will test whether this pattern persists over the years. We write $W \subset R^2$ for the compact set representing the area of interest (often called the window) and consider X_i to be the point process of locations of that occur in W in a range of n time periods. We denote the intensity function of X_i by μ_i. In other words, for every Borel subset A of W, $\int_A \mu_i(x,y)dx\,dy$ is the expected number of points of X_i falling in A. The issue to address is thus whether μ_i is constant over the years, except for a time-dependent scalar multiplication factor.

In order to do so we divide W into two disjoint areas A_1 and A_2 of equal pooled integrated intensity. For testing procedures we introduce the ratio between the number of events in the first area to that in the second area as a function of time, i.e. random variables $Y_i = N(X_i \cap A_1)/ N(X_i \cap W)$, where $N(X_i \cap A)$ is the number of points of $X_i \cap A$ for every Borel set $A \subset W$ and i = 1,…, n.

To test the null hypothesis H_0 that $(Y_i)_{i\in\{1,\ldots n\}}$ is stationary, we develop a test based on earlier work by Kwiatkowski *et al*. (1992). This test is known by the acronym KPSS referring to the first characters of the authors' surnames. Given a time series of length n we define the partial sums process $S_n(i) = \Sigma_{j=1}^{i} (Y_j - m_Y)$ for $i \in \{1, \ldots, n\}$, where m_Y is the sample mean $1/n\ \Sigma_{i=1}^{n}\ Y_i$. Under the null hypothesis, the limiting value $\tau^2 = \lim_{n\to\infty} n\ \mathrm{Var}\ m_Y = \Sigma_{j=1}^{\infty} c_j$ is well-defined provided that the autocovariances $c_j = \mathrm{Cov}(Y_1, Y_{1+j})$ at lag j exist and their sum is absolutely convergent. Under mild regularity conditions, $\tau^{-2}\ n^{-2}\ \Sigma_{i=1}^{n}\ S_n(i)^2$ converges in distribution to the integral of the squared Brownian bridge $\int_0^1 V(t)^2\,dt$ (Herrndorf, 1984). As τ^2 is unknown, we set $s_n^2 = \gamma_0 + 2\Sigma_{j=1}^{\ell}(1-j/n)(1-j/(\ell+1))\gamma_j$, where $\gamma_j = (n-j)^{-1}\ \Sigma_{i=1}^{n-j}\ (Y_i - m_Y)\ (Y_{i+j} - m_Y)$ are the sample autocorrelations and ℓ defines the maximal temporal lag taken into consideration. The weights $1 - j/(\ell + 1)$ were shown by Newey and West (1987) to lead to a non-negative estimator. The authors also proved weak consistency, whereas strong consistency was proved by De Jong (2000). In summary, the KPSS test statistic is given by $T_n = 1/(n^2\ s_n^2)\ \Sigma_{i=1}^{n}\ S_n(i)^2$ which is asymptotically distributed as the integral of the squared Brownian bridge. Critical values of the test are reported in Table 1 of Kwiatkowski *et al*. (1992).

3 Application: the long-term earthquake pattern in Pakistan

Earthquakes are serious environmental disasters. Having a better knowledge on where the earthquakes, e.g. as major events or as aftershocks occur in relation to geological features, may result in identification of hazard zones. Modelling of earthquake data has since long been a focus of research by seismologists and statisticians. Stochastic geometry offers various tools and procedures to contribute to a better understanding by means of spatial testing, spatial modelling and mapping. In that sense, data collected routinely in public databases may reveal patterns that are otherwise unknown.

An earthquake describes both a sudden slip on a fault, and the resulting ground shaking and radiated seismic energy caused by the slip. Deeper causes are volcanic or magmatic activity, or other sudden stress changes in the earth. The release of energy at an

unanticipated moment is registered as the main shock. Main shocks are usually followed by aftershocks that are smaller than the main shock and within 1-2 rupture lengths distance from the main shock. Aftershocks can continue over a period of weeks, months, or years. In general, the larger the main shock, the larger and more numerous the aftershocks, and the longer they will continue.

3.1 Data description

Pakistan is a country that is regularly affected by earthquakes (Fig. 1). The reason for the vulnerability of the country to earthquakes is the subduction of the Indo-Australian continental plate under the Eurasian plate with its two associated convergence zones.

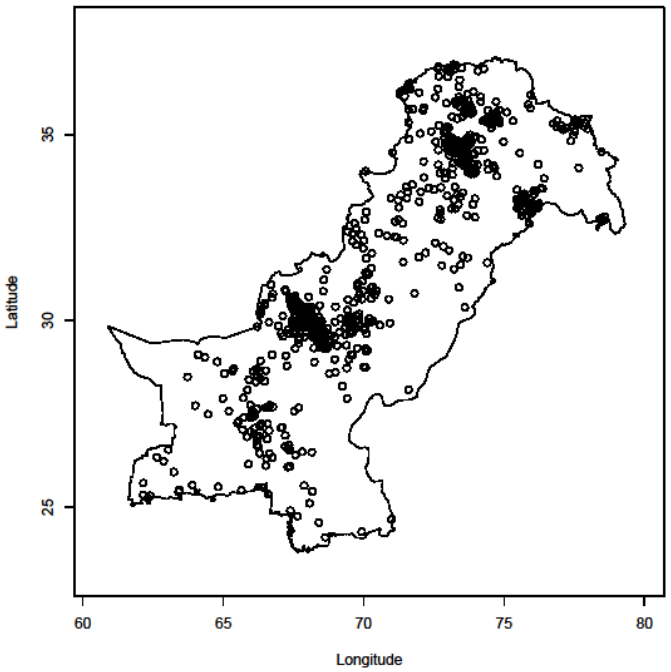


Figure 1. Shallow earthquakes of magnitude 4.5 or higher within the Pakistan territory during the years 1973-2008.

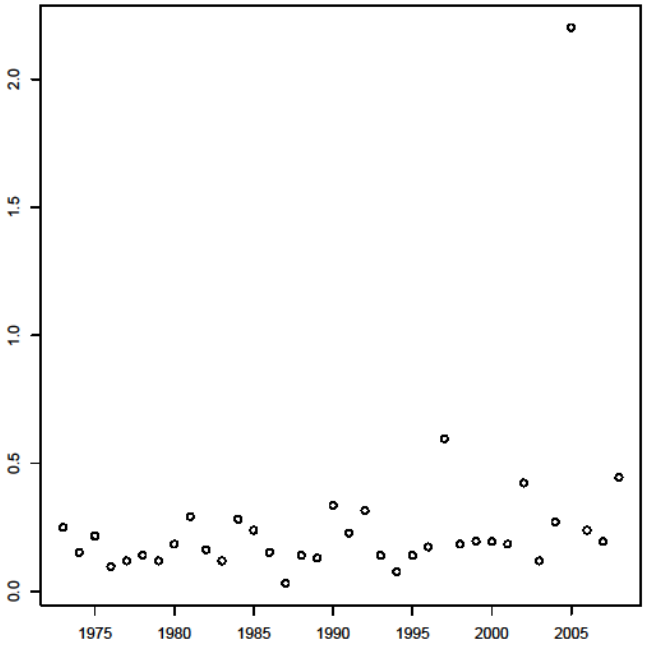


Figure 2. The annual number of shallow earthquakes of magnitude 4.5 or higher per square degree latitude-longitude recorded in Pakistan during the years 1973 - 2008.

Two major earthquakes were recorded in 1997 and 2005 with magnitudes of 7.3 and 7.6 respectively. The 1997 earthquake occurred along the convergence zone running from the South-West to the North-East and resulted in about seventy casualties. The 2005 Kashmir earthquake was devastating with at least 86,000 casualties with a magnitude as at least 7.6 with its epicenter about 19 km north-east of the city of Muzaffarabad. Such big earthquakes are accompanied by many aftershocks. A total of 147 aftershocks were registered in the first day after the initial quake. On October 19, a series of strong aftershocks occurred about 65 km north-northwest of Muzaffarabad.

In addition to such major shocks, that are still relatively rare, many smaller shocks have been recorded (see http://earthquake.usgs.gov/earthquakes for a list of earthquakes since 1973). The majority of tectonic earthquakes originate at depths not exceeding tens of kilometres. Those occurring at a depth of less than 70 km are classified as `shallow'. Earthquakes that originate below this upper crust are classified as `intermediate' or `deep' (Molnar and Chen 1982). Clearly, the impact of an earthquake depends on its epicentre, its depth as well as its magnitude. Minor earthquakes occur very frequently and may not

even be noticed or recorded. Therefore we focus on those having a magnitude of at least 4.5 for which records are believed to be exhaustive. Such earthquakes are well felt, although earthquakes of a magnitude of above 6 become destructive.

Our data consist of the annual patterns of shallow earthquakes of magnitude 4.5 or higher in Pakistan during the period 1973--2008. The Pakistan territory defines the compact set $W \subset R^2$ and X_i the point process of locations of shallow earthquakes of magnitude at least 4.5 that occur in W in year i=1973, …, 2008, hence n being equal to 36. Location and magnitude of each earthquake is recorded. For the major earthquake years 1997 and 2005, also the times at which shocks occurred in the month following the main one are available.

The annual number of such earthquakes per square degree latitude-longitude is given in Figure 2. Note that the clearly visible outlier corresponds to the Kashmir earthquake in 2005 that generated a large number of aftershocks. The number of aftershocks in 1997 was considerably less and more diffuse. In accordance with the Gutenberg-Richter power law, we fit a shifted exponential probability density $\beta\, e^{-\beta\,(m-4.5)}$ for $m \geq 4.5$ and 0 elsewhere. The maximum likelihood estimator is $\beta^* = m_{875} = 2.82$, where m_{875} is the sample mean over the set of 875 pooled magnitudes.

3.2 Spatial intensity

The first analysis concerns the spatial intensity function of the earthquake events. Both visual and geological evidence suggest enhanced earthquake intensity in the northern and mid-western parts of the country. To test whether this pattern persists over the years we first consider the spatial intensity function $\rho(x,y)$, with $(x,y) \in W$. To avoid edge effects, earthquake locations in Pakistan and in neighbouring countries within a distance of about 1^o from the Pakistan border are aggregated into a single pattern.

We then exclude the major earthquake years 1997 and 2005, pool the remaining thirty four years together and calculate the kernel estimator of intensity using an isotropic Gaussian kernel with standard deviation 0.5. High intensity occurs in the north of the country, near the junction of plate boundaries, and in a smaller region in the east. A second zone of high earthquake activity lies in the mid-west of the country. In fact, the epicentre of the 1997 earthquake is located in this area.

To test for stationarity, we divide the country into two parts: A_N is the subset of W lying north of the 31.4^o latitude line, and $A_S = W \setminus A_N$ and we let Y_i be the fraction of shocks above 31.4^o latitude in year i. for i = 1973,…,2008 and applied the testing taking ℓ=4. The test statistic takes the value 0.1297 with a p-value exceeding 10%. When repeating with a value of $\ell = 1$ the test statistic increased to 0.4633, with a corresponding p-value equal to 0.04994. We conclude that there is no statistical evidence of a temporal trend in intensity patterns of shallow earthquakes based on the north-south divide during the 36 years of study.

As a next step, we carried out a sensitivity analysis, as the choice for the dividing value is somewhat arbitrary. Values for the dividing line between 30^o and 34^o degrees are tested (Figs. 3, 4). We note that the test statistic remains stable between 30^o and 33^o, whereas we observe a marked increase in the test value beyond 33^o. This value is apparently the lower boundary of the hotspot area in the north separating it from the less heavily affected area in the south. Based on the interval [30^o, 33^o] we conclude that the test is not particularly sensitive to a specific separating value, whereas moving the separating value to within a clearly different subarea has a more pronounced effect on the test statistic. The corresponding p-values are above 0.10 for the interval between [30^o and 33^o]

whereas they decrease to values below 0.05 when locating the boundary between 33° and 33.8°. This points to lack of spatial stationarity for these higher values, and hence to a changing pattern when a higher boundary is applied.

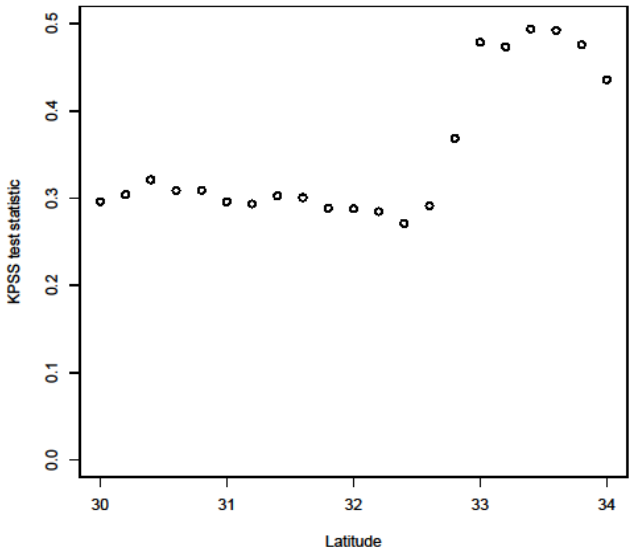


Figure 3a. KPSS test values for different latitudes for boundary between A_N and A_S with $\ell = 4$. The test value remains stable between latitudes 30° and 33° and becomes larger above 33°.

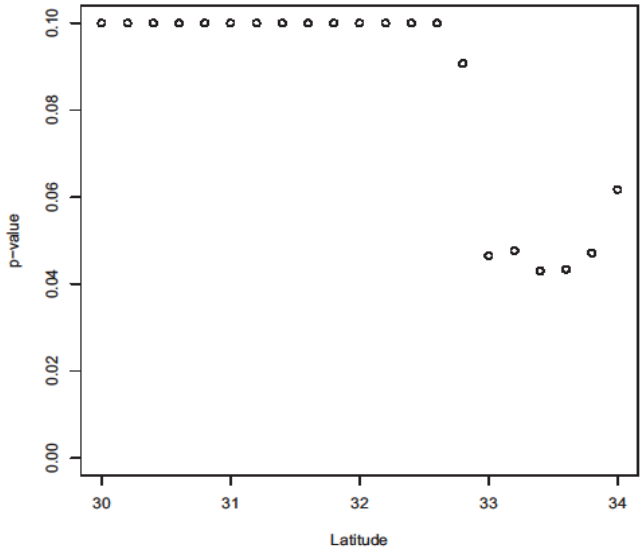


Figure 3b. P-values for the KPSS statistic. Test values are around 0.10 between latitudes 30° and 33° and become smaller above 33°.

As the number of earthquakes can be small within a single year, we finally considered bi-annual data, i.e. by grouping the data from two successive years (Fig. 4). A similar pattern emerges in test values and in p-values, but no significance of p=0.05 is reached. The results thus point to a stationary pattern of earthquakes over the years.

4 Discussion and conclusions

There is a clear benefit in testing for stationarity of a point pattern. The test statistic has a general validity and is applicable to a range of spatial point studies. In the study presented here on the earthquakes, we found it of general interest that the pattern does not show particular changes throughout the years, as long as we excluded the major events. Clearly, we could have divided W into more than two subsets, but in some years the number of events is so low that we choose not to do so. In a more extensive study (Van Lieshout and Stein, *subm.*) we further explored the apparent inhomogeneity in the point patterns and we further explored the occurrence of aftershocks.

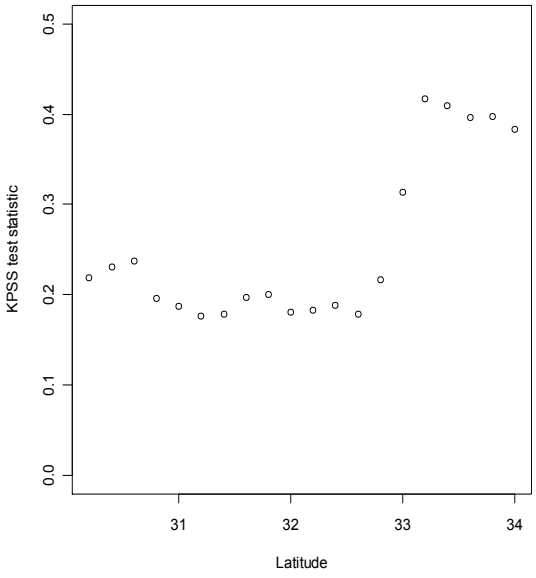


Figure 4a. KPSS test values for bi-annual data for different latitudes boundary between A_N and A_S with $\ell = 4$.

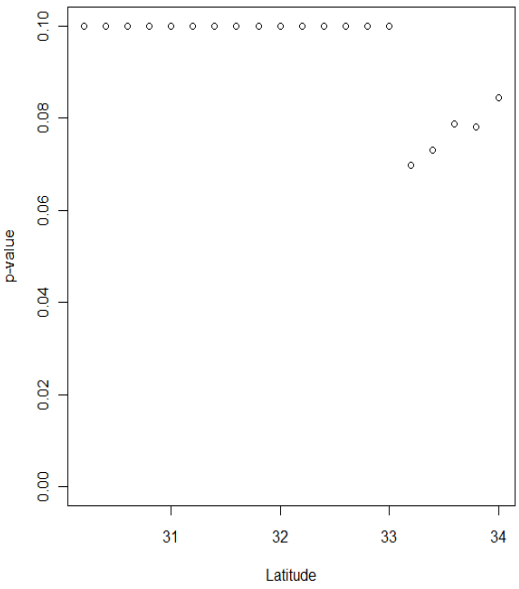


Figure 4b. P-values for the KPSS statistic.

References

Anwar, S., Stein, A. and Genderen, J. van (2011). Implementation of the marked Strauss point process model to the epicenters of earthquake aftershocks. In: Advances in Geo-Spatial Information Science (ed W. Shi), ISPRS Book Series. London: Taylor & Francis.

De Jong, R.M. (2000) A strong consistency proof for heteroskedasticity and autocorrelation consistent covariance matrix estimators. Econometric Theory, **16**, 262-268.

Herrndorf, N. (1984). A functional central limit theorem for weakly dependent sequences of random variables. Ann. Probab., **12**, 141-153.

Holden, L., Sannan, S. and Bungum, H. (2003) A stochastic marked point process model for earthquakes. Natural Hazards and Earth System Sciences, **3**, 95-101.

Kwiatkowski, D., Phillips P.C.B., Schmidt, P. and Shin, Y. (1992) Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics, **54**, 159-178.

Molnar, P. and Chen, W.P. (1982). Seismicity and mountain building. In Mountain Building Processes (ed. K.J. Hsu). London: Academic Press, pp. 41-58.

Newey, W.K. and West, K.D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica, **55**, 703-708.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. Ann. Inst. Statist. Math., **50**, 379-402.

Van Lieshout, M.N.M. and Stein, A. (*Subm.*). Earthquake modeling at the country level using aggregated spatio temporal point processes. Submitted for publication into Mathematical geosciences.

Vere-Jones, D. (1970) Stochastic models for earthquake occurrence. J. Roy. Statist. Soc. Ser. B, **32**, 1-62.

Vere-Jones, D. and Musmeci, F. (1992) A space-time clustering model for historical earthquakes. Ann. Inst. Statist. Math., **44**, 1-11.

Zhuang, J., Ogata, Y. and Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. J. Amer. Statist. Assoc., 97, 369-380.

Ignoring Correlation Leads to Bone Shaped Confidence Regions and other Counter-Intuitive Aspects of Spatial Data Quality

Andrew U. Frank & Gerhard Navratil

TU Wien, Department of Geoinformation and Cartography, Gusshausstrasse 27-29/E127.1, A-1040 Wien, Austria
frank@geoinfo.tuwien.ac.at, navratil@geoinfo.tuwien.ac.at

Abstract

We introduce a straight-forward, practical method to deal with correlation. The approach is based on numerical differentiation and error propagation and consists of two rewriting rules for observation equations: (1) transforming functions in multiple variables to a function in one (vector) variable and (2) converting multiple functions to work on the same vector with all variables. The result is a function for least-squares adjustment that can be integrated in other programs.
Ignoring correlation is a very common assumption and leads to error in tests to derive topology from coordinates, e.g. the discussion on counter-intuitive bone shaped confidence regions for lines.

Keywords: adjustment computation, correlation, line data quality

1 Introduction

We introduce a straight-forward, practical method to integrate correlation in geometric computation using ideas from software engineering. The key contributions are two rules for rewriting of observation equations to functions of one vector input. The result is a short function to solve geometric least squares problems and produces the correlations, which can be integrated in other geometric computations.

The function produces BLUE results with correlations. An example shows how ignoring correlation leads to errors in test for topological relations.

2 Example of Ignoring Correlation: Bone Shaped Confidence Regions

Ignoring correlation is a very common assumption in much what is published in spatial data quality research, primarily justified by the presupposed difficulty of treating correlation.

Chrisman presented 1982 epsilon bands around lines as regions of error (or uncertainty) for point positions in lines (Figure 1), based on older work by Perkal, and used it to assess the quality of area estimates derived from digitized cartographic representations. The term “confidence region” seems to be introduced by Shi and Ehlers (1993). It is later

defined as "A confidence region error model is a band around a measured line segment within which lies the true location of the line segment to a probability larger than a predefined confidence level" (Shi and Liu, 2000, p. 52). Caspar and Schwering improved epsilon bands and proposed bone shaped regions around the lines (Figure 1), computed from the error propagation law with uncorrelated point coordinates.

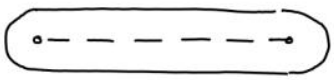


Figure 1: Epsilon band as a confidence region.

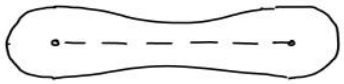


Figure 2: Bone shaped confidence region.

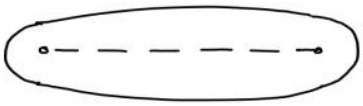


Figure 3: Double bow confidence region.

The bone shaped form is counter-intuitive: why should the error of a point near the middle of the line be smaller than for a point closer to an end? Intuition would suggest a double bow form (Figure 3) indicating that error increases with distance from the given points.

The bone shaped form (Figure 2) seems acceptable when one considers that a point in the middle of a line is computed as the average of the two endpoints and thus its uncertainty is reduced by a factor $\sqrt{(2)}$. Shi and Tempfli (1994) and eventually Shi and Liu (2000) refined the model to take correlation into account. The resulting formulae are very complicated; their results seem counterintuitive. Applying the rules of error propagation to compute the variance of a point in the middle of a line, taking into account correlation, the variance for a point in the middle can vary between 0 (correlation is -1) and σ of the endpoints, which would be definitely a double bow form (Figure 3).

The test whether a point is on a line or left or right of it is a crucial test in many geometric algorithms. It is necessary when constructing a topological data structure, it is used in most 'point in polygon' tests (Guibas and Stolfi, 1985) or when triangulating a set of points. Simplistic tests, using Chrisman's epsilon bands abound, but can be wrong as will be shown with an example.

3 How to Obtain Correlations for the Points Used in a Construction

To assess the quality of the results in a geometric computation with the error propagation law requires the correlation of the points used. Where to get correlations from? How to produce correlations?

Geometric measurements are typically uncorrelated. To guard against gross errors and to improve the quality of the result, surveyors measure more than the bare minimum and use adjustment computation to get the best linear unbiased estimator (BLUE) for the parameters of interest. After adjustment, the results are correlated. The correlations are often computed, but are typically not stored for further use.

As an example consider some constructions based mostly on distance measurements (Figure 4). Unfortunately, only non-trivial nets of measurements show effects of correlations, thus the example network of measurement contains 17 distances (Figure 4). The network of measurements here is constructed from two fixed points (1 and 2), from which points 10, 12, 20, 22, and 32 are derived with about equal precision; these points are correlated. The points 50 and 51 are related by distance measurements to points 10 and

12, the point 41 is related to 20 and 32 and fixed on the line 20 to 32 — again producing similar point precision (lower than for the points 10 to 32). The question is, whether points 41, 51, and 52 are on the line 20 to 22 (i.e., what is the distance between point and line and is it significantly different from 0).

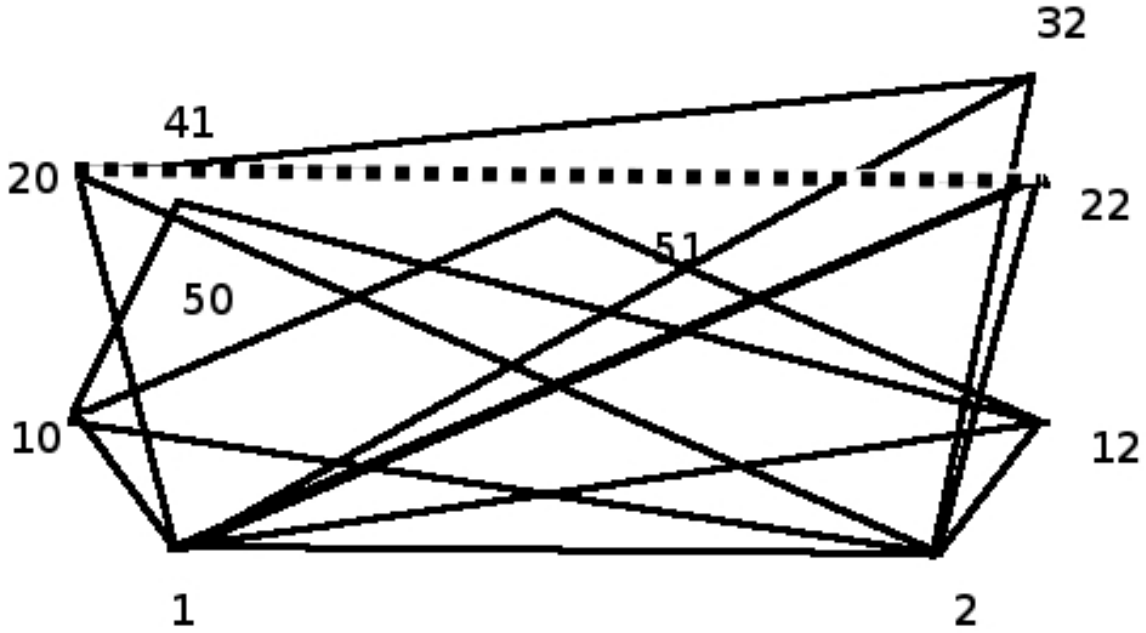


Figure 4: A set of distance measurements to determine points

4 Computing with Correlation

The law of error propagation is succinctly described with the Jacobi matrix *J*, which contains the partial derivatives of a (set of) observation equations. A numeric calculation of the elements of *J* for an adjustment problem with multiple observation equations is in two steps. In a second-order, functional programming language (e.g., Haskell or javascript) these steps can be written as a function to apply to the equation, written as a function.

4.1 Produce Functions with Vector Arguments

Given a formula *f* with some arguments $p_1, p_2, ..., p_n$ producing a result $r = f(p_1, p_2, ..., p_n)$, e.g., the formula for a distance observation. The partial derivatives for many equations can be given in closed form, but numerical differentiation is a more generalizable solution. The formulae must be rewritten to take a vector of values in lieu of some points as inputs and then the vector *A* of the partial derivatives a_i at the point $P_o = [p_{o1}, p_{o2}, ..., p_{on}]$ is:

$$q_i = \{f(p_{o1},p_{o2},...,p_{oi}+\varepsilon,...,p_{on}) - f(p_{o1},p_{o2},...,p_{on})\} / \varepsilon = \{f(P_o+\varepsilon E_i)\text{-f}(P)\} / \varepsilon \qquad (1)$$

where E_i are the rows of the unit (diagonal) matrix.

A typical geometric formula must be reformed from individual arguments, e.g., points, to a single vector input. This single vector parameter is just a concatenation of the input point vectors. The change is comparable to the transformation of functions of multiple

variables to functions with a single variable, which is a tuple of the variables, which is called currying after Haskell Curry (1900-1982). The rewriting can be done with a second order function *vec2param*, which takes a function in 2 parameters and produces a function that has a single parameter (!! is the Haskell list indexing function, starting with index 0)

```
vec2param :: ([a] -> [a] -> t) -> [a] -> t
vec2param op a = op [a!!0, a!!1] [a!!2, a!!3]
```

The code is written in Haskell, a compiled functional programming language (Jones, 2003) but a solution in Mathematica, MathLab, or the open source Octave would not look substantially different; a comparable program could be written in javascript as well.

4.2 Integration Multiple Equations

To produce a model for adjustment, all the formulae must be integrated to operate on a single large vector, which includes all u parameters occurring in the problem (i.e., all coordinate values for the points 10, 11...). The formulae are rewritten from the formula (1) above to pick the right argument values out of the larger single vector (2);

$$f'(x_1, x_2, x_3, x_4)=f''(x_k, x_l, x_p, x_r) \quad (2)$$

where, for example, the indices 1..4 become the four values (k, l, p, r) to pick out the corresponding coordinates.

After rewriting to form f'', all formulae operate on the same vector. The required relabeling is similar to unification in the execution of a logic or functional programming language. The arguments of a function are named with the point and coordinate names, *e.g.*, "10-x" or "22-y" and a vector of all unique parameter names occurring is constructed (*table* in *obsformula2*). It is used by the operation *param2vecix* to pick for each input of f' the corresponding element of the integrated vector. The resulting formulae f'' operate on the single, integrated vector.

```
obsformula2 obs table f' inp = f''
    where
    f'' v = f' (map ((v !!). (param2vecix table)) inp)
    param2vecix tab i = fromJust . lookup i $ tab
```

The vectors consisting of these functions f'' describes the geometric model of the construction, i.e., the Jacobi matrix J.

5 Topological Tests

The variance of the distances of the point 41, 51, and 52 from the line are directly obtained from the matrix N^{-1} including the effects of correlations. They are: 0.3 for 41, 1.9 for 51, and 1.8 for 52. The variance for point 41 is much smaller and therefore a smaller distance from the line will be accepted as 'not on the line' compared to points 51 and 52; this is different from the computation without correlation.

The correlation expresses in numeric form the geometric fact that points on one line are not on another line through a common point. The approach is used here for distances and distances of points from lines for didactic reasons only and generalizes to other observation types. For example, the observation that a line is parallel to another one with a given distance, angles, alignments, or right angles, etc. Any function given as $f(p_1,..)$ can be rewritten with the methods shown.

6 Conclusion

The result is a general function for least squares adjustment for geometric construction, which can be built into other programs; it is built from layers:

1. Construct the list of observation equations from individual descriptions; new observation types (e.g., angle measurements) require 5 to 10 lines of code.
2. Transform to a least squares problem, which is represented as a data type and then passed to a function

   solveAdjustment ::

   LeastSquaresProblem -> LeastSquareSolution; 20 to 30 lines of code.
3. The *solveAdjustment* function calculates the Jacobi matrix; 10 lines of code.
4. A function *gauss_markov* solves the least squares problem and computes all the desired result values; it uses well-tested, highly efficient routines for matrix operations (e.g., LAPACK by Anderson *et al*. (1999)).

A program can use the full stack or only part of it, producing, for example, the data in *LeastSquaresProblem* directly and use only *solveAdjustment.*

The method allows to compute — using the well-known least squares approaches — the variance and covariances for whole chains of geometric constructions. The example here shows that correlations between computed points are high (often above 0.7) and must be taken into account. Testing for geometric conditions, e.g., whether a point is on a line or not, can only be done reliably if the correlations are considered.

The method can be used as a building block for a measurement based cadastre (Buyong, 1992; Navratil *et al.*, 2004; Goodchild, 2004), where data quality for point locations is properly determined and thus a step toward a defensible data quality assessment for geometric positions in GIS. The result contributes to the other long-standing research question to relate GIS and CAD construction programs and adjustment computation (Kuhn, 1989).

References

E. Anderson, Z. Bai, and C. Bischof. (1999), *LAPACK Users' guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

T. Buyong. (1992), *Measurement-Based Multi-Purpose Cadastral Systems*. PhD thesis, University of Maine, Orono.

N. Chrisman. (1982) Theory of cartographic error and its measurement in digital data bases. In *5 th Int. Symp. Comp. Assisted Cartography & Int. Soc. Photogrammetry and Remote Sensing Commission IV*. Crystal City, VA, pp. 159–168.

M.F. Goodchild. (2004), “A general framework for error analysis in measurement-based GIS”. *Journal of Geographical System*s, 6(4):323-324.

L. J. Guibas and J. Stolfi. (1985), “Primitives for the manipulation of general subdivisions and computation of voronoi diagrams”. *ACM Trans. Graph.,* 4(2):74-123.

S.P. Peyton-Jones. (2003), *Haskell 98 language and libraries: the revised report*. Cambridge Univ. Press.

W. Kuhn. (1989), *Interaktion mit raumbezogenen Informationssystemen - Vom Konstruieren zum Editieren geometrischer Modelle,* volume Mitteilung Nr. 44. Institut für Geodäsie und Photogrammetrie, ETH Zürich, Dissertation Nr. 8897.

G. Navratil, M. Franz, and E. Pontikakis. (2004), Measurement-Based GIS Revisited. *AGILE 2004 7th Conference on Geographic Information Science*, Crete University Press, pp. 771-775.

W. Shi and M. Ehlers. (1993), S-BAND, a model to describe uncertainty of an object in an integrated GIS/remote sensing environment. In *International Geoscience and Remote Sensing Symposium, 1993. IGARSS 93. Better Understanding of Earth Environment.*, IEEE, Tokyo, Japan, pp. 1721-1723.

W. Shi and W. Liu. (2000), A stochastic process-based model for the positional error of line segments in GIS. *International Journal of Geographical Information Science*, 14(1):51-66.

W. Shi and K. Tempfli. (1994) Modelling positional uncertainty of line features in GIS. *Proceedings of AS-PRS/ACSM 94*, Reno, Nevada, pp. 696-705.

APPLICATIONS

GSM data analysis for tourism application

Ana-Maria Olteanu[1], Roberto Trasarti[2], Thomas Couronné[1], Fosca Giannotti[2], Mirco Nanni [2], Zbigniew Smoreda[1] & Cezary Ziemlicki[1]

[1] SENSe/Orange Labs, Paris.
firstname.lastname@orange-ftgroup.com
[2] KDD lab, ISTI – CNR – Italy.
firstname.lastname@isti.cnr.it

Abstract

In this work we propose a case study on human mobility as perceived by telecom operators' infrastructures. Based on GSM data of roaming users, we try to understand how visitors of a touristic area use the territory, with particular emphasis on visits to attractions. Several issues due to the nature of available data are explored, providing hints and heuristics to tackle them, and a comparison with external information sources on touristic activity is performed.

1 Background and data description

The extremely large diffusion of mobile location-aware technology, which includes also common mobile phones, provides a large-scale basis for continuously probing the mobility behaviors of population. Such knowledge, then, can enable several (mobility-aware) services to mobility managers as well as to individual users.

The work presented in this paper goes in such a direction, by joining the abundance of implicit knowledge hidden in the GSM data of a telecom operator, with the capabilities of a mobility data analysis system, called M-Atlas (Trasarti *et al*. 2010). The data is collected by a network-based system infrastructure; therefore the recording of the user position is triggered by a set of specific events: (i) start call (outgoing, incoming), (ii) SMS (outgoing, incoming), (iii) handover (antennas update during the call) (iv) location update (LAC: a set of antennas grouped together). Substituting each tower with its spatial location, a dataset of spatio-temporal points is obtained. It's important to notice how GSM data is both *approximated* in space and *incomplete* in space and time; in the following sections we will deal with the uncertainty and the complexity of data representing only fragments of the users' movement w.r.t. fixed position of tower cells, which reduces the granularity of space to a small set of locations.

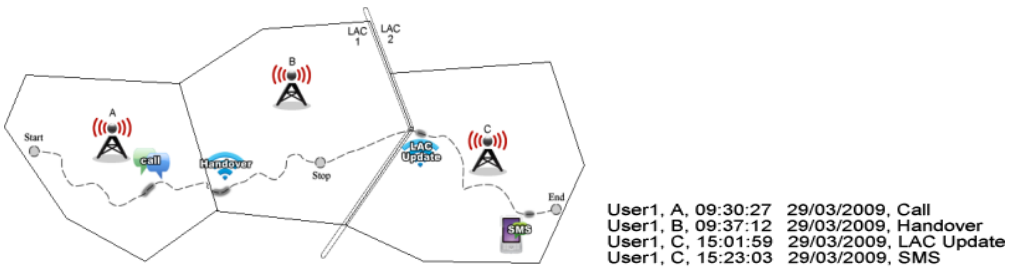


Figure 1. Example of the approximation and incompleteness of the data.

To better clarify this uncertainty, Figure 1 shows an example of a user: the dotted line represents his\her real movement where some portions are highlighted in case of events (a call, an handover, a LAC update, and an SMS) and the recorded information are reported in the bottom of the figure. This example, points out the simplification of the user's movements caused by the recording system. In the next section we present a methodology to deal with this kind of data to estimate the tourism in Paris Region.

2 Tourism in Paris Region

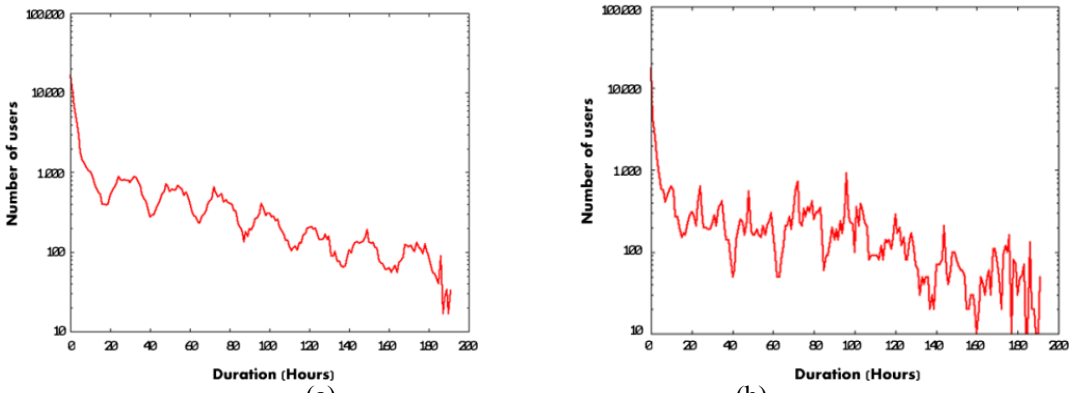
The objective of our analysis is threefold: (i) trace a methodology for dealing with data quality issues in GSM data; (ii) validate the methodology against statistics and background knowledge in tourism field; and (iii) extract a touristic profiles characterization, enabled by our methodology.

2.1 Tourist trajectories reconstruction

In this study we used data coming from QoS GSM network probes collected by Orange France. Hence, from this large amount of data, we approximate the concept of tourist with the technical category of roaming data for three reasons: (i) all the users living in Paris are avoided, (ii) France remains the world's number one tourist destination (79.3 million international visitors), while Paris Region is still the leading French region (Oetp, 2009) and (iii) the subset division is well defined and remains stable in time. The roaming dataset contains ~1.5 millions anonymized users in Paris Region (corresponding to the 5% of total Orange users) during two weeks. Removing the users with less than 3 points the trajectories concatenating the user's points are built; the result is a set of ~1.1 millions trajectories. The first aspect of the *incompleteness* of data is that we cannot generally know when a user starts his visit in Paris or when he/she leaves, for two reasons. First, roaming users may *appear* or *disappear* in the recording phase due to the fact that they can connect to different mobile phone network operators. Secondly, if any event occurs, we do not know whether he/she leaved Paris or not. To overcome these problems we consider only a subset of users which satisfy a constraint: they must have the first and the last point of his\her trajectory in one of the Paris airports. This constraint guarantees the fact that the trajectories are completed at least for the temporal aspects. Using the two major airports in Paris: *Charles de Gaulle* (CDG) and *Orly* we obtain two datasets: D^{CDG} and D^{Orly} .

2.2 GSM Traces as a proxy of tourist behaviour

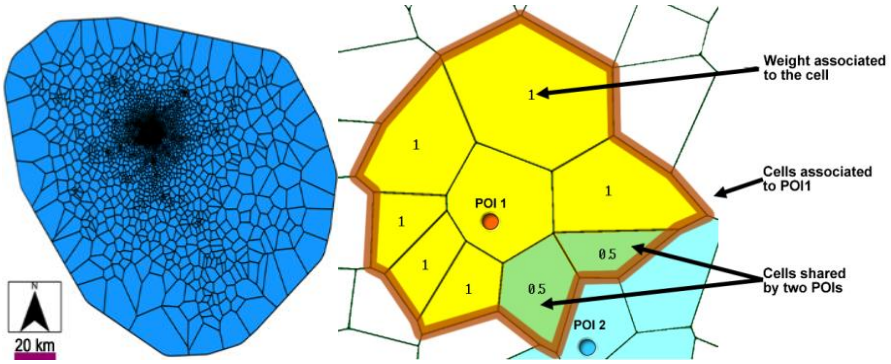
First, we have studied the distribution of duration of stay of the users belonging to datasets D^{CDG} and D^{Orly} to determine if the results are reasonable according with the airports policies (Fig.2). The distributions of duration show the periodicity of the users. The presence of peaks every 24 hours reflects the two already known situations: (i) there is a limited number of flights to come back during the days and (ii) the users tend to use flights at the same hours. This empirically validates the significance of data obtained through the selection and manipulation tasks described above. In this work we focus on describing the tourist' behaviors w.r.t a set of *Tourist Points Of Interests (POIs)*, such as Notre Dame or the Tour Eiffel.



(a) (b)

Figure. 2 (a) Charles De Gaulle and (b) Orly users' distribution of durations during the first week.

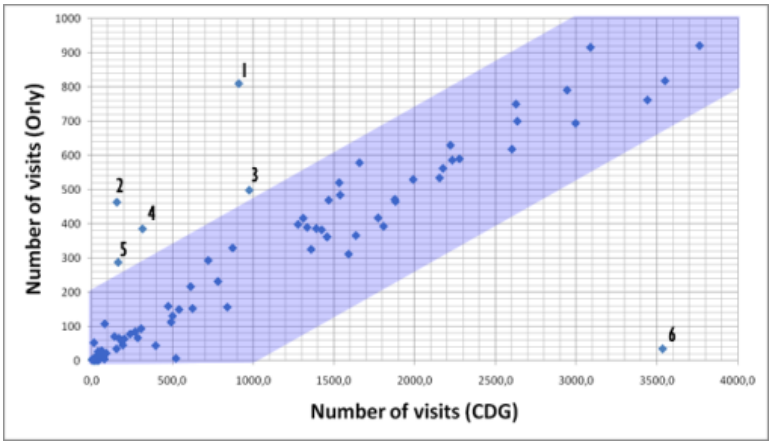
Performing this analysis we consider that we do not have an event every time a tourist visits a place and we cannot trace a single user in all his/her visits. However, this is a general phenomenon that happens independently from the single user and the spatial location, therefore we can expect that our observations provide a good sample of actual visits – as such, the frequencies estimated this way will need a rescaling in order to be comparable with real frequencies. Moreover, since our data is a sequence of tower cell positions, POIs need to be mapped to antennas in order to know which users visit which POIs. To solve this problem of granularity, we adopted a method to partition the space known in literature as *centroid Voronoi tessellation* (Qu *et al*. 1999) using the position of towers cells as centers. In Fig.3 (a) the complete partitioning of the Parisian area is shown, and is evident that the tessellation in the center of Paris is dense due the fact that there are more towers cells, which determine also our precision for representing POIs.



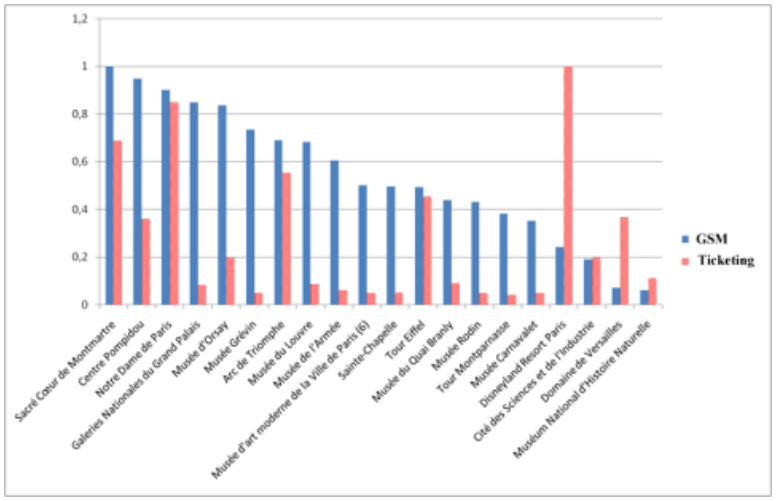
(a) (b)

Figure. 3 (a) The Voronoi tessellation of the Parisian space using the antennas as centers, and (b) the mapping between POIs and cells.

Having these partitions we select for each POI the Voronoi cell that contains it. Due to the degree of uncertainty in the connection policy of the GSM network, all the neighboring Voronoi cells are selected as buffer.



(a)



(b)

Figure. 4 (a) Comparing the number of visits of POIs considering the two dataset D^{CDG} and D^{Orly}. (b) Comparison between GSM data results and the ticketing data.

The process determines a set of Voronoi cells for each POI. This way, each Voronoi cell can be shared by several POIs. Therefore Voronoi cells are *weighted* with a value 1/n, where n is the number of POIs insisting on the cell as shown in Fig.3 (b). To compute the frequency of visits in a POI, the number of users which visit a cell associated to it (considering the weight) is computed. From this comparison, the points in the blue area are selected as points which are coherent for both datasets and then more reliable. In particular, we can spot some outliers (Fig.4 (a)): (1) Les Catacombes (2) Roseraie du Val-de-Marne (3) Musée Zadkine (4) Ecomusée du Val de Bièvre (5) Château de Sceaux and (6) Musée de l'Air et de l'Espace. These points are errors due to the fact that they are near the main road from the airports to the center. In fact, the users do not visit them, but simply pass near them to reach the center of Paris.

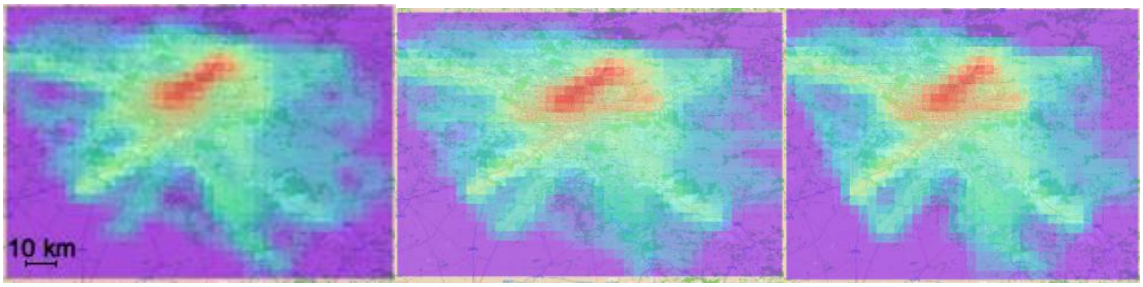
2.3 Validation against ticketing data

In this section we compare our results using GSM data against tourism data (essentially the estimated number of visiting POI) for the first 20 POIs according to (*Oetp, 2009*). The results are summarized in Fig.4 (b). We can notice that the estimation of attendance of the POIs using GMS data is correlated to or, more often, overestimates tourism data. Thus, there are some POIs (e.g. Notre Dame de Paris, Tour Eiffel, Arc de Triomphe, etc.) for which GSM and tourism data are similar. These places are the most visited according to both GSM and tourism data. We believe that the overestimation for the other POIs is

due to the fact that tourists are present in the neighborhood of the POI without actually buying a ticket. Moreover, POIs such as Galeries Nationales du Grand Palais, Sainte Chapelle, Tour Montparnasse, Centre Pompidou, etc. are situated in popular districts in Paris. Note that, for two POIs the number of visitors is underestimated, i.e. Disneyland Paris and Versailles. Versailles is the sixth tourism destination according to tourism data and only at the seventeenth place according to the GSM data, similarly Disneyland Paris is the first tourism destination according to the tourism data does not appear strongly visited in GSM data. This phenomenon can be attributed to two main characteristics of such POIs: first, our data do not capture French visitors, which form a large part of their affluence (more than most other attractions of Parisian area); second, they are rather isolated from the city center and the visitors are most likely to buy tickets in order to visit them, since the exterior offers no attractions. That means that the underestimation of visit frequency based on ticketing in these places is almost negligible as opposed to most other attractions.

2.4 Creation and Analysis of tourist profiles

Due to space limitations we present only an example of how to study data in order to highlight differences w.r.t. categorizations of users. Consider three different categories: (1) *Short stay (1-2 days),* (2) *Medium stay (2-5 days)* and (3) *Long stay (5-7 days).*



(a) (b) (c)

Figure. 5 Density map of movements in the three categories (a) short, (b) medium and (c) long. The density map is computed using points/segments intersections.

We compute the density map of movements for the three categories (resulting in a profile for each one), as shown in Fig.5(a, b, c). The major difference in the behaviors of the three categories is the visit at Disneyland Paris, which becomes more and more popular as the stay duration increases. This represents the fact that it is a preferred location only for the users that have a long visit and it is not really appealing for the short visiting users focused more on the attractions in the center of Paris.

In this experiments we have used different techniques for the map distribution: Kernel-based (Katajisto *et al.*), Time-based (Downs *et al.*), Gaussian-based (Olteanu *et al.*) or considering only the points/segments intersections. The results obtained are very similar, showing no significant differences.

3 Conclusion and future works

This work yielded several results, both at the methodological level and in terms of additional knowledge about tourism in the area of study. The comparison with existing touristic information allowed to spot technical issues in the use of GSM data, yet confirmed the viability of the proposed analysis process. Finally, using such data we were

able to highlight and quantify some behaviors that conventional statistical methods have not been able to identify.

The method proposed has some limitations. Some possible researches to improve our method go to the following directions: (i) we should distinguish between different categories of users: business, tourists, family visits, etc. classifying them by space usage and/or additional background information; (ii) study the activities of the users developing a stop detection algorithm for this type of data; and (iii) we want to extend our approach to all users and not only to airports users. The analyses presented here are under deep development, and next activities include the extension of the study to French users and the use of CDR data.

References

Trasarti R., Giannotti, F., Nanni M., Pedreschi D., Renso C. (2011), "*A Query Language for Mobility Data Mining*". *In:* Int. J. of Data Warehousing and Mining. Vol. 7 pp. 24-45.

Observatoire économique du tourisme parisien (Oetp) (2009), *Fréquentation des sites culturels parisiens en 2008.*

Du Q., Faber V., Gunzburger M.(1999), *Centroidal Voronoi Tessellations: Applications and Algorithms,* SIAM Review Vol. 41 pp. 636-676.

Katajisto J. and Moilanen, A. (2006): *Kernel-based home range method for data with irregular sampling intervals*. Ecological Modelling, 194(4): 405-413.

Downs, J.A. (2010): *Time-geographic density estimation for moving point objects*. Geographic Information Science, Lecture Notes in Computer Science, 6292, Springer-Verlag, 16-26.

Olteanu, A-M., Couronné, T., Fen-Chong, J., Smoreda, Z., *Modélisation des trajectoires spatio-temporelles issues des traces numériques de téléphones mobiles* ; Le paris des visiteurs, qu'en disent les téléphones mobiles? Sagéo, 2011.

A Quality approach to Volunteer Geographic Information

Pau Aragó, Laura Díaz & Joaquín Huerta

Geographic Information Group, INIT, University Jaume I, Castellón, Spain
pauarago@yahoo.es, laura.diaz@uji.es, huerta@uji.es

Abstract

Volunteer geographic information (VGI) has become an alternative geospatial data source. Geospatial data quality coming from official institutions is attending to a set of parameters. VGI data is built in a free collaborative way by people with a different background and expertise. Quality parameters should be adapted to the new way of building geospatial data. VGI's quality description VGI will provide an overview of this informational value to be used in VGI projects and will provide a minimum overview to guarantee its usability. In this paper we propose an adaptation to GIS quality parameters to the idiosyncrasy of the VGI.

Keywords: Volunteer Geographic Information, quality, GIS, geospatial information.

1 Introduction

With the emergence of Web 2.0 Internet provides tools and services that hide the technology allowing an easy way of publishing geospatial information (GI). Therefore, GI has begun to be built into web 2.0 technology within a community or lonely effort. This new phenomena in geography has been named 'neogeography' (Turner, 2006), 'cybercartography' (Tulloch, 2007), or 'voluntary geographic information (VGI)' (Goodchild, 2007).

The Volunteer Geographic Information (VGI) (Goodchild, 2007) is a term used to define the personal contribution of people in collectively building a geospatial information resource. A resource could be a street map like Open Street Map, a geotagged photo or a photo-collection like Flickr (http://www.flickr.com/map/) or Panoramio (http://www.panoramio.com/) or a data validation like Geo-Wiki (Fritz *et al*., 2009).VGI has become a reality thanks to the development of the Web 2.0 (Goodchild, 2007).

Despite the advantages of VGI there is a quality issues associated to this kind of information. A quality problem could be achieved by comparing against similar data its accuracy and quality (Haklay, 2010). However, it is not always possible to compare VGI data with official data, because sometimes an "official" source doesn't exist or VGI is less structured and described.

Therefore we would like to remark one of the challenges for the VGI field which is to generate geospatial information usefully. In other words, assure that generated VGI will meet certain quality parameters that will allow its reusability by other users in other scenarios demonstrating its added value.

2 Quality definition for geospatial data

The quality of geospatial information refers to how well a real object is represented in a geospatial storage. One of the challenges in VGI is to define the quality of the data The quality classification division for professional geospatial information are 10 elements, lineage, positional accuracy, attribute accuracy, logical consistency, completeness, semantic accuracy, usage purpose, temporal quality, variation in quality, metaquality (Van Oort, 2005) (Devillers *et al*., 2010).

The metaquality and variation in quality are elements that are not extracted directly from geospatial layer, but rather derived from quality elements of the list. Also, the usage purpose is strongly related with the other quality elements.

3 Quality adaptation for VGI data

A VGI volunteer project based on citizens' collaborative work, is more focused on uploading spatial object or attributes than in taking care of the spatial quality of the user contributions (Goodchild, 2008). A user contributes to a VGI project according to his/her skill and his/her capabilities.

Official geospatial information project starts defining some scientific criteria, like defining the project's minimum quality level. Therefore geospatial information is built and checked to fit all the features within the previously defined quality level. Applying official geospatial data quality criteria to a VGI project will have poor results as the quality criteria that a user must follow has not been established. The elements taken from point 2 to define VGI quality data approach are, lineage, positional accuracy, attribute accuracy, logical consistency, completeness, semantic accuracy, temporal quality.

3.1 Lineage

Lineage is defined as data history. VGI's basic data history is the user name or Id and the creation date. This information could be completed by defining the data source, GPS, map, magazine, URL, references ... For instance GPS tracks are the main VGI data source in Open Street Map (OSM); moreover in OSM any open data source as a base for content creation could be used. Nevertheless, it is not mandatory but recommended, when geospatial information is published in OSM to describe the data source. In VGI projects such as OSM and Wikimapia, among others, users can publish any type of geospatial object without source references. In order for VGI data to have scientific value, data history annotation must be mandatory. This procedure prevents new data creation without a reference. Therefore, this referenced VGI data will have an added value.

Finally, a VGI project will have different geospatial data level, one top-level with a lineage footprint, which could be checked its lineage in any moment. A second level with no lineage quality, self-sourced or not verified source. This second level becomes a matter of trust on the volunteer.

3.2 Positional accuracy

Depending on the positional accuracy, geospatial data has limitations in its usability. For instance, a road navigation GPS device needs a positional accuracy at least equal to the GPS device's precision or enough accuracy to place the car in the right road or street.

On the other hand, geospatial data for civil engineering requires a subcentrimetical precision.

VGI data is georeferenced with a GPS device or reference cartography. With a GPS device, positional accuracy depends on the GPS's receiver precision. Nevertheless, this precision could be annotated when information is introduced directly from the GPS receiver (Meng, 1998) (Jwo, 2001). If data is exported using the GPX format, signal accuracy is not annotated into the file (GPX schema n.d.). The GPX format is mainly used to export the GPS tracks and waypoints. In those cases where there is no possibility to get the GPS accuracy it will be assumed that stand alone GPS accuracy is +/-100 meters (Xiaoying Kong, 2007) Nevertheless, if GPS information is saved in RINEX (Gurtner, 2007) a format could be use to post process the GPS information to estimate its positional accuracy.

Another possibility is data georeferencing using reference maps or cartography. In this case the map resolution is the highest reachable precision for georeferenced data (Goodchild, 2001). To get accuracy for VGI data according to a reference maps there are two possibilities:

1) With raster layers as the reference source, restricting georeferencing to an edition zoom to fit raster resolution to screen resolution. In this way digitalization accuracy is similar to the raster resolution;
2) Defining the positional accuracy related to the zoom level. To apply this method the system must know the spatial resolution of the reference at each zoom level.

Even when restricting data digitalization, human mistake is always feasible. Human mistake is the error when the user is assigning a position to the data, not because of the map reference, but by his own mistake. In some cases VGI data will provide similar accuracy as the official geospatial information (Haklay, 2010).

3.3 Attribute accuracy

This quality element is difficult to fit into the VGI quality scope. Although it is possible to predefine a list attributes, assigning a right attribute to an object is not guaranteed. Again, a post-process validation may correct a mistake by assigning a new attribute.

It is possible to improve the attribute assignment. An assignment could be done crossing or checking it with related geospatial information such as reference information. For example, when a user is registering a tree species in a VGI platforms. The object has a correct georeference placement. Moreover, the tree name assignment attribute could be checked according to theoretical plant distribution, warning the user when a plant name is being tagged out of is theoretical distribution area. Furthermore, some attributes can be automatically added from referenced geospatial data such as climate, biographical region,...

The VGI have different sources and on of those are the social networks such as Twitter (www.twitter.com), where a user can publish georeferenced short messages. The goodness of the information is measured according the density of similar messages (Schade. S *et al.,* 2010).

3.4 Logical consistency

When a VGI user georeferences a point, logical consistency is not checked. An object like a house, picture, could be misplaced out of its logical position. Logical consistency must be checked and detect errors for misplacing objects (Goodchild, 2008). For instance,

a bad road junction connection is a problem for routing calculation within this geospatial data. Miss-junction error is a way to check logical consistency. The VGI project will run a topological testing when a geospatial object is created, to automatically detect logical consistency. Misplaced objects can be checked with reference layers. Not all the misplaced objects can be automatically detected. When a house is placed inside a wrong block, only the user checking it could correct this mistake.

3.5 Completeness

Completeness is one of the big differences between VGI projects and traditional Geographical information (Haklay, 2010). When a VGI project is complete? Projects like OSM, could be partly completed some day for national country level or even European level for a car road a navigation purpose. When Open Street Map is completed at world-wide level?. Moreover, geospatial information might be growing and growing on OSM with new information base on previous information introduced, bus stops, hospitals, bench, ...

The completeness could be approximated by dividing the coverage area of the VGI project in subareas (quad-tree) depending on the variability number of objects represented in these areas (Maué and Schade, 2008). From a subarea variability representation is calculated using a ratio (1). In this way the completeness is always related with the VGI project evolution.

$$Completeness = \frac{Number\ of\ objectes\ of\ the\ region}{Maximun\ number\ of\ objects\ of\ any\ region} \quad (1)$$

3.6 Semantic accuracy

In a VGI project a user is able to freely annotate an object attribute and in this way the semantic accuracy should be difficult to check. Moreover, the user's interpretation could be different from the creator's interpretation. Even when restricting the user's freedom, the semantic accuracy will be low, because it depends on the user's training and knowledge. One approach to the semantic accuracy is the use of folksonomies (Bishr and Kuhn, 2007). By using folksonomies it is possible to take advantage of multiple contributions. A user can tag a geospatial object created by himself or other users adding its meaning to the object, thus, the more tags an object has the more semantic accuracy is provided to the user. The objectives to take advantage of a large number of users within a VGI project with edition privileges to tag an object. In this way an object receives a greater number of interpretations.

3.7 Temporal quality

Temporal quality depends on how fast the real world is changing and how fast those changes are translated to digital cartography. There are things which have practically had no changes over the past years, like a cathedral or a church. Other things are changing in a short period of time such as new houses and building areas. VGI projects are constantly being edited, published and reviewed, whereas "official" data revision depends on its cost (Goodchild, 2008) and most of the time it is slower than the changes.

Regarding temporal quality in VGI, at least we know its object creation date. Nevertheless, it is possible to record the number of times an object attributes has been revised

or modified. Therefore, it is possible to calculate a ratio of graphical changes and attribute changes for a geospatial object (2).

$$Changeratio = \frac{Number\ of\ updates\ since\ creation}{last\ update\ date - creation\ date} \quad (2)$$

The object changes described above (2) are updated changes in its attribute. The validation and correction change description are in points 3.2 and 3.3.

A more general approach to the temporal quality is the number of volunteer contributions in a time period (3). This contribution ratio is a measurement of VGI project activity. Nevertheless, this measurement is relative to how fast change happens in the real world.

$$Contributionratio = \frac{Number\ of\ new\ contributions}{time\ period} \quad (3)$$

4 Conclusion

In this paper Geospatial data quality has been discussed to suit VGI features. VGI project's peculiarities do not allow for a predefined minimum geospatial data quality set. Quality in VGI project is attained, depending on the volunteer contributions to the project. On the other hand, it is possible to measure contributions quality or constrain it inside some quality parameters. It is possible to improve volunteers 'contribution by providing them with tools to assist them in the data generation process such as digitalization within a predefined zoom level or scale.

The information on a VGI project is created in a collaborative way. In fact it allows semantic accuracy to be improved by using the folksonomies, adding tags by other users different than the creator. If on VGI project there is an active community quality could be improved.

The quality measurement of a VGI project allows the use of VGI geospatial data to be evaluated objectively for a specific purpose. The VGI's project quality will be properly defined and measured; however in a collaborative project information credibility is as important as quality.

Acknowledgments

This work has been partially supported by the European FP7 Project called EuroGEOSS, by the CENIT España Virtual project through the Instituto Geográfico Nacional (IGN) and I+D research project through Fundación Bancaja.

References

Bishr, Mohamed, and Werner Kuhn. 2007. Geospatial Information Bottom-Up: A Matter of Trust and Semantics. In *The European Information Society*, ed. Sara Irina Fabrikant and Monica Wachowicz, 365-387. Lecture Notes in Geoinformation and Cartography. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-540-72385-1_22.

Devillers, Rodolphe, Alfred Stein, Yvan Bédard, Nicholas Chrisman, Peter Fisher, and Wenzhong Shi. 2010. Thirty Years of Research on Spatial Data Quality: Achievements, Failures, and Opportunities. *Transactions in GIS* 14, no. 4 (8): 387-400. doi:10.1111/j.1467-9671.2010.01212.x. http://doi.wiley.com/10.1111/j.1467-9671.2010.01212.x.

Fritz, Steffen, Ian McCallum, Christian Schill, Christoph Perger, Roland Grillmayer, Frédéric Achard, Florian Kraxner, and Michael Obersteiner. 2009. Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover. *Remote Sensing* 1, no. 3 (8): 345-354. doi:10.3390/rs1030345. http://www.mdpi.com/2072-4292/1/3/345.

Goodchild, Michael. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69, no. 4: 211-221. doi:10.1007/s10708-007-9111-y. http://dx.doi.org/10.1007/s10708-007-9111-y.

Goodchild, Michael F. 2001. Metrics of scale in remote sensing and GIS. *International Journal of Applied Earth Observation and Geoinformation* 3, no. 2: 114-120. doi:10.1016/S0303-2434(01)85002-9. http://www.sciencedirect.com/science/article/B6X2F-46DMS49-2R/2/ae1741ff17c81c62b7b65d06204f0757. 2008. Spatial Accuracy 2.0. In , 1-7. Shanghai, P. R. China, June. http://spatial-accuracy.org/system/files/Goodchild2008accuracy.pdf.

GPX schema. n.d. GPX 1.1 Schema Documentation. http://www.topografix.com/gpx/1/1/.

Gurtner, Werner. 2007. RINEX: The Receiver Independent Exchange Format. Version 3. Astronomical Institute University of Berne. ftp://ftp.unibe.ch/aiub/rinex/rinex300.pdf.

Haklay, Mordechai. 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design* 37, no. 4: 682 – 703. doi:10.1068/b35097. http://www.envplan.com/abstract.cgi?id=b35097.

Jwo, Dah-Jing. 2001. Efficient DOP Calculation for GPS with and Without Altimeter Aiding. *The Journal of Navigation* 54, no. 02: 269-279. doi:10.1017/S0373463301001321. http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=74553.

Maué, Patrick, and Sven Schade. 2008. Quality of Geographic Information Patchworks. In . Girona, Spain. http://plone.itc.nl/agile_old/Conference/2008-Girona/PDF/111_DOC.pdf.

Meng, T.H. 1998. Low-power GPS receiver design. In *Signal Processing Systems, 1998. SIPS 98. 1998 IEEE Workshop on*, 1-10. doi:10.1109/SIPS.1998.715763.

S. Shade, G. Luraschi, B. De Longueville, S. Cox, Laura Díaz. Citizens as sensors for forest fires: Sensor Web Enablement for Volunteered Geographic Information. M.A. Brovelli, S. Dragicevic, S. Li, B. Veenendaal (Eds): *ISPRS Workshop on Pervasive Web Mapping, Geoprocessing and Services XXXVIII-4/W13 (*WebMGS 2010*). Como, Italy, August 2010, ISSN 1682-1777.*

Tulloch, David L. 2007. Many, many maps: Empowerment and online participatory mapping. *First Monday* 2, no. 2 (February). http://firstmonday.org/issues/issue12_2/tulloch/index.html.

Turner, Andrew. 2006. Introduction to Neogeography.

Van Oort, Pepijn. 2005. Spatial data quality : from description to application. Delft: NCG Nederlandse Commissie voor Geodesie.

Xiaoying Kong. 2007. GPS Modeling in Frequency Domain. In *Wireless Broadband and Ultra Wideband Communications, 2007. AusWireless 2007. The 2nd International Conference on*, 61. doi:10.1109/AUSWIRELESS.2007.36.

Quality Assessment for Cadastral Geometry

Gerhard Navratil

Vienna University of Technology, Institute for Geoinformation & Cartography,
Gusshausstr. 27-29, A-1040 Vienna, Austria
navratil@geoinfo.tuwien.ac.at

Abstract

Cadastral maps are available for large parts of the world. They usually show boundaries between pieces of land owned by different persons. These maps are maintained for centuries and were subject to significant changes in technology and legal background. The geometrical quality of the maps is thus not always known. Deviations between geometry shown in the cadastral records and reality are common and their size is usually unpredictable. The paper presents a framework in which different types of deviations are separated so that they can be addressed separately. First results for the Austrian cadastre are presented for some types of deviations.

Keywords: Cadastre, Geometrical Quality, Quality Assessment, Error Classification.

1 Introduction

Cadastral data sets are in many cases the only large-scale maps available. Some have been created centuries ago and are continuously updated. This keeps the contents correct and in correspondence with reality. Quality of cadastral data sets is an important issue for the land market and sometimes even political analysis (Manson *et al.*, 2009). Assessment of the quality is not easy, though. Jansone presented a class-based approach for cadastral quality but excluded the task of actually classifying the boundaries (Jansone, 2008). Song used legal standards to assess current data quality (Song, 2008) but standards can be misleading. The standard for the Austrian coordinate-based cadastre specifies a maximum tolerance of 15 cm and assumptions for the tax cadastre are in the range of a few meters. In alpine regions, though, deviations of 150 m and more have been detected. Any realistic quality assessment must compare cadastral data with reality. However, since cadastral systems may incorporate data that are a century old, both may be true: That the cadastral data was wrong from the beginning respectively was falsified during its maintenance and that reality changed. During quality assessment both hypothesis should be separated to understanding the changes related to land and land records.

Better quality estimates for the geometry of cadastral data are necessary. Quality estimates may vary throughout the country because unproductive land like mountain tops was of minor importance for the creators of the original cadastre. In addition, at different times various data capture methods have been used and they provide different geometrical quality. All these aspects need to be addressed. The paper presents a framework in which these questions can be analyzed and uses the Austrian cadastre to show possible strategies.

2 Brief History of the Austrian Cadastre

The Austrian cadastre was designed in the early 19th century. From 1817 to 1868 the stable cadastre was created. In 1883 the transition from the stable cadastre to a system with continuous updates was executed (Muggenhuber *et al.*, 2011). The updates were documented in subdivision maps to allow tracing changes back to the original map from the stable cadastre.

The 20th century saw a number of geometrical changes in the cadastral system. A major undertaking was the change of the cadastral reference system. Already in 1909 the change from the plane coordinate systems to Gauß-Krüger-projection has been proposed and the advancement of surveying technology in the 1st world war proved the necessity for a theoretically sound reference system (Rohrer, 1934). The change from the plane coordinate systems to Gauß-Krüger required a recalculation of the reference points, the definition of a new arrangement for cadastral map sheets, and the redrawing of all cadastral maps. Another change in the cadastral mapping affected the scale of the maps. The original cadastral maps were in the scales 1:720, 1:1440. 1:2880, and 1:5760. This practical problem was resolved by using the scales 1:500, 1:1000, 1:2000, and 1:5000, but again required projection and redrawing. Finally, the introduction of the coordinate-based cadastre led to the limitation to mathematically defined boundary lines. Current developments, inspired by experiences with the digitized version of the cadastral map, even try to restrict boundary line segments to straight lines and arcs.

Currently, the Austrian surveying agency BEV (Bundesamt fuer Eich- und Vermessungswesen) is trying to improve the geometrical quality of the cadastral maps. This is typically done locally based on results from resurveys of the reference frame. The changes typically result in shifts of parts of the cadastral map. Unfortunately, since there is no extended survey, these changes result in new distortions in other places.

3 Classification of Deviations

Original cadastral data were subject to a long process. The data were transformed, projected, sporadically redrawn, and finally digitized. Even after the digitalization the data were changed to ‘improve’ the quality of the data. Each of these steps introduced new errors.

However, this is just half the truth. In the period between observing the boundaries and using the observed data not only the data changed. Reality may change, too. Some changes are caused by geophysical processes:

- Plate tectonics is the processes with the largest spatial extent. Earthquakes, as a major result of plate tectonics, may cause shifts in the range of metres as shown by the Japanese earthquake from 2011.
- Land slides may either be slow or abrupt. In the first case land slides gradually change the position of objects connected to the upper parts of the soil. In the second case land slides may complete change the terrain.
- Coastal lines or riverbeds may move. Rivers in flat terrain tend to meander and thus change the boundary of their parcels. The same is true for sand beaches, e.g., in the Netherlands.

- The change of sea level may have a significant impact of boundaries. A change of a few centimetres in height may lead to significant changes of the average shoreline along shallow beaches.

Figure 1 shows a classification of the deviations based on the separation between observational uncertainty and deviations introduced by either data management or change of the reality. A separation between these aspects is necessary because the processes behind the changes are different and require different models.

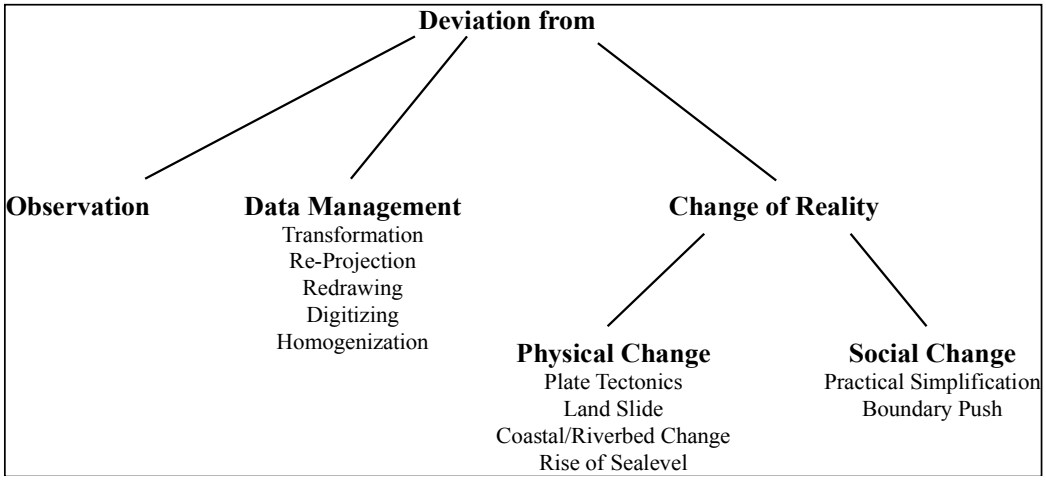


Figure 1. Classification of deviations between cadastral data and reality

The best known aspect is uncertainty in observation processes typically caused by random deviations. Random deviations are the inevitable discrepancies between repeated observations are assumed to be normally distributed and centered around zero. These observations are then eliminated using least squares adjustment (Ghilani, 2010: 104ff). However, there may also be systematic deviations (e.g., caused by adaptation of observation values to the plane coordinate systems) and gross errors like the misidentification of boundary points. Detecting especially gross errors years or even decades after the original survey is challenging because the deviations may result from a change of reality.

Even if the original data is kept correct deviations are inevitable because the position of the boundaries in the real world changes. Physical change includes all deviations resulting from physical processes. They have been already discussed at the beginning of this section. There are changes caused by social interaction, too. Typical examples for such changes are the movement of boundaries by agricultural use, the adaptation of boundaries to simplify the land use, or adverse possession. The detection and classification of these deviations may be difficult, though. Is a simplified boundary line the result of bad placement of a wall or did the land owners agree on the simplification? Years after the change a reliable answer may be difficult.

4 Data Sets Suitable for Testing

A suitable strategy for testing the geometrical quality of cadastral data should be based on a classification like the one in Figure 1. Quality of observations increased over time. Modern measurement equipment has better quality parameters than the measurement tables used in the 19th century. With the use of global navigation satellite systems the geometrical distortions within reference frames are lower than the distortions in reference frames determined with terrestrial observations. Thus the influence of the observations on

the geometrical quality of the cadastral data is significantly smaller in newly created cadastral systems than in old ones. Other influences, however, influence modern and old systems alike. In order to discuss the impacts of different aspects like plate tectonics it is necessary to isolate the deviations caused by this aspect. Some effects may be more visible in some places than in others. Old cadastral systems have the advantage of long-term observational data, which is necessary for slow changes.

The oldest cadastral map in Austria is the 'Urmappe', the result of the original survey. The map sheets are drawn on thick paper and are subject to all problems resulting from the influence on humidity and wear on these maps. In order to archive them, the map sheets have been microfilmed and later scanned. Thus copies of the 'Urmappe' are available and can be used to test the geometrical quality. The 'Urmappe' has an age of between 140 and 190 years. The quality of the original survey, however, is limited since the measurement equipment used for the survey was of a poor quality standard when compared to modern technology.

Changes in cadastral boundaries have been documented in Austria since the end of the 19th century. These documents contain not only the changes in the geometry, but also some measured distances. These distances were originally added to help reconstructing the geometry thereby checking the geometrical correctness. These observations are typically written with centimetre precision. However, sometimes they only have decimetre precision where boundaries are affected by definitional uncertainties (e.g., the shoreline of a lake). Unfortunately, many boundaries have already been changed again because the documented changes are typically in areas that are in development. Since development is an ongoing process, many of these areas may be subject to additional changes. Thus the selection of documents used may be limited.

Finally, the current boundary situation can be observed and compared to the cadastral maps. These resulting deviations are obviously affected by all types of influences separated above. Problems with the original survey, however, should be ignored, when discussing problems of data management or social processes between land owners. A resurvey can provide valuable insights but the different reasons for deviations need to be separated in order to understand the processes that led to these deviations. This separation may be difficult.

5 First Tests and their Results

A first attempt to assess the geometrical quality of cadastral maps tested a similar but less complicated system. In the 19th century a governmental agency was installed to manage the forests owned by the crown. The agency was responsible for maintenance, protection, and economical utilization. This required forest maps showing areas of different age groups. The responsibility for protection also included the protection of boundaries and thus the maps show the spatial extent of the forest parcels and thus show the same geometry as the cadastral maps. The measurements necessary for these maps were taken with similar equipment than used for cadastral maps.

The advantage of the forest maps over the cadastral maps is that the subdivision of the forest parcels was not changed in the last 150 years. Changing the subdivision would be difficult because the different areas are covered with trees of significantly different age and these trees should be cleared at different times. The subdivision shall support the clearing and therefore it remained unchanged. The subdivision is marked in the field by

stone markers and they have not been changed either. Therefore, resurveys of the original maps are easily possible.

Figure 2 shows one of the results of the analysis, in this case the relation between the distance of two points and the deviation of their vectors from the map and the resurvey. It shows that the differences between the vectors are not increasing with the distance between the points.

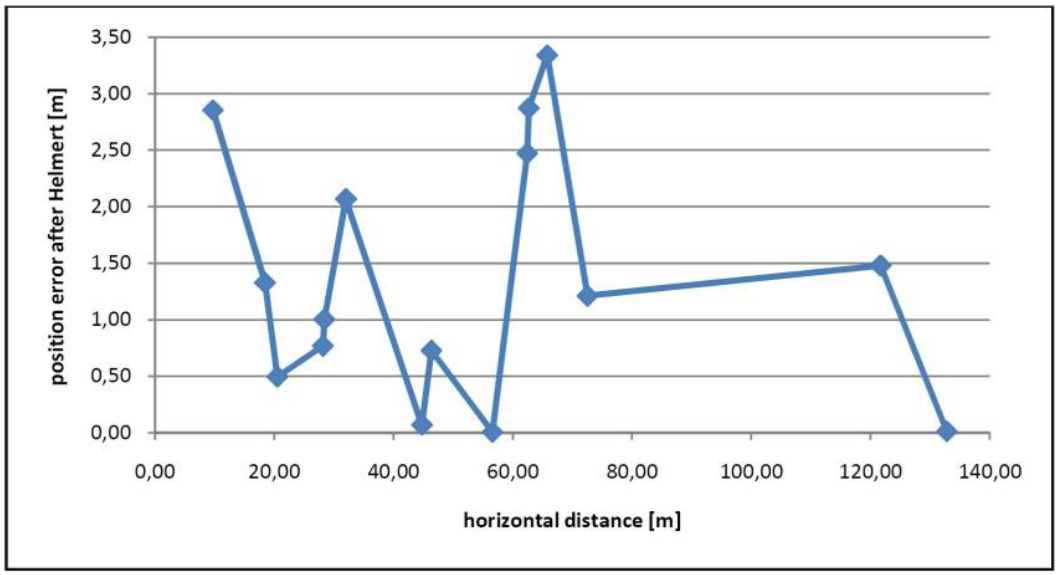


Figure 2. Relation between deviation and distance between the points (Zachhuber, 2009)

The investigation concentrated on finding systematic errors in the original measurements. The selected area south of Vienna was geologically stable and influences from processes like sliding soil were excluded. One of the questions was, for example, if the slope has an influence of the quality of the original observation. However, no significant influences could be detected (Zachhuber, 2009). Thus it seems that the quality of the personnel and the measurement strategy compensated the inferior quality of the measurement equipment. The quality of the data is only determined by the mapping scale.

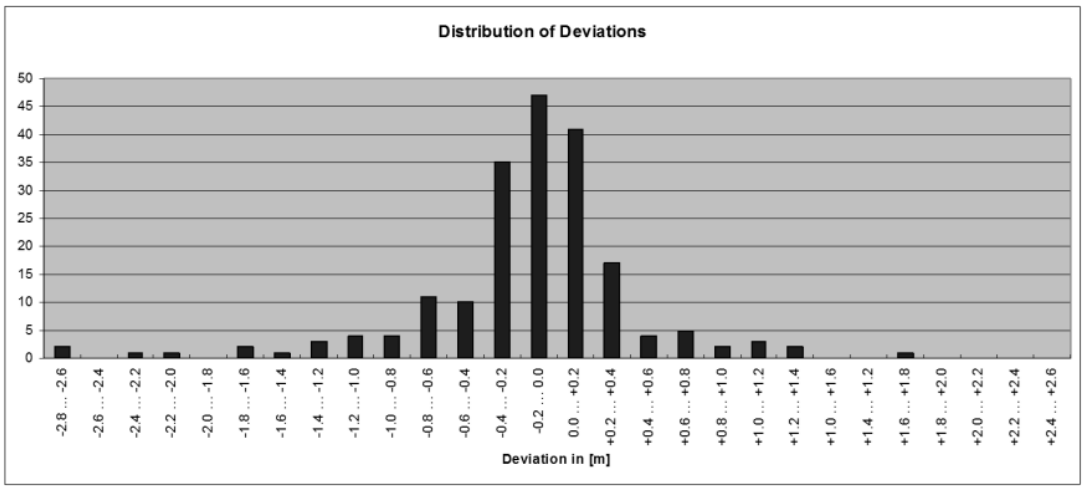


Figure 3. Distribution of deviations between original measurement and distance computed from the digital cadastral map in a test region in Vienna (based on Navratil *et al.*, 2010)

A second approach to assess the quality of cadastral data concentrated on data management issues in the cadastre. The idea was that cadastral updates show the geometry at

the date of survey. These documents do not only include the topology and a graphical representation of the geometry, but also measured distances between specific points, typically between boundary markers. These distances can only change if the representation of the boundary marker locations changes. This again requires an update document. In places where no such documents exist, measurements stored in the document and measurements taken from the current cadastral map should be equal. Differences between these two values must be caused by changes in the map representation, i.e., in one of more of the data management processes. Figure 3 shows the distribution of deviations found in a test area in Vienna. The sample consists of almost 200 observations of at most 100 m.

6 Conclusions and Outlook

The paper presented some first ideas for assessing the geometrical quality of cadastral data. It showed a systematic separation of different influences. These influences can be discussed separately using suitable test data. The resulting statements about geometrical quality are better than what is currently available.

Work on a systematic description of the different processes used for collecting data and managing the maps is in progress. The results should provide a basis to discuss the quality of the processes.

Another approach must include measurements of current boundaries. This can done in two different kinds of areas: Firstly, clearly defined parcels documented in the 20th century shall be resurveyed. The deviations will provide information on the documentation quality itself. Secondly, parcels that are not clearly defined, e.g., in alpine regions, shall be surveyed to investigate movement by social processes.

References

Ghilani, C. D. (2010), *Adjustment Computations*. John Wiley & Sons, Hoboken, New Jersey, 647p.

Jansone, A. (2008), "An Approach to Cadastre Map Quality Evaluation". *In:* K. Elleithy (ed.) *Innovations and Advanced Techniques in Systems*, Springer, pp. 105-110.

Manson, S. M., Sander, H. A., Gosh, D., Oakes, J.M., Orfield, M.W., Craig, W.C., Luce, T.F., Myott, E., Sun, S. (2009), "Parcel Data for Research and Policy". *Geography Compass*, Vol. 3(2): 698-726.

Muggenhuber, G., Navratil, G., Twaroch, Ch., Mansberger, R. (2011), "Development and Potential for Improvements of the Austrian Land Administration System". *In: Proceedings of the FIG Working Week*, Marakkech, Tunesia, FIG.

Navratil, G., Hafner, J., Jilin D. (2010), "Accuracy Determination for the Austrian Digital Cadastral Map (DKM)". *In:* D. Medak, B. Pribicevic, J. Delak (eds.) *Fourth Croatian Congress on Cadastre*, Zagreb, Croatia, Croatian Geodetic Society, pp. 171-181.

Rohrer, H. (1934), "Zum neuen Projektionssystem Österreichs". *Österreichische Zeitschrift für Vermessungswesen*, Vol. 32(5/6): 89-97, 116-123.

Song, W.H. (2008), *Cadastral Map Renovation - An Analysis of the South Corean Perspective*. Master thesis, University of Twente.

Zachhuber, P. (2009), *Investigation of correlations between coordinates from the 'Austrian Federal Forests' and the land surveying office*. Bachelor thesis, Vienna University of Technology.

Data quality of free of charge climate datasets: A comparison of NOAA temperature and precipitation data with validated sources

Péter Zalavári[1] *& Hermann Klug*[2]

[1] Centre for Geoinformatics (Z_GIS), University of Salzburg
Schillerstr. 30, Building 15, 3rd Floor, 5020 Salzburg, Austria
peter.zalavari@sbg.ac.at
[2] Centre for Geoinformatics (Z_GIS), University of Salzburg
Schillerstr. 30, Building 15, 3rd Floor, 5020 Salzburg, Austria
hermann.klug@sbg.ac.at

Abstract

Many researchers have underpinned global climate change affecting especially spatio-temporal changes in precipitation and temperature in the Alps. However, the variability among the degree of changes is as high as the diversity of the datasets available. This paper explores precipitation and temperature datasets available from different sources: free of charge NOAA NCDC (National Oceanic and Atmospheric Administration, National Climatic Data Center) are compared with validated datasets from ZAMG (Austrian Weather Service), Austrian hydrographical surveys, data from the Ministry of Life Science, WorldClim (Global Climate Data), Histalp (Historical Instrumental Climatological Surface Time Series of the greater Alpine Region), HAÖ (Hydrological Atlas of Austria). Having pre- and post-processed the datasets, we use a spatial database to compare the data sources. Analysis have been done on datasets come from equal or closely related stations on a daily, weekly, monthly, and yearly basis. As a result, we conclude that free of charge datasets might be considered for e.g. climate change impact analysis and to be integrated in hydrological models, but one need to take into account the restrictions outlined in this paper.

Keywords: GIS, climate change, Alps, statistics, R

1 Introduction

Analysing the past climate change referring to changes in temperature and precipitation values are necessary to understand and to underpin climate change (Solomon *et al.*, 2007; EEA, 2009). Global climate change impacts on water resources have been reported by numerous researchers (OECD, 2007). The most well-known climate change references on a global scale are the IPCC reports (Solomon *et al.*, 2007) and the Millennium Ecosystem Assessment (Hassan *et al.*, 2005). Even more dramatic are the reported climate induced changes on hydrology in the European Alps, with increasing temperatures twice the global average since the last century (EEA, 2009; Beniston, 2005; OECD, 2007). Local examples from Austria report on decreasing groundwater recharge of 25%

within the last 100 years (Harum *et al.*, 2007). Also in Slovenia the measure of incoming and outgoing water is in a de-creasing trend (Brancelj, 2009).

A growing number of climate change studies as well as applied research require high-quality climate data. But quality proved data from the European national me-teorological services cost a huge amount of money that low budget research studies and NGOs cannot afford. Although there is still a lack of available data and a strong need to better sharing of information, some organizations and research cen-tres provide datasets free of charge. Better data access and information sharing can facilitate the advanced analysis and use of climate data.

To support the above mentioned goal and to estimate the spatial distribution of the precipitation and to analyse water availability different climate data was col-lected from different sources. Using Python scripts the datasets are automatically pre-, and post-processed and stored in a PostgreSQL database (Zalavari *et al.*, 2010). This method facilitates the further use of the data.

Quality control of data – after data collection – is the following step in every re-search study. Validated climate data sets are necessary for any precise and correct research.

Because the NOAA datasets are free of charge and the website doesn't consider too much quality information we decided to compare climate values from NOAA with validated values from the Austrian Weather Service. We form the hypothesis that the daily measurements at one unique station but datasets coming from differ-ent sources share the same value or very close measurements. As a conclusion we will answer our research question about « How good perform the free of charge NOAA datasets in comparison to validated datasets from national data providers? » As a consequence we decide whether or not to use these datasets for spatio-temporal water scarcity analysis in the Alpine environment.

2 Methods

2.1 Data collection

The NOAA repository (http://gis.ncdc.noaa.gov/geoportal/) is a worldwide data pool, intending free and unrestricted access for research purposes, education, and other non-commercial activities for altogether 18 surface meteorological parame-ters (including temperature and precipitation). Historical data are generally avail-able for 9000 stations from 1929 to the present, while data from 1973 to present is almost complete for every station.

The HISTALP database consisting of monthly homogenised temperature, precipitation and other records (air pressure, sunshine, etc.) for the Greater Alpine Region (Auer et. al, 2005) The longest temperature series extend back to 1760, precipitation to 1800. This dataset is a collection of quality improved, long-term instrumental climate data freely available (http://www.zamg.ac.at/histalp/).

The third used dataset in this study is the validated national Austrian Weather Service (ZAMG).

2.2 Statistical Analysis

For the analysis we used a free programming langue and software environment for statistical computing. R is widely used for statistical software development and data analysis.

An evaluation of the quality of the freely available (NOAA, HISTALP) datasets was carried out by comparing the daily measured values and the weekly, monthly, and annual average values against the daily measured ZAMG values and against respectively aggregated values.

Because the HISTALP data is available only on a monthly basis we aggregated monthly and annual mean values in the ZAMG dataset. Analysing the annual temperature between the ZAMG and HISTALP we chose the same station (Salzburg Airport).

3 Results

Having analysed the different datasets we achieved two main results as outlined in the following chapters.

3.1 Comparison between daily values of the ZAMG and NOAA datasets

In the first step we analysed the daily measured values between the same meteorological stations but data acquired from two different data providers: ZAMG and NOAA. As an example shown in Figure 1, the Salzburg Airport station (USAF ID Number 11150; lat: 47.80139, lon: 13.00167) shows some considerable differences up to one degree Celsius. The density of the two series is similar with little differences (Figure 2). Maximum and minimum values are more frequent in the ZAMG series. A difference larger than 2 °C is rare. For most of the months the mean difference are between 0.2 - 0.7 °C. As an example, Table 1 provides the summary statistics for September 2004.

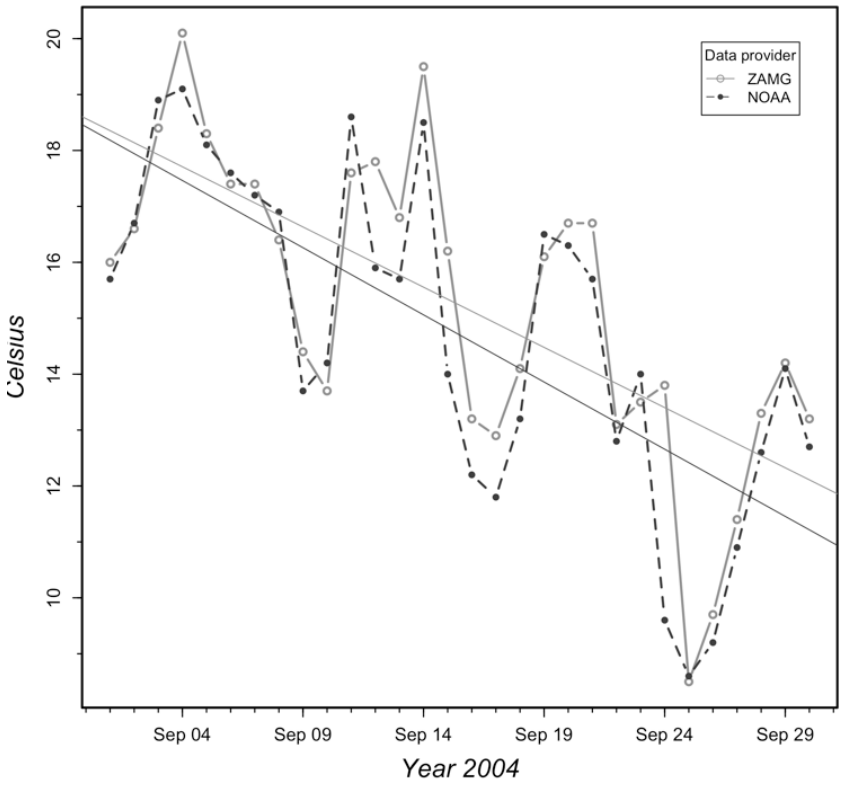


Figure 1. ZAMG vs. NOAA datasets at Salzburg airport

Table 1. Monthly mean values from NOAA and ZAMG for September 2004 at the station

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
ZAMG	8.50	13.35	16.05	15.23	17.25	20.10
NOAA	8.60	12.72	14.95	14.70	16.85	19.10

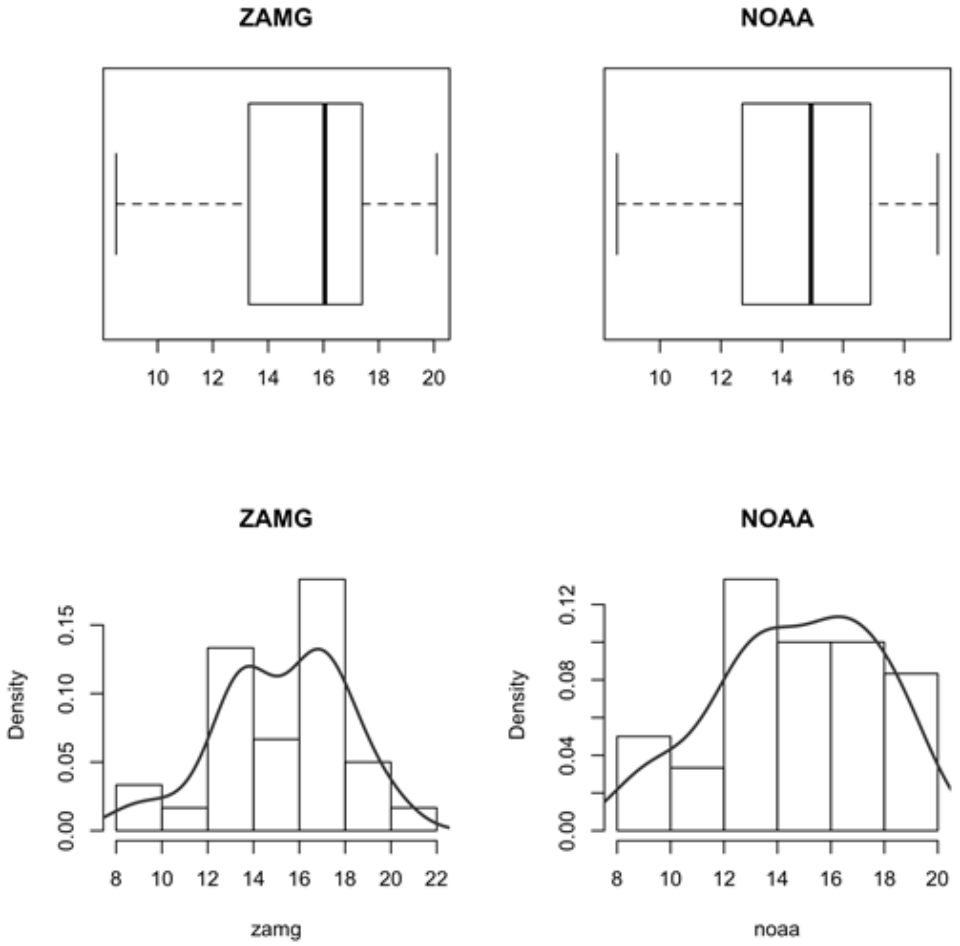


Figure 2. Boxplot and histogram results for 2004 September

3.2 Comparison between annual values of the ZAMG and HISTALP datasets

The differences are larger compared to the daily-based comparison. The annual average temperature changes between 1981 and 2004 are depicted in Figure 3. The figure shows that the trends of the significant increasing temperature are stronger and the values are 1 °C Grad lower with the HISTALP data for most of the year. In terms of the increasing temperature trend, the daily NOAA series show a less significant increasing trend of approximate 0.5 °C. In contrast, the HISTALP series show an increasing trend of more than 1 °C considering the last 20 years.

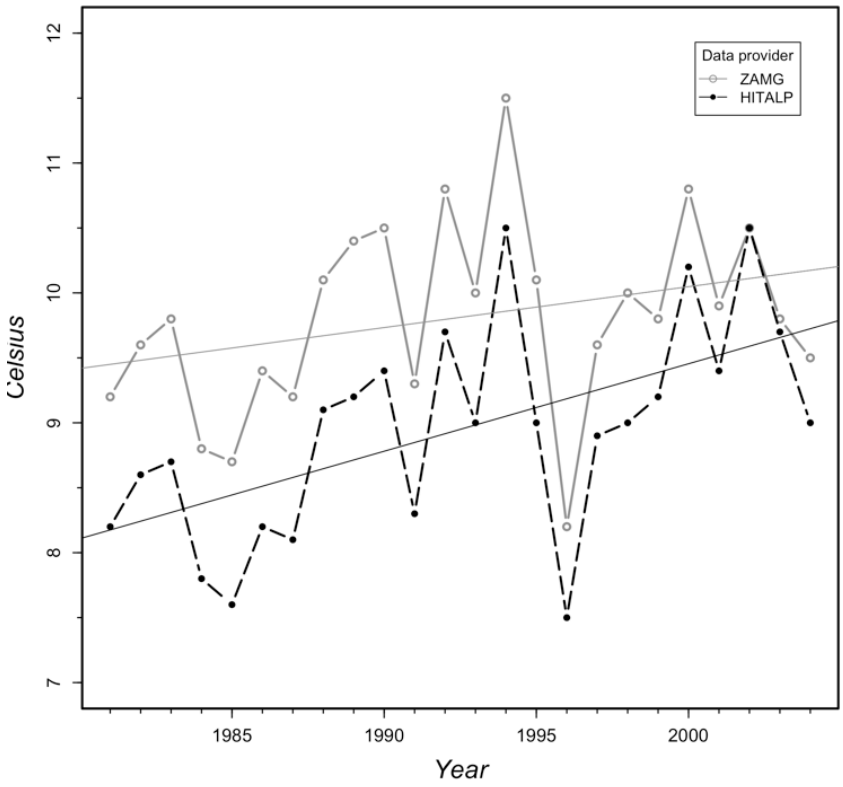


Figure 3. Annual mean temperatures for the Salzburg Airport meteorological station

4 Conclusion and Outlook

The results clearly show some differences between the datasets from the three different sources. It was predictable between the NOAA and HISTALP because of the homogenisation of HISTALP data but unexpected for the same NOAA and ZAMG meteorological station.

The degree of the difference between the ZAMG and HISTALP annual mean temperature values is around 1 °C in the chosen time interval.

Highlighting the benefits of daily resolution and the less variance from the ZAMG data we expect that the NOAA climate time series are adequate for the potential use in climate change impact analysis on the hydrological cycle. However the potential use of this dataset is promising but more study is needed to analyse the correlation between different datasets at a different time interval and place.

Acknowledgments

This research was partially funded by the Interreg IVb project "AlpWaterScarce".

References

Alpine Conference Action Plan (2009), Alpine Conference meeting "Making the Alps an exemplary territory for prevention and adaptation to climate change", 12.03.2009, http://www.alpconv.org/NR/rdonlyres/193D7A9E-0F5E-475D-A48D-E3276F11D292/0/AC_X_B6_en_new_fin.pdf

Beniston (2005), Sensitivity Analysis of Snow Cover to Climate Change Scenarios and Their Impact on Plant Habitats in Alpine Terrain, In: Journal of Climatic Change, pp 299-319.

Brancelj, A. 2009, talk in the framework of the 1st annual meeting of the AlpWaterScarce project, April 27, 2009, Vienna, Austria

EEA 2009, Regional climate change and adaptation. The Alps facing the challenge of changing water resources. EEA Report No 8/2009, ISSN 1725-9177

Harum, T.; Poltnig, W.; Ruch, C.; Freundl, G. and Schlamberger, J. (2007), Variability and trends of groundwater recharge in the last 200 years in a south alpine groundwater system: impact on the water supply. Poster presentation at the International Conference on Managing Alpine Future in Innsbruck, 15–17 October 2007.

Hassan, R., Scholes, R. and Ash, N. (Eds) (2005), Ecosystems and Human Well-Being: Current State and Trends. Island Press

OECD 2007, Climate Change in the European Alps. Adapting winter tourism and natural hazards management. ISBN: 92-64-03168-5

Solomon, S., Qin, D., Manning, M., Marquis M., Averyt, K., Tignor, M.M., Miller, H.L., Chen, Z. (2007), Climate Change 2007. The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Published for the Intergovernmental Panel on Climate Change, ISBN 978-0-521-88009-1

Zalavari, P., Klug, H., Weinke E. (2010), Three steps towards spatially explicit climate change alanysis. In: Car, A., Griesebner, G., Strobel J. (eds.) Geospatial Crosroads @ GI_Forum 2010, Proceedings of the Geoinformatics Forum Salzburg. Heidelberg: Wichmann

Searching for spatial data resources by fitness for use

Ivana Ivánová, Javier Morales, Mesele Atshbeha Gebresilassie & Rolf A. de By

Faculty of Geo-information Science and Earth Observation (ITC),
University of Twente, P.O. Box 217, 7500 AE, Enschede, The Netherlands
ivanova@itc.nl

Abstract

Search for a spatial data resource and determining its fitness for use with respect to the needs of an application is a challenging task for many users, especially for those who are not experts in the GI field. In traditional approaches, users are expected to know what type of spatial data resource they need, and which clearinghouse or geo-portal to search. In case of success, they are still left with the decision on fitness for use, based on complex metadata, for the few cases where such exists. We propose a system for guided search for spatial data resources. This system enhances current search engines with the decision intelligence on fitness for use. The proposed system, a 'guided search engine', works with profiles that contain data about users and relationships among them, and the usage of spatial data resources. With these profiles the proposed system can recommend a spatial data resource that fits the user's needs. We illustrate capabilities of the proposed system on a request by a fictional user of a spatial data resource frequently present in the neo-geography world.

Keywords: fitness for use, spatial dataset, quality, spatial data search, recommendation service

1 Introduction

Let us introduce Garry Mayer. He is an entrepreneur who wants to know which areas in the city of Enschede are suitable for establishing a new restaurant with an existing parking lot. How would Garry find spatial datasets that are fit for his use? Connected to the Internet, he wishes only to type 'unused areas in Enschede with parking space' into his favorite search engine. He expects to receive a spatial dataset with attribute information that he needs (e.g., type of area, owner, amount of parking space).

A spatial dataset that complies with user requirements can be considered of acceptable quality, i.e., it is fit for use. The optimal choice and delivery of spatial datasets to end-users requires a solution to two fundamental problems. The first problem is:

1) How can end-users specify the requirements on spatial datasets that they want to obtain?

Users can provide a detailed list of data quality elements, e.g., as defined by ISO (2002) with acceptable values or ranges thereof. However, to expect a detailed list of values for each data quality element means targeting a very small spatial data quality user group. Most of the time, users only know what the application they want to use the spatial

data for is. In that case, a comparison of default quality requirements with quality metadata of the spatial dataset needs to be automated.

2) How do users assess the fitness for use of a spatial dataset?

Producers of spatial datasets often assume that users are able to determine a spatial dataset's fitness-for-use before using it. They expect users to look at the production quality information and other parts of metadata of the spatial dataset and compare it with the list of quality requirements preferably using the same (standard) quality elements. Users often determine a spatial dataset's fitness for use by comparison of an already selected spatial dataset with some other spatial dataset they may possess (Boin and Hunter, 2009). In case of an 'incorrectly selected' spatial dataset, such evaluation may cause important dissatisfaction.

In this paper, we propose a strategy to target spatial data users that are not experts in spatial data quality and its evaluation. We expect that users (such as Garry Mayer) will appreciate hints or suggestions during spatial data search. We sketch a system that will provide 'guided spatial search' based on reasoning over metadata to determine the spatial dataset's fitness for requested use. In the remainder of the paper, we will first look at currently available means for determining a spatial dataset's fitness for use, next we will discuss current approaches for spatial data search, and we will conclude with a sketch of the 'guided spatial search engine'.

2 Current means of determining spatial dataset's fitness for

Technical committee 211 Geographic information/Geomatics by International organization for standardization (ISO/TC 211) provides standards that define spatial dataset's quality metadata (ISO, 2002; ISO, 2003b), and guidelines for data quality evaluation (ISO, 2003a) with a list of applicable data quality measures (ISO, 2006). ISO/TC 211 defines in (ISO, 2002) five quantitative data quality elements (*positional accuracy*, *thematic accuracy*, *temporal accuracy*, *completeness*, and *logical consistency*) and three overview data quality elements (*lineage*, *purpose*, and *usage*).

Despite detailed definitions, left with *ISO standard* metadata information, users as Garry Mayer may feel lost and overlook quality statements as the whole. To avoid that, Devillers *et al*. (2007) suggested improving the visualization of spatial data by using business intelligence approach and tools. Improved visualization of data quality should help users in deciding on fitness for use. Fisher, *et al.* (2010) suggested re-definition of medatada from: "Data about data" as in (ISO, 2003b) to "information that helps users to assess the usefulness of a dataset relative to their problem" (Fisher, *et al.*, 2010, p.11). This would require enriching technical medatada of a dataset with reports on socio-political context of a dataset's creation, critical reviews of and opinions on the dataset and means for users to provide feedback on the dataset. By thinking along the risks involved in taking decisions based on the use of data that does not fit the application requirements, Hunter and de Bruin (2006) proposed a case study-based approach for users to see consequences of a spatial dataset's misuse.

As Boin and Hunter (2009) confirmed, consumers of spatial data make little or no use of standard spatial data quality metrics. Conclusions drawn from consumer survey on the use of spatial data quality information are in the direction of improved data quality communication carefully considering users' expectations while designing the interface for communicating spatial data quality. Boin and Hunter (2009) suggested, in line with

(Fisher *et al*., 2010), adding to data quality information other user experience with and opinions on the dataset.

3 Search for spatial data resources

With the advent of spatial data infrastructure (SDI) in the 1990s, the search for spatial datasets became a familiar task. Search is labeled by many as one of the critical factors in the path to optimal exploitation of spatial datasets. Through the years many steps have been taken to enable the search, from clearinghouses all the way to catalogue services. In spite of the effort put into facilitating the search, and the significant number of spatial data catalogues readily available, it remains a cumbersome task to find out what resources are available in the web from all types of spatial data providers, and which of the available resources best (or sufficiently enough) suit the user's needs.

There exist a number of reasons for this. For the first part, spatial data search is always fragmented. The user is expected to have prior knowledge on available catalogues. For the second part, the user does not have mechanisms to search and filter resources based on fitness for use, but rather only on availability, which implies location/extent and/or keyword. Also, the search process is manual and unaided. There are no indexed resource descriptions, no knowledge about how a resource has been used or reused, and no comparative reasoning on similar resources. Exposing available spatial datasets to mainstream search engines (or upcoming recommender systems) is not an easy task.

An interesting development in this direction is the addition of spatial and temporal parameters to OpenSearch, the so-called OpenSearch-Geo extensions (Turner, 2007; Fonts *et al.*, 2010). OpenSearch is the collection of technologies for sharing search results in a way that facilitates analysis and reasoning over the results (Clinton, 2005). Still, search for a spatial dataset remains based on extent and keywords, and does not include means for assessment of a dataset's fitness for use.

To make the search process for a spatial dataset more user-oriented (as indicated in Section 2), an active search approach needs to be put in place. An active, 'guided search engine' may have a simple interface, and yet still be capable of searching across all spatial datasets (or their parts) and reasoning over dataset quality and previous dataset usage. As an added value, such 'guided search engine' would use the above capability for recommending a dataset based on the objectives of the user.

We propose to improve the intelligence of spatial data resource search systems by constructing a mechanism that ensures proper interpretation of the dataset's metadata as a response to user requests, and that 'prepares' the decision on dataset's fitness for use.

4 User profile and spatial data resource profile

To help users to decide on a dataset's fitness for a certain use, we suggest embedding the decision intelligence into the spatial data search system. Such a system is based on user and spatial dataset profiles stored by the search engine in a database. The user profile contains information about the user in his process of searching for spatial data resources based on data quality parameters. Components of the user profile are data about *user identity*, *searched spatial extent*, and *user requirements* on datasets. User requirements on datasets can be expressed by the *name of the application* for which the searched dataset will be used, or by *acceptable level of the dataset's quality* using elements defined by

(ISO, 2002). Spatial dataset profile contains link to spatial dataset's metadata, data about its rating, frequency of accessing it, volume of spatial data used (e.g., the whole spatial dataset or its parts), usage of related resources, and the location independent re-use (i.e., search for the same quality spatial dataset in another location). User and spatial dataset profiles are gradually updated using iterations to gather detailed user requirements (see Section 6).

5 Sketch of 'guided search engine'

To provide spatial data resources that are fit for users, a number of components has to be added to the standard search approach. Figure 1 depicts the components and information flows of the so-called 'guided search engine'. An OpenSearch (Turner, 2007) based *search plugin* is embedded in the user's application for asynchronous communication with the *Search Engine*. This plugin allows to obtain user profiles, and also to deliver search results, and to actively track the user interaction with those results. Interaction with the user provides an active and passive long- or short-term (L/S-T) feedback to the engine. The *Search Engine* contains a *recommender*, a *usage tracker* and a store for *profiles* and *L/S-T* feedback, and the actual *search*/*indexer*. The *recommender* keeps track of user behavior and builds a profile of the relationship between users and spatial datasets as follows: user–spatial dataset (rating, activity, and application), spatial dataset–application (spatial dataset quality), and user–user (similarity). This data matrix is exploited in combination with user profiles in a process called collaborative filtering (Griffith *et al*., 2008) to reason over resources and recommend spatial datasets to users. The search results that the recommender delivers to users contain virtual service calls to the actual spatial datasets to keep track of usage of the resource. To support such tracking, a so-called *Virtual Service Interpreter (VSI)* is deployed. VSI receives requests from users and filter information tracking from the actual service request. Once that has been done, the VSI relays the request to the service provider and communicates to the *usage tracker* on the characteristics of the request. The *usage tracker* builds a profile of the usage of a

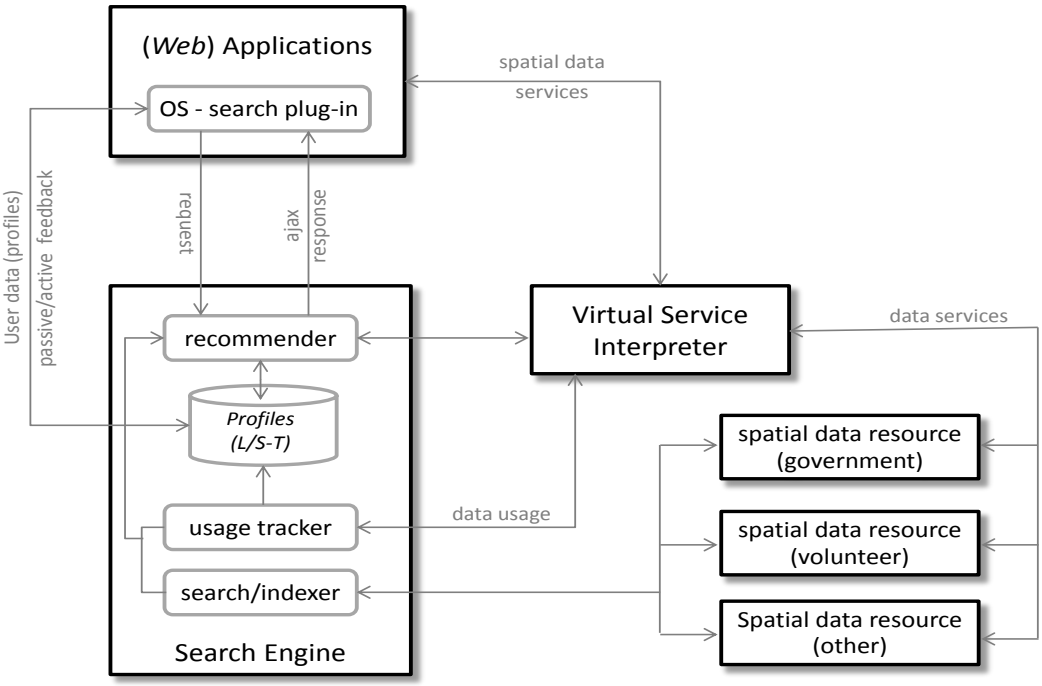


Figure 1 Guided search engine

resource based on the user applications, data quality values and volume of usage. This information is exploited by the *recommender* to refine and filter search results. The *search/indexer* implements the actual search as required by the recommender and actively crawls and indexes spatial data resources using location, content and quality.

The computation of spatial dataset's (or its parts') fitness for use is done using cosine similarity. This computation requires the following inputs: a multi-dimensional user requirements statement and a multi-dimensional metadata record describing existing spatial dataset. These inputs are represented using the bag-of-words (bow) format. A bow U is a set of binary tuples $\{\langle t_1, w_1\rangle, \ldots, \langle t_n, w_n\rangle\}$ where t_i are descriptive terms and w_i are weights that represent the importance of the term in the description. The similarity between bows is determined by comparing the deviation of angles between the vectors that each bow represents. These vectors are obtained by plotting the bows on a N-dimensional vector space model where each term represents an axis. Given two vectors $\vec{U}$ and $\vec{T}$ representing user requirements and a spatial data resource. Their cosine similarity is obtained using the vectors magnitudes and their scalar product as follows:

$$similarity\ (\vec{U},\vec{T}) = \frac{\vec{U}\cdot\vec{T}}{\|\vec{U}\|\times\|\vec{T}\|} \quad \text{where} \quad \|\vec{U}\| = \sqrt{\sum_{i=1}^{n} w_i^2}$$

Resulting similarity values are ranked and the result is presented to the user. As an example of this computation, let us consider Garry Mayer's free text request as the following bow: $P_u = \{\langle Enschede, 2.0\rangle, \langle unused\ areas, 1.0\rangle, \langle parking\ space, 0.5\rangle\}$. To make sense of the request, besides location identification, the first iteration of the similarity analysis is used to deduce the application area related to the request. For this, ISO's topic category list (ISO, 2003b) is used. Weights of terms with respect to the category are determined by computing their degree of closeness using latent semantic analysis (Landauer *et al.*, 2007). The bow for the transportation category is:

$$P_{trans} = \{\langle road, 1.0\rangle, \langle tunnel, 1.0\rangle, \langle vehicle, 1.0\rangle, \langle station, 1.0\rangle, \ldots$$
$$\ldots, \langle building, 0.7\rangle, \langle car\ park, 0.7\rangle, \langle polygon, 0.5\rangle, \langle area, 0.3\rangle, \ldots\}$$

The similarity values resulting from the intersection check between the users request bow and the various category bows (*i.e.*, $P_u \cap P_{trans} \neq \emptyset$). Once the application area is identified, the search engine deploys the default profile (including quality values) values for that application and interacts again with the user. This iterative process results in a spatial dataset that fits the application derived from Garry's input free text request. A subsequent iteration refines the search based on default data quality requirements associated with the application. Once the spatial dataset is picked, the search session is recorded by updating user and spatial dataset profiles, as described in Section 4.

6 Conclusion

Organized geo-portals expect minimum metadata information attached to a spatial data resource. However, we cannot assume data quality being explicit all the time in all spatial data resources available. We propose a 'guided search engine' that fills the gap between application experts (such as Garry Mayer) and geo-information experts (those that understand ISO/TC211 quality definitions). Our approach enables to gradually build quality requirements for various applications from user tracking and feedback.

End-users want an information product, and if such product does not exist, it may be produced from existing spatial data resources. This can be a dataset or a computing pro-

cess. A computing process needs resources to produce output, which itself is a dataset. We believe that the 'guided search engine' proposed can also accommodate geo-information expert users or web processing services as users. We recognize that rating parameters and values for tracking user satisfaction and experience with spatial data resource can be an implementation challenge.

References

Boin, A.T. and Hunter, G. J. (2009) What communicates quality to the spatial data consumer?, *In*: Stein, A., Bijker, W. and Shi, W. (eds.). *Quality aspects in spatial data mining*, CRC Press, Boca Raton, USA, 2009, pp. 285-296

Clinton, D. (2005). *OpenSearch Specification*. OpenSearch.org, Version 1.1, http://www.opensearch.org/Specifications/OpenSearch/1.1 (accessed on: 28. March, 2011)

Devillers, R. Bédard, Y. Jeansoulin and R. Moulin, B. (2007). Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. *International journal of geographical information* science, Vol. 21(3): 261-283, 2007

Fisher, P., Comber, A. and Wadsworth, R. (2010). What's in a Name? Semantics, Standards and Data Quality, *In*: Devillers, R. and Goodchild, H. (eds.). *Spatial Data Quality. From Process to Decissions*, CRC Press, Boca Raton, USA, 2010, pp. 3-16

Fonts, O., Huerta, J., Díaz, L. and Granell, C. (2010). OpenSearch-geo: The simple standard for geographic web search engines. *In*: IV Jornadas SIG Libre, 10-12 March, 2010, http://geoportal.dlsi.uji.es/OpenSearch/docs/SIGLibre-OpenSearch-article.pdf (accessed on: 23, June, 2011)

Griffith, J., O'Riordan, C. and Sorensen, H. (2008). Identifying and Analyzing User Model Information from Collaborative Filtering Datasets, *In* Uchyigit, G. and Ma, M. Y., (eds.). *Personalization Techniques and Recommender Systems*. World Scientific Publishing, New Jersey, USA, 2008, pp. 165-188

Hunter, G.J. and de Bruin, S. (2006). A Case study in the use of risk management to assess decision quality. *In*: Devillers, R. and Jeansoulin, R. (eds.). *Fundamentals of spatial data quality*, ISTE Ltd., London, UK, 2006, pp. 271-282

International Organization for Standardization (ISO) (2002), *ISO 19113:2002 Geographic information – Quality principles*, ISO, Geneva, Switzerland, 29p.

ISO, (2003a). *ISO 19114: 2003 Geographic information – Quality evaluation procedures*, ISO, Geneva, Switzerland, 63p.

ISO (2003b) *ISO 19115: 2003 Geographic information – Metadata*, ISO, Geneva, Switzerland, 140p.

ISO (2006) *ISO/TS 19138: 2006 Geographic information – Data quality measures*, ISO, Geneva, Switzerland, 68p.

Landauer, T. K., McNamara, D. S., Dennis, S. and Kintsch, W. (2007). *Handbook of latent semantic analysis*. Lawrence Erlbaum Associates.

Turner, A (2007). *OpenSearch-Geo Extension*. Version 1.0, http://www.opensearch.org/Specifications/OpenSearch/Extensions/Geo/1.0/Draft_2 (accessed on 23, June, 2011)

GEOVIQUA: a FP7 scientific project to promote spatial data quality usability: metadata, search and visualization

Joan Masó[1], Ivette Serral[1] & Xavier Pons[2]

[1] Centre for Ecological Research and Forestry Applications (CREAF), C-Building, Universitat Autònoma de Barcelona. 08193 Cerdanyola del Vallès. Spain
Joan.Maso@uab.cat; ivette@creaf.uab.cat
[2] Department of Geography, Universitat Autònoma de Barcelona. 08193 Cerdanyola del Vallès. Spain.
Xavier.Pons@uab.cat

Abstract

GeoViQua is a recently started FP7 project (ENV.2010.4.1.2-2; nr 265178) focused on adding rigorous quality specifications to the Global Earth Observation System of Systems (GEOSS) spatial data in order to improve reliability in scientific studies and policy decision making. Quality visualization and search tools will be integrated to the public GEOPortal to assist in the spatial data discrimination for scientific purposes; the project will also contribute to the definition of a GEOLabel concept reflecting scientific relevance, quality, acceptance and societal needs. To achieve all this, several pilot cases spread over the whole Earth Observation chain are performed, considering remote sensing acquisition and data processing, and its application to the main GEO (Group on Earth Observations) Societal Benefit Areas: climate, biodiversity, ecosystems, agriculture, water, etc. Other data and methods are considered as well, such as interpolation techniques and their quality implications on cartographic products.

Keywords: Spatial quality, visualization, geo-search, GEOLabel, Earth Observation, SBA.

1 Introduction

GeoViQua, the contraction of GEOSS Visualization and Quality, is a 7th Framework Project initiative funded by the European Commission under the ENV.2010.4.1.2-2 theme - Integrating new data visualisation approaches of earth Systems into GEOSS development. GeoViQua started on February 2011 and it is expected to be finished by 2014 with the aim of adding spatial data quality focusing on its elicitation, visualisation and search, under the umbrella of GEOSS, the Global Earth Observation System of Systems. GEOSS is coordinated by the Group on Earth Observations, GEO (GEO, 2005), a public infrastructure that interconnects a diverse and growing array of instruments and systems to allow monitoring and forecasting changes in the global environment. GEOSS supports policy-makers, resource managers, scientific researchers and many other experts and decision makers in their daily work with Earth Observation (EO) data. The GEOSS Common Infrastructure (GCI) provides a clearinghouse (GEOSS registry and data cata-

logue) and a GEOPortal to discover and visualize data in integrated, standardize and interactive way allowing it to be broadly used by the scientific community when dealing with representation and modelling of Earth Systems. GCI is composed of standardised components tested in the Open Geospatial Consortium (OGC) Architecture Implementation Pilots (AIP) (Husar, 2008). The GCI is subject to continuous development and improvement of its functionality, from infrastructure and back-end services to context-driven applications and human-computer interfaces. A major target for this development is the wide variety of Societal Benefit Areas (SBA) that GEOSS addresses: Health, Disasters, Weather, Energy, Water, Climate, Agriculture, Ecology and Biodiversity (GEO, 2005).

GeoViQua main objective is to improve the GCI providing the user community with innovative quality-aware visualisation and advanced geo-search capabilities making them available through the GEOPortal and other end-user implementations. To this end, GeoViQua will attach standard quality parameters to the current metadata making it available to users and experts, producing more reliable studies about Earth systems and their dynamics, and tagging spatial information by means of a quality label: the GEOLabel.

To achieve all tasks, GeoViQua is supported by a group of 10 partners: Catalan, French, Italian and German research centres (CREAF, CEA-LSCE, CNR-IMAA, Fraunhofer-IGD, 52North); a Catalan and two English Universities (Universitat Autònoma de Barcelona -UAB-, Aston University and University of Reading); a Dutch company (S[&]t); and the European Space Agency (ESA).

2 Procedure

2.1 Quality elicitation

Many datasets provided through GEOSS and other systems lack the quality information needed to allow users to decide about their fitness for a purpose. This information has then to be derived from:

- Encouraging major data producers to provide information about data provenance in human or machine readable forms.
- Developing tools to process and compare datasets, as “ground-truth” remotely-sensed data with in-situ observations, thereby deriving quantitative information about data uncertainty.
- Developing tools and procedures to allow users of datasets to provide feedback about the utility of datasets for their purpose.

In GeoViQua, information about data quality will be extracted from metadata, data itself, validation processes with in-situ sensors, provenance information and user feedback. Current and extended standards over data quality description will be used or developed to define ‘quality indicators’, including quality and provenance parameters as those proposed by the GEO strategy on data quality: the Quality Assurance Framework for Earth Observation -QA4EO- (QA4EO task team, 2010).

2.2 Standards quality codification

Currently, most spatial data producers register information about data quality along with metadata. However, this information remains largely ignored by data users, as not all geodata end users are spatial data experts, leading to the risk of making poor decisions based on spatial data. The usability of quality information is hindered by the lack of

simple, powerful and appropriate tools that can effectively manage, store, manipulate, query, update or display quality information.

The metadata model in ISO19115 allows data quality to be encoded using the DQ_DataQuality element, which contains several fields, as a DQ_Element containing DQ_QuantitativeResult elements. The standard also allows conformance results and lineage information. GeoViQua will use the ISO19157 model (that was previously in ISO19115:2003) as the starting point for representing data quality, although it will be necessary to propose some extensions to allow data quality to be linked with results on a pixel-by-pixel basis where this can be realistically done. Best practices examples to illustrate how to employ the extended ISO19157 focussing on quantitative quality indicators will be developed.

UncertML, developed by Aston University, is a proposed standard that provides a well defined encoding for uncertain quantities, and a mechanism for weak typed approaches through the use of dictionaries. UncertML includes support for all commonly used quality indicators and aligns with QA4EO focus on rigorously defined statistical representation of data quality (Williams *et al*., 2008).

GeoViQua will extend UncertML to allow descriptions of further types of uncertainty, including qualitative descriptions consistent with ISO19157, and promoting the use of UncertML within the EO context to become a best practice paper in OGC, and exploring its standardisation within the ISO process.

2.3 Quality description embedded and linked with data

Currently, information about quality is usually stored separately from data (*e.g.*, in an XML metadata document in a metadata catalogue, as occurs in the current GEOSS catalogues). GeoViQua will connect such quality metadata with the dataset and with the services that provide data visualisation to allow quality information to be used, propagated and re-documented in any derived product.

Spatial data quality becomes more critical when several geoprocessing services are chained (Masó *et al*., 2011). One of the key deficits of current approaches to data quality propagation in geospatial processes is that they fail to consider provenance information - the sequence of processes by which a data has been prepared for usage (for example, resample, interpolation, line generalisation or atmospheric correction). Data quality may be acquired in a static state by quantitative or qualitative testing or in an operational state by tracking the processing steps that have been applied. While in evaluating the static state one is concerned about the accuracy (in terms of extent and magnitude of errors) and other quality components that can be measured or quantified, the operational state evaluation aims to understand data quality by reviewing the provenance. The operational state is then concerned with evaluating the current version of data in relation to the original sources and processing chain. A provenance record can be invoked to ascertain, for instance, when, how and where data was produced; dataset derivation information; types of resources used; workflow of data derivation, etc. Provenance metadata entries can also be used to provide information on run time, resource consumption, computational anomalies, and past executed workflows.

2.4 Quality visualisation and search

Visualisation of spatial data is a very important process for end users to assess data suitability for specific targets. Simplistic visualisation tools often do not provide enough

information about quality for a user to make a decision. Users also need to discover new datasets, and quality information is extremely important in this context to allow users to find the best dataset for their purpose (van der Wel *et al*., 1994). Within GeoViQua, graphical representation of metadata quality parameters will be developed helping the user to know data collection structure and patterns and thus to easily mine data.

Geo-searches usually produce multiple results: users are therefore lost in assessing which datasets are the best suitable for them in terms of quality. GeoViQua will ensure that users are able to search in datasets conform to certain quality standards, (*e.g.*, limiting search results to those datasets that are associated with a GEOLabel). Query by location with metadata, quality statistical charts, and quality representations through symbolization are some of the techniques to be explored. Search results will be linked to relevant quality information so users can merge, prioritise and threshold quality metrics to build specific, re-usable models of ”fitness-for use”.

These data visualization and search strategies, related to data quality information, will be integrated, through existing standards, on the GEOPortal to be accessible to everyone, as well as on mass market "Google Earth-like" map tools and other 3D viewers and mobile devices.

2.5 The GEOLabel definition and application

Within GeoViQua a spatial quality indicator, GEOLabel, will be defined to ensure at maximum users knowledge about quality when using spatial data, an issue particularly critical in scientific studies.

GEOLabel requirements will be firstly mined and determined, integrated into all GeoViQua components, validated and applied by the pilot cases, and finally disseminated to the community. This theoretical and methodological definition will be complemented in collaboration with the FP7 EGIDA project and the GEO task ST-09-02 committee, responsible of promoting GEOLabel in GEOSS.

2.6 From theory to reality: pilot cases assistance

GeoViQua sets seven PC to cover a variety of techniques, source data, spatial resolutions (from global to local), time resolutions (from near real time to decades) and GEO SBA among other useful cases (Table 1). This broad range of applications will enable to access much-needed contextual expertise and information on real-world challenges, and promote constant dissemination of GeoViQua research and development activities. These considered PC include the following scenarios:

1. *Agriculture*. Provides local collection of data on semiautomatic classification of agricultural irrigated crops. The Joint Research Centre provides data from LUCAS (Land Use/Cover Area frame Statistical Survey) and Eurostat (UAB, JRC). (Serra *et al*., 2009)
2. *Global Carbon Cycle*. Offers the opportunity to develop and integrate new data visualisation approaches of carbon fluxes in GEOSS including crucial quality data lacking in most current carbon portals. (CEA-LSCE). (Peylin *et al*., 2002)
3. *Climate*. Combines numerical models with observations, requiring a deep understanding of the uncertainties inherent in both approaches. (University of Reading, UAB). (Pons and Ninyerola, 2008)
4. *Air quality*. Focuses on the provision of information from in-situ air quality sensors (PM_{10} and/or NO_2) to mobile users. (52North). (Pebesma *et al*., 2007)

5. *Land Use*. Addresses the relation between landscape dynamics and biodiversity for several SBA. Errors and uncertainties propagate when combining data from different sources, spatial resolutions, dates and transect models, when categories do not nest or correspond, etc. (CREAF). (Krauss *et al*. 2010)
6. *Water Cycle*. Contributes, applying EO methodologies in the whole water cycle, to the sustainable management of fisheries and aquaculture. (UAB, University of Reading, ESA). (Díaz-Delgado *et al*. 2010)
7. *Remote Sensing chain*. Provides Landsat series processed images, from row images to final products, helping to validate GeoViQua proposals and aiding to deepen inside the knowledge about the organization and use of metadata and data to improve tools that deal with provenance. (ESA, UAB). (Serra *et al*., 2003)

Table 1. Pilot cases classification, where 1 stands for Agriculture, 2 for Global Carbon Cycle, 3 for Climate, 4 for Air Quality, 5 for Land Use, 6 for Water Cycle, and 7 for Remote Sensing chain.

	Technique									**Source**					**Societal Benefit Area**							**Time**				**Scale**		
Pilot Cases	Detection	Digitizing	Classification	Series comparison	Generalization	Interpolations	On field validation	Summaries	Numerical models	Raw Satellite	In-situ sensors	Aerial photograph	GIS, DEM	Satellite products	Health	Climate	Water	Weather	Ecosystems	Agriculture	Biodiversity	Near real time	Daily-weekly	Annual	Multiyear	Global	Regional	Local
1		✓	✓				✓			✓		✓					✓			✓				✓			✓	✓
2				✓		✓		✓			✓			✓		✓			✓		✓			✓		✓		
3				✓		✓	✓	✓	✓	✓	✓		✓			✓		✓					✓	✓	✓	✓	✓	
4	✓					✓	✓				✓				✓							✓					✓	✓
5		✓		✓	✓							✓							✓		✓			✓			✓	✓
6	✓				✓			✓		✓	✓			✓	✓	✓	✓			✓	✓			✓	✓		✓	✓
7	✓		✓						✓	✓					transversal								✓				✓	✓

4 Conclusion

- Based on user requirements, GeoViQua will formalize and implement a methodology to extract, encode, embed and link data quality information with data.
- It will provide technology and implementation foreground to the quality assurance protocols proposed by QA4EO and deliver interfaces to visualise quality related parameters in a range of data visualisation tools, including those serving mobile devices. It will also provide quality enable geo-search strategies.
- Such tools and web services will become part of the European contribution to GEOSS being integrated in the GCI architecture, and will follow the evolving directives on standards as given by the OGC and the IEEE Standards Interoperability Forum (SIF).

Acknowledgments

Authors of this communication thank the support of the European Commission through the FP7- 265178- GeoViQua (ENV.2010.4.1.2-2). Xavier Pons is recipient of an

ICREA Academia Excellence in Research grant (2011-2015). Ivette Serral is recipient of a MICINN-PTA grant from the Spanish Ministerio de Ciencia e innovación (2011-2014).

References

Christian, E. J. (2008), "GEOSS Architecture Principles and the GEOSS Clearinghouse". *IEEE Systems Journal.* 2: 333-337

Díaz-Delgado, R., Ameztoy, I., Cristóbal, J., Bustamante, J. (2010), "Long time series of Landsat images to reconstruct river surface temperature and turbidity regimes of Guadalquivir Estuary". *In:* Int. Geosc. & Rem. Sens.Symposium (IGARSS): 233-236

GEO (2005), "The Global Earth Observation System of Systems 10-Year Implementation Plan". [www.earthobservations.org/docs/10-Year%20Implementation%20Plan.pdf]

Husar, R.B., Hoijarvi, K., Falke, S.R., Robinson, E.M., Percivall, GS (2008), "DataFed; An Architecture for Federating Atmospheric Data for GEOSS". *IEEE Syst. J.* 2(3). 333-337

Krauss, J., Bommarco, R., Guardiola, M., Heikkinen, R.K., Helm, A., Kuussaari, M., Lindborg, R., Öckinger, E., Pärtel, M., Pino, J., Pöyry, J., Raatikainen, K.M., Sang, A., Stefanescu, C., Teder, T., Zobel, M., Steffan-Dewenter, I. (2010), "Habitat fragmentation causes immediate and time-delayed biodiversity loss at different trophic levels". *Ecology Letters*. 13: 597-605

Masó, J., Pons, X., Schäffer, B., Foerster, T., Lucchi R. (2011), "Haiti Earthquake: Harmonizing post-event distributed data processing". *IEEE EarthZine*. 3 [www.earthzine.org/2011/03/18/haiti-earthquake-harmonizing-post-event-distributed-data-processing]

Pebesma, E., De Jong, K., Briggs, D. (2007), "Interactive visualisation of uncertain spatial and spatiotemporal data under different scenarios: an air quality example". *International Journal of Geographical Information Science*. 21:515–527

Peylin, P., Baker, D., Sarmiento, J., Ciais, P., Bousquet, P. (2002), "Influence of transport uncertainty on annual mean and seasonal inversions of atmospheric CO2 data". *Journal of Geophysical Research-Atmospheres*. 107. Article number 4385

Pons, X., Ninyerola, M. (2008), "Mapping a topographic global solar radiation model implemented in a GIS and refined with ground data". *Int. J. of Climatology*. 28: 1821–1834

QA4EO task team (2010), "A Quality Assurance Framework for Earth Observation: Principles". [Online, Available http://qa4eo.org/docs/QA4EO_Principles_v4.0.pdf]

Serra, P., Pons, X., Saurí, D. (2003), "Post-classification change detection with data from different sensors: Some accuracy considerations". *Int. J.of Rem. Sens*.. 24: 3311-3340

Serra, P., Moré, G. Pons, X. (2009), "Thematic accuracy consequences in cadastre land-cover enrichment from a pixel and from a polygon perspective". *PE&RS*. 75: 1441–1449.

van der Wel, F. J. M., Hootsman, M. R., Ormeling, F. (1994), "Visualisation of data quality". *In:* MacEachren, A. M., and Taylor, D. R. F. (eds.). *Visualisation in Modern Cartography*, Pergamon, London, U.K. p. 313-331.

Williams, M., Cornford, D., Bastin, L. (2008) "Describing and Communicating Uncertainty within the Semantic Web". *In:* Uncertainty Reasoning for the Semantic Web Workshop, 7th International Semantic Web Conference, 26 October 2008, Karlsruhe, Germany.

Gas pipeline route selection using Dempster Shafer theory of evidence

Zahra Bahramian[1] & Mahmoud R. Delavar[2]

[1]PhD. Student, GIS Division, Department of Surveying and Geomatic Eng., College of Eng., University of Tehran, Tehran, Iran.
zbahramian@ut.ac.ir
[2]Center of Excellence in Geomatic Eng. and Disaster Management, Department of Surveying and Geomatic Eng., College of Eng., University of Tehran, Tehran, Iran.
mdelavar@ut.ac.ir

Abstract

Pipeline route selection is a critical issue because of its environmental, social and economical impacts on the region. It is a multi criteria decision making (MCDM) approach based on weighting the criteria by experts. Since it is not expected that decision makers have enough knowledge regarding whole aspects of the problem and effective criteria, therefore the knowledge-based pipeline route selection is an uncertain problem. Uncertainty in experts' viewpoints must be handled to select the least cost pipeline route. In this research, Dempster-Shafer theory (DST) of evidence was used to handle the uncertainty in the least cost pipeline route selection. DST allows the expression of ignorance in uncertainty management and the Dempster rule of combination provides an important approach to aggregate indirect evidence and incomplete information. Also, a sensitivity analysis was performed using Monte Carlo simulation by changing the experts' allocated weights. The reliability of the proposed approach was successfully verified.

Keywords: Dempster Shafer theory of evidence, Dempster's rule of combination, pipeline route selection, multi criteria decision making, uncertainty.

1 Introduction

Pipeline route selection is a critical issue because of its environmental, social and economical impacts on the regions. It can be considered as a multi criteria decision making problems. Some studies have been undertaken for the optimum pipeline routing using some selected criteria (Delavar and Naghibi, 2003; Nataraj, 2005; Restriction regulations of natural gas pipelines in the vicinity of buildings and facilities, roads, transmission power lines, railroads and oil pipelines, 2006; Yildirim and Yomralioglu, 2007; Matos *et al.*, 2008). However, none of these studies consider uncertainty and heterogeneity between experts' viewpoints for pipeline route selection. On the other hand, pipeline route selection is a MCDM problem based on weighting the criteria allocated by the experts. However, since it is not expected that decision makers have enough knowledge regarding whole aspects of the problem, the knowledge-based route selection is an uncertain problem. Therefore, DST can be used to handle this kind of uncertainties in the MCDM problem. Some studies have been done to handle uncertainty using DST (Amiri *et al.*, 2007;

Jahankhah, 2010; Sentz and Ferson, 2002; Tangestani and Moore, 2002). In this research, according to the criteria weighted by DST of the evidence, the least cost gas pipeline route is selected from Orumiyeh to Maku in the Western Azarbayejan Province, Iran. The method has been successfully implemented for determination of the least cost gas pipeline route.

The structure of the rest of this paper is as follows: Section 2 provides an introduction to DST and Dempster's rule of combination. Section 3 demonstrates the proposed approach and its implementation. Section 4 draws the conclusions.

2 Dempster Shafer theory of evidence and Dempster's rule of combination

The Dempster-Shafer belief structure was introduced by Dempster (1967) and by Shafer (1976). Dempster-Shafer theory, an extension of Bayesian probability theory, allows for the expression of ignorance in uncertainty management (Gordon and Shortliffe, 1985; Lee *et al.*, 1987). Let X be the universal set: the set representing all possible states of a system under consideration. The power set 2^X is the set of all subsets of X, including the empty set. The theory of evidence assigns a belief mass to each element of the power set. Formally, a function m is called a basic belief assignment (BBA) (Shafer, 1976):

$$m:2^X \rightarrow [0,1] \tag{1}$$

where it has two properties (Shafer, 1976): The mass of the empty set is zero and the masses of the remaining members of the power set add up to a total of 1:

$$m(\emptyset) = 0 \tag{2}$$

$$\sum_{A \in 2^X} m(A) = 1 \tag{3}$$

The mass $m(A)$ of A, a given member of the power set, expresses the proportion of all relevant and available evidence that supports the claim that the actual state belongs to A but to no particular subset of A.

From the mass assignments, the upper and lower bounds of a probability interval, plausibility measure $pl(A)$ and belief measure $bel(A)$, can be defined as follow (Shafer, 1976):

$$bel(A) \leq p(A) \leq pl(A) \tag{4}$$

$$bel(A) = \sum_{B|B \subseteq A} m(B) \tag{5}$$

$$pl(A) = \sum_{B|B \cap A \neq \emptyset} m(B) \tag{6}$$

$$pl(A) = 1 - bel(2^X - A) \tag{7}$$

where $p(A)$ is the probability of A. While belief represents the degree of hard evidence in support of a hypothesis, plausibility indicates the degree to which the conditions appear to be right for that hypothesis, even though hard evidence is lacking. The range between the two is called the belief interval, and represents the degree of uncertainty in establishing the presence or absence of that hypothesis (Eastman, 2001).

DST can combines evidence from different sources and arrive at a degree of belief (represented by a belief function) that takes into account all the available evidences. The Dempster's rule of combination provides an important approach to aggregate indirect

evidence and incomplete information. The combination (called the joint mass) is calculated from the two sets of masses m_1 and m_2 (Shafer, 1976):

$$m_{1,2}(\emptyset) = 0 \tag{8}$$

$$m_{1,2}(C) = (m_1 \oplus m_2)(C) = \frac{1}{1-K}\sum_{A\cap B=C\neq\emptyset} m_1(A)m_2(B) \tag{9}$$

$$K=\sum_{A\cap B=\emptyset} m_1(A)m_2(B) \tag{10}$$

where K is a measure of the amount of conflict between the two mass sets.

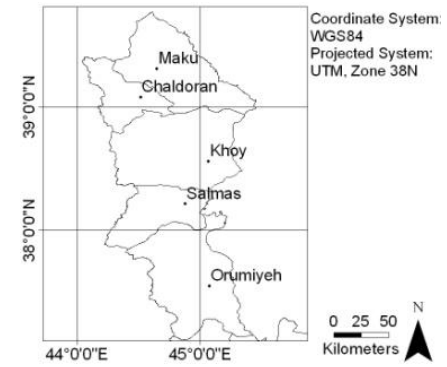
3 Methodology and case study

In the proposed approach, to select the least cost gas pipeline route, at first, the influential criteria were selected and the required data were collected and prepared. The weights of the criteria according to the experts' opinions were determined. Then, Dempster combination rule was used to combine experts' viewpoints and take into account all the available evidences. The least cost gas pipeline route was obtained by combination of input layers using their computed weights. At the end, a sensitivity analysis was performed using Monte Carlo simulation by changing the experts' allocated weights.

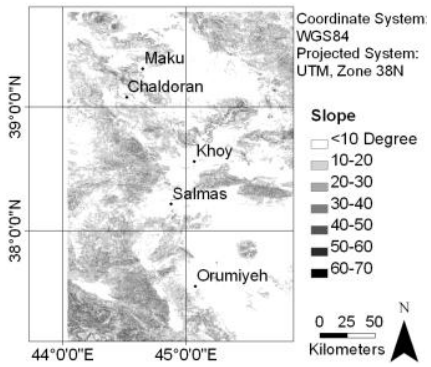
In this research, the optimum gas pipeline route from Orumiyeh to Maku in Western Azarbayejan state, Iran, was determined (Figure 1a). Western Azarbayejan Province has an area of 37059 km^2, occupying 2.25% of the Iranian territory. It has 5319921 m gas pipeline route and uses about $3200*10^6$ m^3 natural gas per year. Orumiyeh and Maku are the first and third cities in this Province in terms of area.

According to the similar research (Iranian restriction regulations of natural gas pipelines, 2006 ; Delavar and Naghibi, 2003; Matos *et al.*, 2008; Nataraj, 2005; Yildirim and Yomralioglu, 2007), six criteria were considered in the computation process including slope, and distances from urban areas, roads, rivers, faults and mines. Therefore, six input map layers were used including urban areas, roads, rivers, faults and mines obtained from National Geoscience Database of Iran (www.ngdir.ir) and slope that was derived from SRTM DEM with 90 m grid dimension.

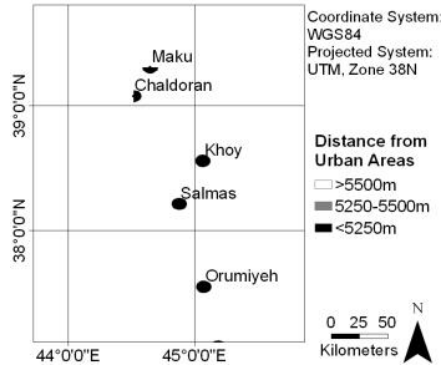
The next step was the classification of the study area according to the criteria. The slope was classified per 10 degree (Figure 1b). According to Iranian restriction regulations of natural gas pipelines (2006), gas pipeline route cannot be located nearer than 250 m of urban areas and the studied area must be classified to two classes in term of suitability. In order to have more certainty about our selection, the study area was classified to three parts including areas closer than 250 m (threshold 1), between 250 m and 500 m (threshold 2) and more than 500 m from urban areas (Figure 1c). The employed thresholds 1 for roads (Type 1, 2 and 3), rivers, faults and mine are 60, 50, 30, 250, 250 and 250 m, respectively and thresholds 2 are twice the thresholds 1. For other layers, similar operations were done (Figures 1d, e, f and g).



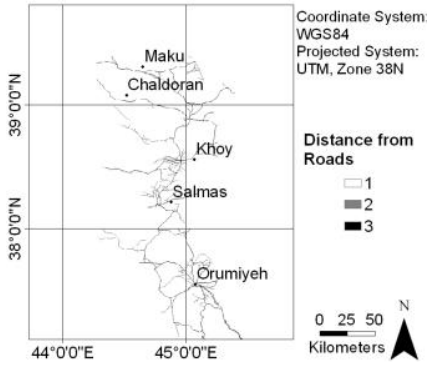
(a) Study area



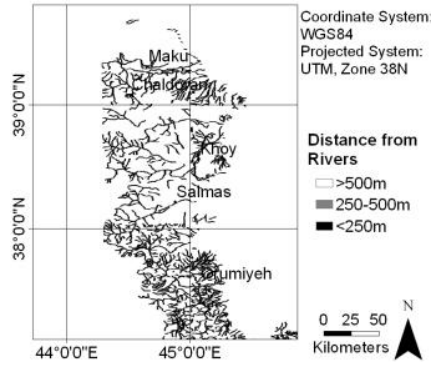
(b) Slope map (degree)



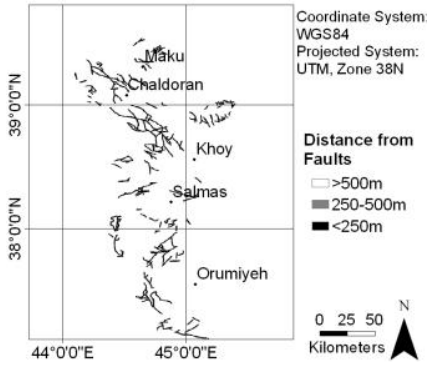
(c) Distance from urban areas (meter)



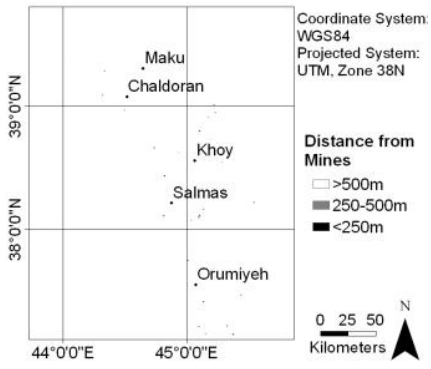
(d) Distance from roads (meter)



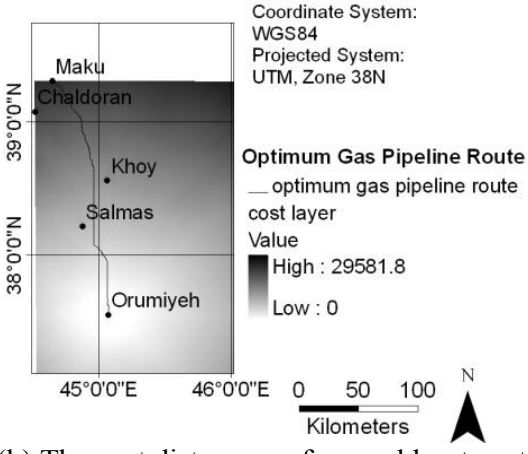
(e) Distance from rivers (meter)



(f) Distance from faults (meter)



(g) Distance from mines (meter)



(h) The cost distance surface and least cost path for pipeline route

Figure 1. (a) Study area, (b) Slope, (c) Distance from urban areas, (d) roads, (e) rivers, (f) faults, (g) mines and (h) The cost distance surface and least cost path for pipeline route

Three experts evaluated the weight of each criteria for gas pipeline routing in terms of their contribution (Table 1). The viewpoint of each expert was considered as an evidence (E 1, E 2 and E 3). Since it is not expected that the experts have enough knowledge regarding whole aspects of the problem, therefore the knowledge-based route selection is

an uncertain problem. Uncertainty in experts' viewpoints must be handled to select the least cost pipeline route. In this research, DST of evidence was used to handle the uncertainty and combined the evidences from different sources. Therefore, using Dempster's rule of combination, a degree of belief was calculated that takes into account all the available evidences. Table 1 also represents the combination of views provided by expert 1 and expert 2 (E 1,2) and the final weights E 1,2,3 (combination of E 1,2 and expert 3's view).

Table 1. The weights assigned to the criteria according to the expert's opinion

Criteria	E 1	E 2	E 3	E 1,2	E 1,2,3
Slope (degree)	0.25	0.3	0.25	0.1316	0.036
Distance from Urban Areas (m)	0.3	0.35	0.25	0.1842	0.051
Distance from Roads (m)	0.1	0.1	0.15	0.0175	0.003
Distance from Rivers (m)	0.1	0.1	0.1	0.0175	0.002
Distance from Faults (m)	0.15	0.1	0.15	0.0263	0.004
Distance from Mines (m)	0.1	0.05	0.1	0.0088	0.001

The cost distance surface between any location on the surface and the point of origin (Orumiyeh) was generated based on input layers and their final weights. Then the least cost path for the gas pipeline between Orumiyeh and Maku was calculated (Figure 1h).

To assess the reliability of the least cost pipeline route selection, a sensitivity analysis was undertaken by changing the experts' weights. The experts' weights for the six criteria were varied within a range of 10% to do such a simulation. The results (Figure 2) showed that the determined pipeline route was robust and the perturbation of the decision weights had a small impact on it and is almost independent of changes in the decision weights associated with the selected criteria.

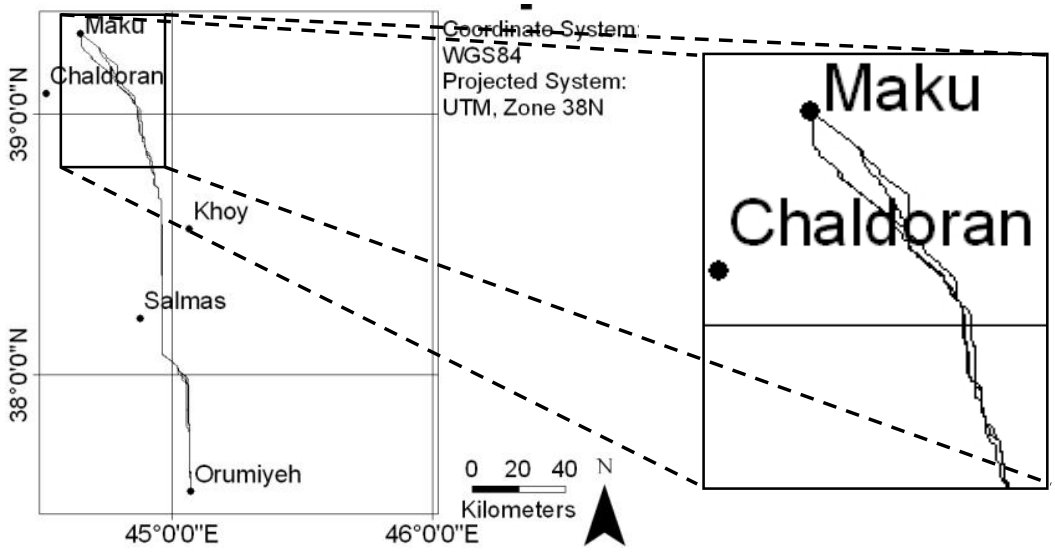


Figure 2. The results of sensitivity analysis by changing the experts' weights within a range of 10%

4 Conclusion

Pipeline route selection is a MCDM problem based on weighting the criteria by the experts. However, since it is not expected that decision makers have enough knowledge regarding the whole aspects of the problem, uncertainty exist in the experts' perspective. In the proposed approach, DST was used to make the decision under uncertainties. Using a Monte Carlo simulation, the reliability of results was assessed with respect to a change in the experts' weight within a range of 10%. The results verified the validity of the determined least cost pipeline route under uncertainty in experts' viewpoints.

References

Amiri, A. R., M. R. Delavar, S. M. Zahrai and M.R. Malek, 2007, Earthquake Risk Assessment in Tehran Using Dominance-Based Rough Set Approach, ISPRS conference, National Cartography Center, Tehran, Iran.

Delavar, M.R. and Naghibi, F. (2003), "Pipeline routing using Geospatial Information System Analysis". Proceedings of the 9th Scandinavian Research Conference on Geographical Information Science, ScanGIS'2003, 4-6 June, Espoo, Finland, pp. 203-213.

Dempster, A. P. (1967), "Upper and lower probabilities induced by a multi valued mapping". *Annals of Mathematical Statistics*, Vol. 38, pp. 325-339.

Eastman, J.R.(2001), Guide to GIS and Image Processing, Volume 2, Clark lab,pp. 34-38.

Gordon, J., and Shortliffe, E.H. (1985), "A Method for Managing Evidential Reasoning in a Hierarchical Hypothesis Space". *Artificial Intelligence*, 26: 323-357.

Jahankhah, M. (2010), Geospatial Information Fusion Using Theory of Evidence for Tehran Seismic Vulnerability Assessmnent, MSc Thesis (in Persian with English abstract), College of Eng., University of Tehran. Tehran, Iran.

Lee, N.S., Grize, Y.L. and Dehnad, K. (1987), "Quantitative Models for Reasoning Under Uncertainty in Knowledge-Based Expert Systems", *International Journal of Intelligent Systems*, 2: 15-38.

Matos, D. F., Menezes, P. C. P., Cruz, C. B., Garcia, K. C. and Damazio, J. M. (2008), "CEA-GIS based automatic tool for selection of gas pipeline corridors-SIGGAS project". IAIA08 Calgary, Assessing and Managing Cumulative Environmental Effects, Nov. 6 – 9, 2008, 13p.

Nataraj, S. (2005), "Analytic Hierarchy Process as a Decision Support System in the Petroleum Pipeline Industry". *Issues in Information Systems*, **4** (2) (2005), pp. 16–21.

Sentz, K. and Ferson, S. (2002), "Combination of evidence in Dempster Shafer theory". TR 0835, Sandia National Laboratories, Albuquerque, New Mexico.

Shafer, G. (1976), A Mathematical Theory of Evidence, Princeton University Press, Princeton and London.

Tangestani M. H. *and* Moore F. (2002), "The use of Dempster–Shafer model and GIS in integration of geoscientific data for porphyry copper potential mapping, north of Shahr-e-Babak, Iran". International Journal of Applied Earth Observation and Geoinformation 4*:* 65–74

Yildirim, V, and Yomralioglu, T. (2007), "GIS based pipeline route selection by ArcGIS in Turkey". Twenty-Seventh Annual ESRI International User Conference, San Diego Convention Center San Diego, California. June 18-22, 2007, 6p.

References from websites:

Restriction regulations of natural gas pipelines in the vicinity of buildings and facilities, roads, transmission power lines, railroads and oil pipelines, 2006.(Available: http://igs.nigc.ir)

www.ngdir.ir

Propagation of spatial imprecision in imprecise quantitative data in agronomy

Karima Zayrit[1], Eric Desjardin[1], Cyril de Runz[1] & Herman Akdag[2]

[1] CReSTIC, Université de Reims Champagne-Ardenne
Karima.Zayrit@univ-reims.fr, Eric.Desjardin@univ-reims.fr, Cyril.de-Runz@univ-reims.fr
[2] LIP6, Université Paris 6
Herman.Akdag@lip6.fr

Abstract

One of the stakes of Observox, an observatory of agricultural practices, is to deal with imperfect spatial information and to always associate a quality evaluation to acquired or computed data. So, we introduce the notion of fuzzy geographical entities. Then, we consider both spatial and quantitative information in order to obtain fuzzy local quantitative information. This paper proposes a new operator which gives the fuzzy quantity of spatially disseminated chemical products for each location.

Keywords: Imprecision, fuzziness, propagation, agriculture.

1 Introduction

In the past 30 years, the use of GIS has grown and today it is the standard for managing spatial – as located on Earth – and spatiotemporal data. Their use goes from archaeology (Conolly and Lake, 2006; De Runz and Desjardin, 2010) to agronomy (the context of this work).

The spatial feature of studied entities is often as imprecise (and/or uncertain) as its quantitative and descriptive features. According to literature (Klir and Yuan, 1995; Smets, 1995; Fisher *et al.*, 2006; De Runz *et al.*, 2008), fuzzy set theory and fuzzy logic are a good approach to deal with this kind of data imperfection. Then, one can build entities where both the spatial and the quantitative features are fuzzy.

The fuzzy set theory allows overlap between fuzzy shapes. The question is: what is the value of fuzzy quantitative attributes in a location where two or more fuzzy spatial shapes overlap? The answer to this question is the heart of this article.

Actually, in order to build an observatory on agricultural practice in the Vesle Basin, we have to deal with multiple sources of information that introduce imprecision in the object. From this situation, spatial and quantitative information may thus be imprecise.

Indeed, in the spatial context, there is two main ways for modelling imprecision (Bejaoui *et al.*, 2009). In the first hand, the crisp models extend or transform precise spatial concept in order to represent spatial imprecision as for instance the Egg-Yolk model (Cohn and Gotts, 1996). In the second hand, the models are based on uncertain mathematical theories as the ones, such as (Navratil, 2007), using fuzzy sets (Zadeh, 1965), either those, for example (Worboys, 1998), exploiting rough sets, or those, as for instance

(Pfoser *et al.*, 2005), using probabilities.. The fuzzy models give us a unique and soft framework that allows us to represent imprecision and to better conceptualize the reality (see Smets (1995)). As the aim of our system is to give interpretable information in each location of the monitored space, fuzzy data modeling data is used by us.

This paper exposes our opinion and choices in order to answer to these questions in the context of agronomical data exploitation. It introduces an operator for the propagation of spatial imprecision to imprecise quantitative information. It also presents a global structure for the management of fuzzy geo-entities. This structure is based on a fuzzy data storage impact analysis.

Section 2 is devoted to the imprecise geo-entity modelling in the framework of fuzzy set theory. Then, the propagation of imprecision in overlapping areas is studied (section 3). Finally, the conclusion is presented in section 4.

2 Fuzzy modeling of agronomical entities

In the sustainable development context, the AQUAL project (a State-Region Project in the Champagne-Ardenne, France) highlights the need of a monitoring environment for the study of agricultural practices and their pressure on the water resources in the Vesle basin. It is called Observox and it exploits data coming from heterogeneous sources: satellite images, land registry, statistical data, Corine Land Cover and other European data. The construction of a unique set of entities implies the combination of information coming from all the sources. The built entities thus induce some imprecision in the definition of spatial features and quantitative attributes (Shi, 2010).

On the other hand, Fisher in (Fisher, 1996) presents a comparative study between crisp sets and fuzzy sets in order to model landscape. The formers simplify the modeling but could amplify errors. The latters make the models and the treatments more complex. In (Fisher *et al.*, 2006), the authors present a taxonomy of uncertainty in spatial context where the vagueness is associated to the fuzzy set theory. According to (Duckham *et al.*, 2001), vagueness is a special type of imprecision. Vagueness and imprecision could be both represented by fuzzy sets (Bouchon-Meunier, 1995; Klir and Yuan, 1995; Smets, 1995) introduced in (Zadeh, 1965).

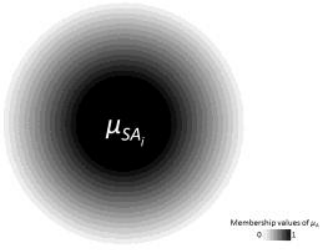
According to this, in agronomical studies as well as in geography, the geographical entities could be modeled as fuzzy geographical entities. Those entities have a label, a fuzzy spatial shape and a set of fuzzy quantities (each quantity corresponds to a specific attribute such as population or a specific chemical). The definition of a geographical entity may be defined as follows.

Let Ω be the set of studied geographical entities $\{A_1,...,A_n\}$. Let be Q the set of monitoring quantitative information $(Q_1,...,Q_m)$ if one supervises m different information $(P_1,...,P_m)$ as for instance m different molecules or products . Let us define a fuzzy geographical entity A_i in Ω as an object described by:

- A label or concept LA_i member of an ontology.
- A fuzzy set FSA_i describing its spatial representation. The membership function μSA_i of FSA_i is defined on $\mathbb{R}^2$.
- A fuzzy quantity FQ_jA_i for each quantity Q_j (of P_j) in Q. The membership function μQ_jA_i of FQ_jA_i is defined on $\mathbb{R}^+$.

An example of an A_i is shown in Table 1.

Table 1. A fuzzy geographical entity A_i (only one quantitative information is shown)

Type of information	Attribute	Value example
Label/Function	LA_i	Vineyard
Spatial definition	FSA_i	
Quantity of P_j (*e.g.* Isoproturon)	FQ_jA_i	

If Q_j is a precise quantity (with a value a), it could be represented by a singleton in the fuzzy set theory as follows: if $q=a$ then $\mu Q_jA_i(q)=1$ else $\mu Q_jA_i(q)=0$ such as q belongs to $\mathbb{R}^+$). This principle is presented in figure 1.

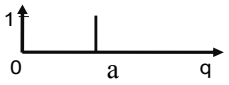


Figure 1. Illustration of a precise quantity a represented in the fuzzy set theory.

In the context of OBSERVOX, Q is the set of studied chemical (or at a micro-scale, the set of phytosanitary molecules). It could be for example a fuzzy prescribed dose or an estimation of quantity which was actually spread.

The next section is devoted to the sensibility of quantity values in a space location.

3 Propagation of imprecision

Let us consider x a location. We consider that the confidence in FQ_jA_i should be put into perspective with the membership degree $\mu SA_i(x)$ in order to define FQ_jA_i,x with its membership function $\mu Q_jA_i,x$ as proposed in (1). In this definition, when a fuzzy geographical entity A_i does not participate to the definition of x ($\mu SA_i(x) = 0$), the quantity of product P_j diffused at x by Ai is certain and null.

$$\begin{aligned} &if\ \mu SA_i(x) \neq 0\ then\ \mu Q_jA_i,x\ (q) = T(\mu SA_i(x), \mu Q_jA_i(q)) \\ &else\ if\ q = 0\ then\ \mu Q_jA_i,x(q) = 1 \\ &else\ \mu Q_jA_i,x\ (q) = 0 \end{aligned} \quad (1)$$

with q in $\mathbb{R}^+$ and T an aggregation function, usually a *t-norm* such as the multiplication or the minimum.

The imprecision, conceptualized using a classical fuzzy number for quantities and by fuzzy area for spatial feature, is the propagated in the consideration of fuzzy quantities at a specific location. As our goal is to consider all the quantities of a specific product at each location of the space, an aggregation operator is now needed for obtaining the combined information. Then, we use the Zadeh‘s extension principle that allows to extend

usual operation in the fuzzy set context such as in our context the sum (due to the additive aspect of product diffusion).

Thus if we deal with an additive information P_j, using this hypothesis and Zadeh's extension principle we define FQ_j,x the overall quantity at the position x by following the equation (2) for the definition of its membership function $\mu Qj,x$.

$$\mu Q_j,x(q) = \sup_{q=z+t} \quad (min_{A_i,A_k\ in\ \Omega^2, i\neq k} \quad (\mu Q_j A_i\ ,x\ (z), \mu Q_j A_k\ ,x(t))) \qquad (2)$$

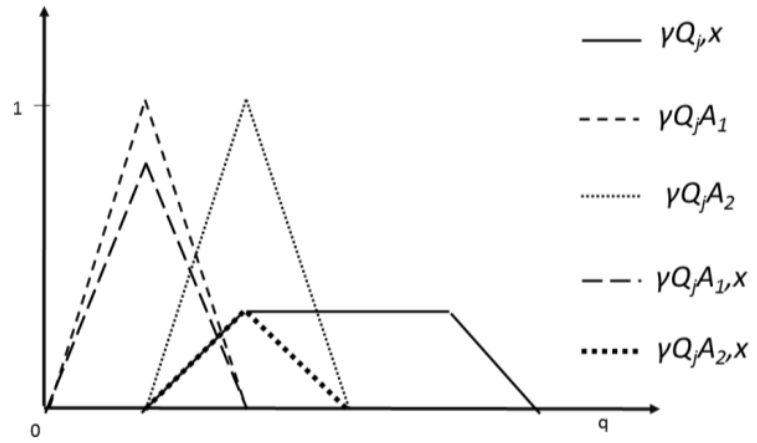


Figure 2. Illustration of the imprecision propagation on quantitative value Q_j of P_j for a specific location x, with $\Omega=\{A_1,A_2\}$, $\mu SA_1(x)=0.8$ and $\mu SA_2(x)=0.4$.

In order to test the feasibility of our approach, we illustrate it using two overlapped fuzzy geographic entities (A_1 and A_2) at a specific location x (figures 2 and 3). The goal is in this example to determine the total quantity of a chemical Pj (corresponding to Bentazone) at x.

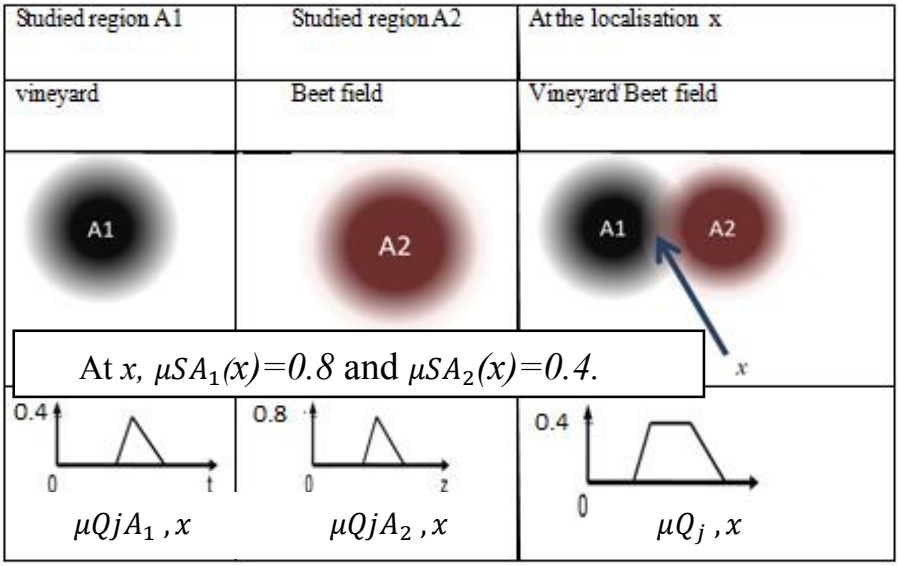


Figure 3. Illustration of the imprecision propagation: a spatial/quantity view.

This principle allows us to compute the quantity of each monitored molecule in every location of the studied region. The confidence in the computed fuzzy quantity is lower (or equal) than the original confidence in each fuzzy geographical entities.

4 Conclusion

In this paper, we propose a study of the imprecision propagation from spatial information to quantitative one. We firstly introduced our context and our approach of a fuzzy geographical entity. Next, we proposed a new operator of imprecision propagation.

This paper is a starter for the future construction of an agricultural practice observatory. In our future work, we will use conceptual approach that allows us to automatically obtain a fuzzy spatiotemporal data storage solution (Zoghlami *et al.*, 2011), but we also want to study the propagation of quantitative imprecise information into other topological relations between fuzzy spatial objects.

This paper is a preliminary study before building the observatory. It presents our choice at the beginning of the project. In our future work, we will develop our approach by defining new fuzzy agronomical indices in the observatory.

Acknowledgements

We would thank the Seine-Normandy Water Agency, Champagne-Ardenne Region Council, France and European Union, through the FEDER, for their funding of the project CPER AQUAL.

Bibliography

Bouchon-Meunier, B. (1995), *Logique floue et applications*, Addison Wesley, Paris

Bejaoui, L., Bédard, Y., Pinet, F., Schneider, M. (2009), "Qualified topological relations between spatial objects with possibly vague shape". *International Journal of GIS*, Vol. 23, No. 7, pp. 877-921.

Conolly, J., Lake, M. (2006), *Geographical Information Systems in Archaeology*, Cambridge University Press

De Runz, C., Desjardin, E. (2010), "Imperfect spatiotemporal information modeling and its analysis in a generalization process in a GIS: application to archaeological information". *In* Jeansoulin, R., Papini, O., Prade, H., Schockaert, S. (eds.), *Methods for Handling Imperfect Spatial Information*, Springer Verlag, Studies in Fuzziness and Soft Computing, vol. 256, pp. 341-356

De Runz C., Desjardin, E., Piantoni, F., Herbin, M. (2008), "Toward handling uncertainty of excavation data into a GIS". *In Proceedings of CAA 2008,* Budapest, Hungary

Duckham, M., Mason, K., Stell, J., Worboys, M.F. (2001) "A formal approach to imperfection in geographic information". *Computers, Environment and Urban Systems*, vol 25, pp. 89-103

Fisher, P., Comber, A., Wadsworth, R., (2006). "Approaches to uncertainty in spatial data". *In* Devillers, R., Jeansoulin, R. (eds.) *Fundamentals of Spatial Data Quality*, ISTE, London, pp. 43-59

Fisher, P.F. (1996), "Boolean and Fuzzy Regions". *In* Burrough, P., Frank A.U. (eds) *Geographic objects with indeterminate boundaries,* CRC Press, pp. 87–94.

Klir, G.J., Yuan, B. (1995), *Fuzzy sets and fuzzy logic: theory and applications*. Prentice-Hall, 592p.

Navratil, G. (2007) "Modeling data quality with possibility-distributions". *In International Symposium on Spatial Data Quality - ISSDQ'07*, Enschede, Paysbas,

Pfoser, D., Tryfona, N., Jensen C.S. (2005) "Indeterminacy and Spatiotemporal Data: Basic Definitions and Case Study". *GeoInformatica*, vol. 9, no. 3, pp. 211–236

Shi, W.Z., (2010), *Principle of Modeling Uncertainties in Spatial Data and Spatial Analyses.* CRC Press, 412p.

Smets, Ph. (1995) "Probability, Possibility, Belief : which for what ?". *In* De Cooman G., Ruan D., Kerre E.E. (eds.) *Foundations and Applications of Possibility Theory*, World Scientific, Singapore, pp. 20–40.

Worboys, M.F. (1998), "Imprecision in finite resolution spatial data". *GeoInformatica*, 2, pp. 257–279.

Zadeh, L.A. (1965), "Fuzzy Sets". *Information Control*. Vol 8, pp.338–353

Zoghlami, A., de Runz, C., Akdag, H., Zaghdoud, M., Ben Ghezala H. (2011), "Handling imperfect spatiotemporal information from the conceptual modeling to database structures". *In International Symposium on Spatial Data Quality (ISSDQ)*, Coimbra, Portugal, to appear.

Spatial Data Model for Local Government with the Inspire Rules

Jose Carlos Martinez Llario[1], Rafael Sierra Requena[2] & Eloina Coll Aliaga[3]

[1] PhD in Geodesy and Cartography. Universitat Politècnica de València.
jomarlla@cgf.upv.es
[2] Engineer in Geodesy and Cartography. Universitat Politècnica de València.
rasiere@upvnet.upv.es
[3] PhD in Geodesy and Cartography and degree in Computer Engineering. Universitat Politécnica de València.
ecoll@cgf.upv.es

Abstract

This article focuses on the study of a new spatial data model for the Spanish Local Government in accordance with the Directive 2007/2/EC (INSPIRE).
The new data model will contain local and national spatial data compliant with the INSPIRE rules. The INSPIRE generic conceptual model and its data specifications guidelines are based on ISO 19100 and OGC international standards.
Therefore the implementation of this spatial data model will enable municipalities to improve data interoperability, harmonization and to publish spatial data sets and web services. The establishment of this spatial data model improves integration of data on local SDI to allow any query of citizens and technicians.
The principal goal for the Spanish local Government is to have a GIS technology to manage the territory with the same data structure and network services as other municipalities in the EU.

Keywords: INSPIRE, SDI, GIS, ISO 19100, spatial data model, Local Government

1 Introduction

This paper presents a study of the spatial data of the Spanish Local Government and its application to the national and international normative on geographic information.

The Spanish government is formed by the following levels of territorial organi-zation: state, 17 autonomous communities or regions (NUTS2 [1]), 50 provinces (NUTS3) and 8116 municipalities. The objective of the Spanish government is to promote the use of GIS (Geographical Information System) and Spatial Data Infrastructures (Rajabifard, A., Williamson, I.P, 2001; Nebert and Douglas 2004) for municipalities through projects like LocalGIS [2] or Spanish normative [3].

The INSPIRE (Infrastructure for Spatial Information in Europe) Directive 2007/2/EC [4] and following regulations define the structure of spatial data and services for a future European SDI. The INSPIRE policy of harmonization and interoperability of spatial data is consistent with international standards of the OGC (Open Geospatial Consortium) and the ISO/TC 211 [6].

The INSPIRE Commission has established the structure of spatial data (spatial data model) to define the data transfer and the way to provide data and web services (view, search and query). The European Union will store the data from all countries in a centralized geoportal (Bernard *et al.*, 2004; Maguire and Longley, 2004) that will allow queries for any study: environmental, hydrological, demographic, etc.

However, the local spatial data can be numerous and very different for each department (planning, environmental, fiscal, economic, tourism, etc.) and multiscale. (Jiang *et al.* 2005). Furthermore, the data of municipalities has the larger scale (minimum 1:5000) and better precisions (at least 0.5 m) should be provided to INSPIRE.

Therefore the study focuses on local data structure to be harmonized and interoperable with other government with an important part as national and interna-tional normative of spatial data.

One of the most important goals for the results will be a spatial and relational data model to interact with different data models and different GIS applications (Desktop GIS, spatial database, SDI, management modules) allowing a better resource management in the local Government (Harvey and Tulloch, 2006).

2 ISO 19100 Standards and OGC

2.1 Introduction to international standards series ISO 19100

The international technical committee ISO/TC 211 began to work in 1995 with the aim of establishing a standard for digital geographic information. The most important job is to ensure the quality process and results through the standards.

ISO/TC 211 standards deal with the structuring of spatial data, methods and information processing tools and services to access or transfer information. Standards are developed by working groups of the ISO/TC 211 in collaboration with other institutions like OGC or GSDI (Global Spatial Data Infrastructure) [7] to minimize overlap.

The standards description begins with the general rules for ISO/TC 211 standards on ISO 19101-19109. These general rules explain the reference model, unified conceptual modeling (UML), encoding, management and profiles.

The most interesting of ISO 19100 family standards (Ariza Lopez and Rodriguez Pascual, 2008 ISO 19100) for our study are those related to spatial or temporal data models (ISO 19107-19110) and spatial referencing (ISO 19111-19112). There are other interesting rules for local government to assess the quality of spatial data (ISO 19113, 19114, 19138).

The Spatial data types are defined in ISO 19107 (primitives, aggregates or topological) with the class “GM_Object”, but the spatial object types are not complete because at this stage of the study coverages (ISO 19123) are not considered. The other standards for service implementation, data encoding and management are not considered at this first phase of our study.

3 The INSPIRE data model

3.1 INSPIRE Directive

The directive 2007/2/CE was adopted to establish an Infrastructure for Spatial Information in the European Community (INSPIRE). This directive establishes the basis of

harmonization and interoperability of spatial data on services for the founding of European SDI.

New regulations have been taken in addition to the directive:

- Regulation No 1205/2008 as regards **metadata** [13]
- Regulation No 976/2009 as regards **Network Services** [14]
- Regulation No 268/2010 and 1088/2011 as regards **access to spatial data sets and services** [15], [16]
- Regulation No 1089/2010 and No 102/2011 as regards **interoperability of spatial data sets and services** [17], [18]

These regulations establish common technical specifications for network services (view, sharing and search). They also provide the detailed definitions of the spatial data sets and metadata for European SDI (INSPIRE). Every spatial data set and service specification developed by INSPIRE rely on international standards of the family ISO 19100, OGC, CEN/TR 15449 and the GSDI cookbook.

3.2 Generic Conceptual Model

The INSPIRE Generic Conceptual Model focuses on the harmonization of specification level that includes semantic aspects and rules needed to support in-teroperability, aspects such as metadata or services are beyond the scope of these rules. The INSPIRE reference model is defined to provide a structure that will consistently describe components and their relationship to INSPIRE data specifications.

a) ***Application schema:*** it provides the structure of spatial data, this specifies the types of spatial objects and their properties (attributes, relations, operations and constraints). In INSPIRE it is meant to conform to the General Feature Model as specified in ISO 19109, expressed in a formal conceptual schema language UML 2.1 in English.

b) Feature catalogue: it is a different representation of the information that is written with text legible and translated to all official languages of the EU. The catalogue allows the access and queries to individual elements of the application schema. [17]

c) Dictionary: it is used to manage names, definitions and descriptions of all spatial object types that are used in INSPIRE application schemas and feature catalogues. The dictionary is one of the instruments for the cross-theme harmonization of concepts in INSPIRE. [19]

Currently INSPIRE has a specification of data with different application sche-mas for each spatial data theme of the annex I: Administrative Units, Cadastral Parcels, Geographical Names, Hydrography, Protected Sites, Transport Networks, Addresses, Coordinate Reference Systems, Geographical Grid Systems.

4 Spatial Data Model for the Spanish local government

The new spatial data model for the local government should be built in accord-ance with the standards ISO 19100 and INSPIRE. The most important task is to adapt the spatial data model to INSPIRE that already follows the ISO 19100 standards as they are the rules demanded by the EU to form local SDI.

There are much more Spanish geographic information than in the INSPIRE data model for the design of new spatial data model. The first phase of this research was dedicated to study the existing geographic information in Spain (Coll *et al*. 2005). It is interesting to analyze the amount of information available within each of the existing themes and compare the data with those required by INSPIRE.

The spatial data model designed for use or link with the new data model are:

a) BTA: This is the harmonized topographic base for the Spanish territory scales 1:5.000 – 1:10.000 mantained by the cartographic agency of each autonomous region and in the future will be the basic map of reference [24].
b) BTN25: This is the National Topographic Base, scale 1:25.000, on Spanish territory by the IGN (National Geographic Institute – Ministry of Public Works and Transport) and since for a long time it has been the basic map of reference. [25]
c) Cadastre: This data manages the ownership of property from urban and rural zones, scales 1:5000 (rural) and 1:2000 - 1:500 (urban). It is elaborated by the Spanish Directorate General for the Cadastre - Ministry of Finance. [26]
d) SIOSE: It is the land use database of the Autonomous Communities and national related with the CORINE Land Cover project. [29]
e) SIA: It contains all the spatial data of Spanish hydrology stored in the Water Information System (SIA) of the Ministry of Environment [31] and under the European policy of the Directive 2000/60/CE [32].
f) INE: It is the institution responsible for national statistics, population census, housing census, voter registration, addresses, etc.[27]
g) LocalGIS: It is a GIS software for processing, management and query of local government information through spatial data and SDI. [2]

All this Spanish geographic information must be analyzed (Müller and Hartmut, 2006) and if it is important for local government should be structured in new features related to the INSPIRE features. If the new features match INSPIRE features ("AdministrativeUnit, “CadastralParcel”) a different or new property data should be added to INSPIRE features.

The new data model shall be composed of elements grouped into common the-matic and relationships between their features and relationships with other thematic elements. The themes and INSPIRE features will remain, so only new elements, new properties, or new relationships will be added.

The tools used for our work are free software databases such as PostgreSQL with PostGIS spatial extension (Martinez Llario and Coll,) or Jaspa [38]. In addi-tion, we believe there are enough free applications to carry out such SDI projects. An example might be “qGIS” as Desktop GIS, “MOSKitt GEO” [39] as spatial database designer, “Mapserver” for web map server, “GeoNetwork” as metadata catalogue or “Deegree” for web services, etc).

The new spatial model will be compliant with the ISO to be fully compatible with INSPIRE, but having all the information necessary for local government in a format accessible to any application Desktop GIS and SDI through database. This undoubtedly enhances the competitiveness of the applications and its use in e-government (Goodchild, 2011, Enermark, 2009).

Finally, the relational data model should allow to manage the municipal information and likewise be structured like the rest of government (state, communities, Europe) to achieve the information flow between them and citizens with e-government.

5 Conclusion

The local SDI development in our country is growing with new technologies. However, the increase of applications does not guarantee the quality of municipal mapping products. Therefore, our proposal aims to make a common spatial data structure to all municipalities that try to ensure minimum standards of data quality, harmonization and

interoperability. This will help the competitiveness of enterprises and the introduction of new computer developments as free software.

The local SDI will be favored because they have a more powerful and coordi-nate engine approach to their citizen-oriented applications. With the new data model will emerge the need for quality municipal geographic information for any technical project that is developed in their municipality. Furthermore, the data model will allow local government to integrate different spatial data bases and its coordination with other administrations as Cadastre or INE.

The Integration with the INSPIRE will help the National Geographic Institute (IGN) to structure the data in order to be integrated directly into its database and transferred to the European Union.

We hope when the new spatial data model is completed all the municipalities that use it will be able to use the applications and tools for SDI and management of e-government.

Acknowledgments

This work has been supported by the research project "Creation and cartographic feeding of Spatial Data Infrastructures" in the local government by means of a data model that integrates cadastre, planning and cultural heritage CSO2008-04808 from the Spanish Government (CICYT) and the European Union (ERDFunds).

References

Ariza Lopez, F.J, Rodriguez Pascual, A.F, (2008), *Introducción a la normalización en la información geográfica: la familia ISO 19100*. Universidad de Jaén. ISBN: 978-84-612-2075-5.

Bernard, L., Kanellopoulos, I., Annoni, A., Smits, P., (2004), *The European geoportal—one step towards the establishment of a European Spatial Data Infrastructure*. Computers, Environment and Urban Systems, Volume 29, Issue 1.

Coll Aliaga, E.; Martinez Llario, J.C., Irigoyen, J., Velasco, E. (2005). Geographic Information and Local Government Management (GISMUN).

Goodchild, M.F. (2011), *Information technology as mega-engineering: the impact of GIS.* In S.D. Brunn, editor, Engineering Earth, pp. 37–47. New York: Springer. [499].

Harvey, F., Tulloch, D., (2006) *Local-government data sharing: Evaluating the foundations of spatial data infrastructures.* International Journal of Geographical Information Science, International Journal of Geographical Information Science, Volume 20, Number 7 pp. 743-768(26).

Jiang, J., Chen, J., Han, G. (2005), *A Model for Integrating Multi-scale Spatial Data for e-Government and Public Service*. FIG Working Week 2005 and GSDI-8. Maguire D.J., Longley P.A.,(2004), *The emergence of geoportals and their role in spatial data infrastructures.* Computers, Environment and Urban Systems 29 (2005) 3–14.

Martínez-Llario, J.C, Coll, E., *GIS analysis using different spatial databases.*

Müller, Hartmut. (2006). ‚*Spatial Data Infrastructure in Germany – Principles and Initiatives.University of Applied Sciences, i3mainz – Institute for Spatial Information and Surveying Technology.*

Nebert, Douglas D. (2004). *The SDI Cookbook. GSDI Global Spatial Data Infrastructure*.

Rajabifard, A., Williamson, I.P, (2001): *Spatial data Infrastructures concept, sdi hierarchy and future directions*. Proceedings of GEOMATICS, 2001.

Web References

[1]European Parliament (2003). *Regulation (EC) No 1059/2003 of the European Parliament and of the council of 26 May 2003 on the establishment of a common classification of territorial units for statistics (NUTS).*

[2]*LocalGIS Web.* http://www.planavanza.es/avanzalocal/Soluciones/Paginas/LocalGis.aspx

[3]Spanish Estate (2010), *Ley sobre las infraestruturas y los servicios de información geográfica*.

[4]European Parliament (2007). *Directive 2007/2/EC of the European Parliament and of the council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE).*

[6] Internacional Standard Organization. *The ISO TC/211 Standards*.

[8]Eurogeographics : *Guidelines for Implementing the ISO 19100 Geographic Information Quality Standards in National Mapping and Cadastral Agencies*.

[13]European Parliament. *Comission Regulation (EU) No 1205/2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards Metadata.*

[14]European Parliament. *Comission Regulation (EU) No 976/2009 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards the Network Sevices*.

[15]European Parliament. *Comission Regulation (EU) No 268/2010 of the European Parliament and of the Council as regards the access to spatial data sets and services.*

[16]European Parliament. *Comission Regulation (EU) No 1088/2010 of the European Parliament amending Regulation (EC) No 976/2009 as regards download services and trans-formation services.*

[17]European Parliament. *Comission Regulation (EU) No 1089/2010 of the European Parliament and of the Council as regards interoperability of spatial data sets and services.*

[18]European Parliament. *Comission Regulation (EU) No 102/2011 amending Regulation (EU) No 1089/2010 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards interoperability of spatial data sets and services*.

[19]INSPIRE. *Feature Concept Dictionary*.

[20]INSPIRE. *Metadata Implementing Rules*.

[21]INSPIRE. *D2.5: Generic Conceptual Model, Version 3.3*

[22]INSPIRE. *Web of Access to documents of data especifications.*

[24]*Especificaciones técnicas BTA versión 1.0*.

[25]Instituto Geográfico Nacional (IGN). *Web. Base Topográfica Nacional 1:25.000. BTN25.*

[26]*Spain Cadastre Web*. http://www.catastro.meh.es/

[27]*National Statistics Institute*. http://www.ine.es

[29]*Siose Web*. . http://www.siose.es

[31]*SIA Web.* http://www.mma.es

[32]Parliament (2000). *Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water polic.*

[38] *Web of Jaspa*

[39] *Web of Moskitt geo*.http://www.prodevelop.es/en/products/MOSKitt/MOSKittGeo

POSTERS

Analyzing the most adequate GPS kinematic observable for linear elements control methodologies of cartographic products

Antonio T. Mozas-Calvache, Manuel A. Ureña-Cámara & Juan J. Ruiz-Lendínez

University of Jaén. Campus Las Lagunillas, s/n. 23071 - Jaén (SPAIN)
antmozas@ujaen.es, maurena@ujaen.es, lendinez@ujaen.es

Abstract

The widespread use of digital cartography inside the Information Society has created the need for a more precise quality control. This need involves the development of new methodologies in order to quantify positional control quality. Nowadays these methodologies tend to use linear elements. All the positional control quality methodologies must use more accurate sources than those features to control. Among all the surveying techniques, kinematic surveying using GPS code observable is the favourite one. However, we can achieve an increase of precision using the phase observable which allows us to control higher scales. This work presents the results of the comparison using different three-dimensional linear positional control methodologies of a kinematic survey (phase and code post-processed). In this way, we extend the use of cartographic quality control methodologies to GPS quality control. As a result, the precision and efficiency of each observable allows us to present a proposal of the most appropriate in order to achieve the quality control of maps according to their scale.

Keywords: Quality, GPS observable, positional control, linear elements.

1 Introduction

Positional control methodologies in cartography are generally based on the comparison between a sample set of points on the ground and the product. These methodologies have recently been extended with a new process which uses linear elements as the main source of control (Mozas and Ariza, 2010). These new techniques are based on the surveying of lines with high precision and their comparison with the features represented in cartography. These studies use a GPS mounted on a vehicle which runs along the motorways and processes the survey using differential-code (K-DGPS) from nearby reference stations to the surveyed zone.

This study uses positional control quality methodologies with linear elements (PCQM-LE) to analyze the differences between code and phase observables in GPS kinematic surveying. This analysis allows us to determine the efficiency of each observation and the possible increase in precision of phase observable (K-phase). All the observables and surveying kinematic methodologies using GPS are analyzed for determining which is the most appropriate based on its efficiency and the required positional precision (mainly due to the scale of the controlled map).

The height component (Z) is studied independently in this work, as it has traditional been analyzed, but using the same methodologies than those applied to planimetric components (X and Y). Longitudinal profiles of GPS surveys are generated in order to determine the quality of the height component.

This study allows us to: (i) determine the possibility to apply control cartographic quality at higher scales, (ii) determine the most suitable observable based on the efficiency and accuracy in order to propose the most appropriate based on map scale.

The objectives described in the previous paragraph are achieved using two GPS kinematics surveys of about 50 km along the N-323a motorway near the city of Jaén (Spain). These data are processed using code and phase observables independently in order to compare them using PCQM-LE. Both observations correspond to the same zone but were surveyed on different days at different times in the day, one was obtained at night (OBS1) and the other in the morning (OBS2). These different times are chosen in order to test the influence of the hour of observation because of possible changes in surveying precision or traffic density. The use of the same survey processed with both observables eliminates the systematic errors between code and phase, such as the station error, inclination of the surveying stick, etc.

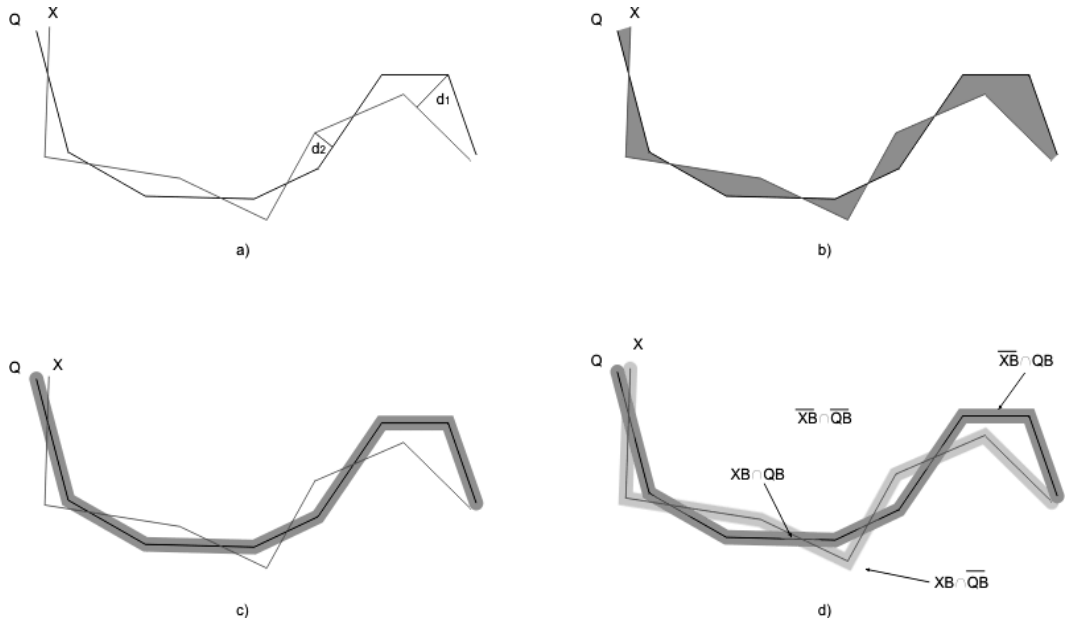


Figure 1. Control positional quality methods: a) HDM, b) EBM, c) SBM and d) DBM.

The main methodologies used to control positional quality using linear elements are described by Giordano and Veregin (1994) and Mozas and Ariza, 2010. The methods used in this study can be briefly described as:

(i) HDM, Hausdorff Distance Method (Abbas *et al.*, 1995): This method is based on determining the maximum distance between the lines X and Q (d1 and d2 in Figure 1a) and calculating the mean distance between the vertexes of one line to the other and vice versa. Using HDM we obtain the maximum and the average error between the vertexes of the lines.

(ii) EBM, Epsilon-Band Method (Skidmore and Turner, 1992): This method estimates the epsilon value by determining the total displacement area between two homologous lines (X and Q) (Figure 1b) and dividing this area by the total line length. EBM allows us to determine the average displacement originated by the enclosed area and the average number of crossings between lines.

(iii) SBM, Simple Buffer Method (Goodchild and Hunter, 1997): This method is based on buffering the most accurate line (Q in Figure 1.c), and then determining the percentage of the line to control (X), which is contained inside the buffer. A probability distribution of the quality of the X line is obtained by increasing the buffer distance of the Q line. SBM estimates the uncertainty originated by the inclusion of the controlled line into the control-line buffer.

(iv) DBM, Double Buffer Method (Buffer Overlay Statistics method by Tveite and Langaas, 1999): This method uses one buffer for each line (X and Q in Figure 1.d) and the analysis of the area covered by each possible buffer intersection. The buffer distance is increased in order to determine a probability distribution of the areas and a mean distance between buffers. Using DBM we obtain the average displacement of the buffers.

2 Methodology

The first step was surveying the different sections of a motorway using kinematic GPS observation with double-frequency mounted on a vehicle. The motorways were selected taking into account the proximity to the reference GPS station, road safety and the lack of obstacles which would interfere with the GPS signal or produce multipath effects (Ruiz *et al.*, 2009). Once the surveying was finished, the simultaneous observations from the reference station were obtained for post-processing. In our case, we used the reference station at the University of Jaén (UJaen). This station assures a maximum baseline distance of about 15 km.

In order to process the observations we have used two methods, the first one is differential code processing and the second one corresponds to processing based on phase observable, which correspond to an objective of this study. Therefore, we have processed the observations with both methods independently, using the processing software provided with the equipment. The surveys produce a sequence of 3D points (x,y,z) which defines lines followed by the mobile GPS. The observation time for each point was added for maintaining the acquisition order.

After that, two sets of homogeneous lines were defined in order to apply PCQM-LE. One is formed by lines to control (X lines in Figure 1) and the other by control lines (Q lines in Figure 1).

Therefore, an edition of the processed lines is required. In this step, the lines of code and phase survey were compared and edited from a general point of view. We apply this edition to obtain homogeneous characteristics between each homologous line from the two sets, that is: lines starting and ending at similar points have a similar length, respect logical coherence, etc. In our study, we have the particularity of simultaneous observation of the two sets. In this way we can use the observation time of the processed points in order to obtain the homogeneous and correlated set of lines. Using these two sets of lines we determined the efficiency of each method, which is the first result of our study.

As a result, we obtain two databases. Both of them have two sets of homologous lines one for the OBS1 surveying and other for the OBS2 with planimetric coordinates of the edited lines. On the other hand, two lines databases are obtained from the height of OBS1 and OBS2. These lines are determined using longitudinal profiles defined by the cumulative distances for each surveyed point of the kinematic chain (X-axis) and the height component (Y-axis). The cumulative distance is considered as the independent variable and the height is the dependent one. Finally, we applied the PCQM-LE using the CPLin software (Mozas *et al.*, 2007).

3 Results and discussion

The first result obtained from the methodology was the efficiency for each GPS observable, explained in the previous sections and applied to kinematic GPS surveys. Table 1 shows the results of efficiency in absolute and percentage.

Table 1. Post-processing Phase-Code statistics of OBS1 and OBS2.

Survey		Surveyed motorway	K-DGPS	K-Phase	Edited code vs. phase
OBS 1	**Length**	50 km	41,9 km	25,3 km	25,3 km
	%	100%	83,8%	50,6%	50,6%
OBS 2	**Length**	50 km	40,5 km	23,7 km	23.7 km
	%	100%	81 %	47,4%	47,4%

As shown in Table 1, there was an important loss of information after the phase processing with respect to the code observable. This loss of information was derived both from signal problems and the time-gap needed to recover the surveying or zones where ambiguities cannot be solved using phase processing. Finally, the set of lines obtained after the edition and comparison step between code and phase has the shortest length of the total surveyed length of phase processing.

We have applied the PCQM-LE to these two sets obtaining the results shown in Table 2 and Figures 2 and 3. The PCQM-LE have been used taking into account that the phase lines act as control lines (Q), or the ones with best positional precision, and the code lines are considered as lines to control (X).

Table 2. Results of HDM and EBM method.

Survey	HDM: Mean distances		HDM: Maximum distance		EBM: Average displacement	
	XY	**Z**	**XY**	**Z**	**XY**	**Z**
OBS1	0.081m	0.184m	0.828m	0.995m	0.073m	0.157m
OBS2	0.097m	0.224m	0.936m	1.176m	0.096m	0.209m

The HDM allows us to determine the mean distance between the lines and the maximum value. The results (Table 2) show that the mean distances are 0.081 m and 0.097 m for OBS1 and OBS2 respectively. Regarding Hausdorff distance the mean distance per length of the line shows 0.82 m and 0.93 m respectively.

On the other hand, the mean distance values are confirmed by the EBM, which shows a value of 0.073 m and 0.096 m for OBS1 and OBS2 respectively. These distances represent the mean value between the lines of code and phase processing.

With respect to the SBM, using a medium distance of 0.1 m to 1 m, we have obtained the probability function shown in Figure 2. This figure shows that the 95% of the line is inside the buffer if the buffer distance is about 0.3 m for both OBS1 and OBS2. The previous value is similar to the differences specified in the user's manual between both observation methods of the equipment. Finally, the mean displacement measure determined using DBM presents curves which ascend up to 0.09 m approximately maintaining this value in the remaining buffer distances (Figure 3). This mean displacement is similar to the values of the HDM and EBM.

Evaluating all the results, we can assure that the difference between the lines processed with code and phase have a mean displacement ranging from 0.08 to 0.1 m, with a

maximum difference that goes from 0.82 to 0.93 m, and assuring a maximum difference of 0.3 m with a 95% probability.

With respect to the difference between the OBS1 and OBS2, we can see in Figures 2 and 3 that the first survey (night time) has always lower differences than the second survey in all the PCQM-LE. Anyway this difference is too low to be significant with the sample used.

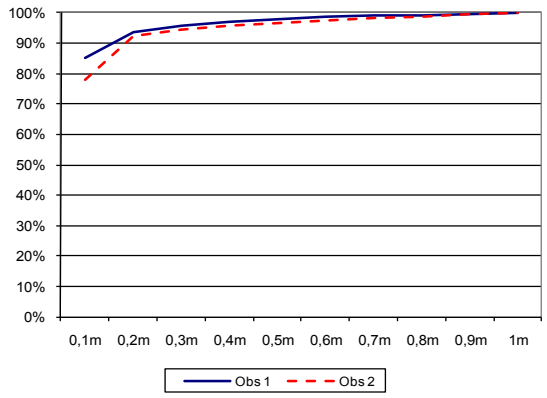


Figure 2. Results of SBM for planimetry: X: buffer's width; Y: % inclusion.

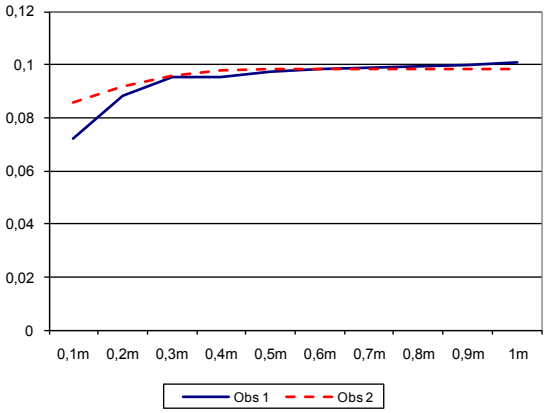


Figure 3. Results of DBM for planimetry: X: buffers' width; Y: Mean displacement (m)

The analysis of the height component using longitudinal profiles is also presented in Table 2. HDM and EBM were the only quality positional control methodologies applied to the profiles, because SBM and DBM are based on buffers and these buffers must be obtained from the same kind of variable. The results of HDM show values close to 0.2 m for the average distance and near 1 m for maximum distances. Regarding mean displacement (EBM) the values are ranged from 0.15 m to 0.2 m, too. These values duplicate the average error obtained for the planimetric analysis. However, they coincide with the user's manual specification of GPS.

The results of the analysis of OBS1 and OBS2 are following resumed: (i) Code observable is more efficient than phase observable; (ii) Phase observable is more precise than code with a mean displacement in planimetry of 0.1 m (0.3 m with a 95% of confidence); (iii) Displacements always have higher values in height (Z) than in planimetry (X,Y).

Table 3. GPS observable recommendation for positional control using lines.

Map Scale	**Coordinate accuracy (class 1 of ASPRS, 1990)**		**GPS Observable for post-processing**			
	XY (0.25mm x M)	**Z (1/3 contour interval)**	**No coordinate transformation**		**Coordinate transformation**	
			XY	**Z**	**XY**	**Z**
1:50.000	12.5m	16.6m	Code	Code	Code	Code
1:25.000	6.25m	8.4m	Code	Code	Code	Code
1:10.000	2.50m	3.2m	Code	Code	Code	Phase
1:5.000	1.25m	1.6m	Code	Phase	Phase	Phase
1:2.000	0.50m	0.6m	Phase	Phase	Phase	Phase

With the previous ideas in mind, Table 3 presents an advice about the most adequate observable based on the scale of the controlled map. As it is possible that cartography is defined in a different coordinate system that the one uses by GPS, Table 3 presents the recommendation for an error free transformation and with a transformation having a precision better than 0.2 m.

4 Conclusions

This study presents a comparison between the different GPS observable and survey kinematic methodologies both planimetrically and in height. We also propose the most adequate methodology based on the scale of the controlled map using linear elements (PCQM-LE). This proposal is obtained from two examples of real survey of each observable of a section of a motorway of more than 50 km.

The results of the relative precision among the observables have been determined applying positional quality control methodologies using linear elements to GPS surveys. They show that the mean differences between code and phase post-processed are near 0.1 m in planimetric positional precision and 0.3 m with a confidence level of 95%. However, in height, those mean differences are near 0.2 m.

The analysis of the efficiency of the surveys shows that the code observable has a faster recovery of the signal after losses than the phase (even using OTF algorithms). This indicates that the differential code survey has more observed length of the motorway by a factor of 33% with respect to the phase.

Finally, this study confirms the viability of the use of the code and phase observables to survey linear elements for position control quality. The selection between code or phase post-processed is determined by the controlling map scale, which defines the positional precision of the survey.

Acknowledgments

This work has been conducted within the research project ConPoCar by the National Ministry of Science and Technology under grant no. BIA2003-02234.

References

Abbas, I., Grussenmeyer, P., Hottier, P. (1995), "Contrôle de la planimétrie d'une base de données vectorielle: une nouvelle méthode basée sur la distance de Hausdorff: la méthode du contrôle linéaire". *Bul. S.F.P.T.*, Vol. 1 (137): 6-11.

American Society for Photogrammetry and Remote Sensing ASPRS (1990), ASPRS Accuracy Standards for Large-Scale Maps. *P.E.&R.S.*, Vol. 56 (7): 1068-1070.

Giordano, A.; Veregin, H. (1994), *Il contollo di qualitá nei sistema informativi territoriali*, El Cardo, Venecia, Italy, 138p.

Goodchild, M., Hunter, G. (1997), "A simple positional accuracy for linear feat*ures". Int. Journal Geographical Information Science*, Vol. 11 (3): 299-306.

Mozas, A. T., Ureña, M. A., Ariza, F. J. (2007), "CPLin: Una herramienta para el control posicional de la cartografía mediante elementos lineales". *Mapping*, Vol. 116: 81-87.

Mozas, A. T., Ariza, F. J. (2010), "Methodology for positional quality control in cartography using linear features". *The Cartographic Journal*. Vol. 47(4): 371-378.

Ruiz, J. J., Mozas, A. T., Ureña, M. A. (2009), "GPS survey of road networks for the positional quality control of maps". *Survey Review*, Vol. 41 (314): 374-383.

Skidmore, A., Turner B. (1992), Map Accuracy Assessment Using Line Intersect Sampling. *Photogrammetric Engineering and Remote Sensing*, Vol. 58 (10): 1453-1457.

Tveite, H., Langaas, S. (1999), "An accuracy assessment meted for geographical line data sets based on buffering". *I. J. Geographical Information Science*, Vol. 13 (1): 27-47.

(accessed on 23, June, 2011)

Handling imperfect spatiotemporal information from the conceptual modeling to database structures

Asma Zoghlami[1,2], Cyril de Runz[1], Herman Akdag[1,3], Montaceur Zaghdoud[2] & Henda Ben Ghezala[2]

[1] CReSTIC, Université de Reims Champagne-Ardenne
asma.zoghlami@etudiant.univ-reims.fr, cyril.de-runz@univ-reims.fr
[2] RIADI-GDL, ENSI Tunis
montaceur.zaghdoud@ensi.rnu.tn, henda.benghezala@ensi.rnu.tn
[3] LIP6, Université Paris 6
Herman.Akdag@lip6.fr

Abstract

Many real world systems and applications require a management tool that provides support for dealing with imperfect data. The aim of this paper is to handle the imperfection of spatiotemporal data from the conceptual modeling to the database conception. We propose to add new pictograms in PERCEPTORY in order to build imperfect spatiotemporal class diagrams such as those made using Fuzzy UML. Using those models, we organize the database as a three layer organization: data layer, metadata layer, multivalued layer. Those interlinked layers give a more accurate interaction.

Keywords: Conceptual model, imperfection, pictograms, database, multi-layer, multivalued set.

1 Introduction

As said in (Goodchild, 2006), "the quality of spatial data, as indeed of any data, is crucial to its effective use". As data imperfection is a part of (spatial) data quality, Geographic Information Systems (GIS) has to deal with, for instance, uncertain, imprecise and/or incomplete knowledge.

This paper considers imperfect information modeling in GIS, either on its descriptive, temporal or spatial levels. Imperfect information that characterizes the knowledge is most often manifested by the vagueness but also sometimes by uncertainty or lack of data. Much work about handling imprecise information in Information Systems and in spatial field has been done (Devillers and Jeansoulin, 2006; Jeansoulin *et al.*, 2010). Therefore, a lot of conceptual data models have been extended to model fuzzy data (Ma *et al.*, 2010); fuzzy set theory is the best possible tool that represents data imprecision.

The starting point of our conceptual data modeling is an UML type model which is the class diagram. It is just through this model that different levels of fuzziness were introduced into the concept of class, object and relations between classes. Thus, the concepts of fuzzy class, fuzzy association, fuzzy aggregation and fuzzy generalization were set in the work of (Ma, 2008).

PERCEPTORY, presented in (Bédard *et al.*, 2004), is a modeling tool that is mainly based on UML and extended with spatial and spatiotemporal stereotypes. These stereotypes have been developed through two distinct PVL (plug-in for visual language): the spatial PVL, for the representation of spatial data and the spatiotemporal PVL that is used to model spatial data and/or temporal data (Brodeur *et al.*, 2000). However, it does not provide a sufficiently explicit way to cope with information imperfection, such as uncertainty and imprecision.

Even though one considers that we can represent imperfect data using the dictionary associated to the model, the automation of the translation into GIS is not easy. This may be possible using an approach that splits data, imperfection and meta-data storage.

This paper aims to highlight the imperfect information in the PERCEPTORY model through the introduction of new visual symbols to manage imperfect spatial, temporal and descriptive data.

Then, we manage these imperfections from the conceptual data model to the database through a multi-layer approach. In this approach, a multi-valued layer is added to the more classical, *data layer* and *meta layer*.

This work is organized as follows. Section 2 is devoted to the insertion of visual variables in PERCEPTORY for the management of data and knowledge imperfection. Then, the structure of the built system is exposed (section 3) with a focus on the explanation of the imperfect management layer so called multivalued layer. The last section (section 5) presents the conclusion.

2 Highlighting imperfect knowledge in PERCEPTORY: new visual variables

There are a lot of ways to highlight the imperfection of knowledge in conceptual models. This section introduces the ones we have to deal in the objective of building an agronomical observatory, called Observox, in the Vesle Basin (the project challenges are described in (De Runz and Desjardin, 2009)).

2.1 Imperfect spatial symbol

Spatial pictograms in PERCEPTORY allow us to define the geometry chosen for spatial elements of a class model. The main geometries are: point, line and polygon. If one can't define with precision the boundaries of a spatial object, there is some imprecision on the geometric shape of this object at the class level. Thus in the PERCEPTORY model, we propose to use the three following basic geometries with dashed outline as shown below in figure 1.

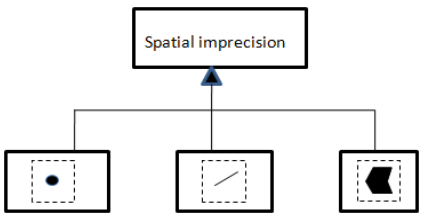


Figure 1. Spatial vagueness on the form of spatial objects

2.2 Imperfect temporal symbol

The temporal modeling in PERCEPTORY is based on the concepts of existence and evolution. The existence of an object corresponds to its period of life. Objects, having an instant existence, are represented by a pictogram indicating a date while objects that have a sustainable existence are represented by a pictogram indicating a time interval. At this level, there may be some imprecision in the definition of a date or of a time interval. The question is: when was an object present and when did it disappear? The two temporal pictograms are used with dashed outline to express this imprecision.

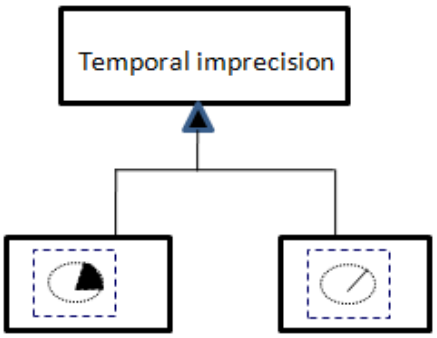


Figure 2. Temporal imprecision

2.3 Imperfect object attribute symbol

Some attributes in the class model may be defined by fuzzy sets or belief masses. To model this level of imprecision, the keyword **IMP** is introduced and placed in front of the attribute name.

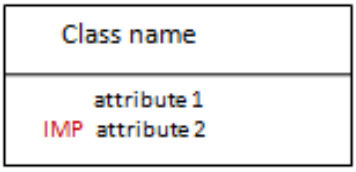


Figure 3. Modeling of imperfect attribute

2.4 On modeling of class relationship imperfection

The modeling of relationships between classes with their imperfection, we use the UML object diagram in which we associate to the link between the two classes a membership degree.

The relation between two classes (A and B) may also have different value depending of the class instances. In order to model that, we link each instance of A to each instance of B with a membership degree. The figure 4 shows an example of modeling an uncertain relationship between two classes A and B according to class instances.

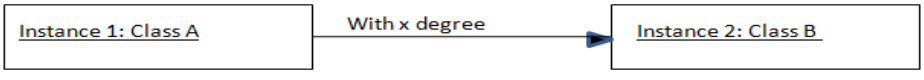


Figure 4. Modeling uncertain relationships between two classes through class instances

2.5 Membership degree of an object to a class

An object can belong to a class with a membership degree. This is shown in the object diagram through the introduction of the word "with a membership degree" after the instance name.

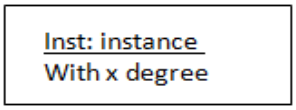


Figure 5. An object belonging to a class with a membership degree

3 Structuring the database: a multilayer approach

3.1 A multilayer approach

At the implementation level, a first classical layer is implemented. This layer, called *data layer*, contains data in a crisp modeling: the geometric data with the shape and the location of an object, the descriptive data referring to all the descriptive attributes of an object, the temporal data... In our database, the spatial data are represented according to the vector mode in which the objects are represented by points, lines and polygons instead of the raster mode because the vector approach has a lower storage cost.

The *data layer* is followed by a *meta layer* that represents the metadata. Metadata usually concerns the content, data sources, data identification, data quality, spatial representation, spatial reference and any other useful characteristic that may qualify the data. It can also store specific ontology, database schema, *etc.*

A third layer will allow us to link the *data layer* and the *meta layer* to a modeling tool that takes into consideration their imperfection. This layer allow to represent the imprecision and uncertainty through a multivalent approach *(*De Runz *et al., 2010)*.

The principle of the multivalent approach lies in the introduction of several truth values that modulate the information in order to focus on the natural language imperfection. Thus, linguistic expressions such as "very little", "a lot", etc can be used (Akdag *et al*., 2008). By building a link between data, metadata and imperfection modeling, we try to provide a more accurate view of the processed information by linking together these three layers through putting a link interface between them (figure 6).

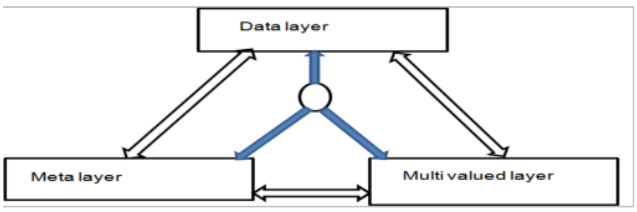


Figure 6. Relations between layers

3.2 Multivalued layer

The multi valued layer deals with all the spatial, temporal and descriptive imperfections that may be present on the two more classic layers (*cf.* figure 7).

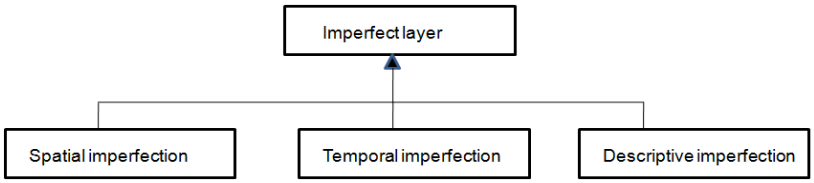


Figure 7. Managing imperfection in the multi valued layer

According to (Fisher, 1999; Dubois and Prade, 2009), the modeling of imperfect (spatial) data may be done using a lot of theories (probabilities, possibilities, fuzzy sets, belief functions, etc.). All of those theories use a paradigm of the attribution of weights (between [0;1]) to each element of the studied domain ($\mathbb{R}^2$ or $\mathbb{R}^3$ for space, $\mathbb{R}$ for time, $\mathbb{R}^+$ for quantitative information, *etc.*).

In order to reduce the cost and the complexity of storage and also in order to maintain the possibilities of exploitation, an approach based on the α-coupe principle – the domain of values for which the weight is higher or equal to α – is developed. Then, the modeling data have been putting into a multivalued form. This view is then adaptable to the more frequent uncertain representation. It allows users to choose the mode of uncertainty representation for every data. The interoperability between theories should after be done by the systems (it is one of our future goals).

The imperfection layer has an impact on the data layer through dealing with the imperfect relations between objects, imperfect object class relations, imperfect attributes, etc. To deal with possible uncertain relations between objects belonging to different database classes, one must associate a membership degree to the object identifiers in a new database table that indicates at what degree they may have a relation between them.

3.3 Links between layers (example)

A geographical entity (De Runz *et al.*, 2010) is composed by a fuzzy spatial area and a set of fuzzy quantities. To handle the vagueness at the database level, the imprecise information is stored in a specific table connected to the geographical entity table through an intermediate table which references the fuzzy quantity values stored in the fuzzy quantity table (see figure 8).

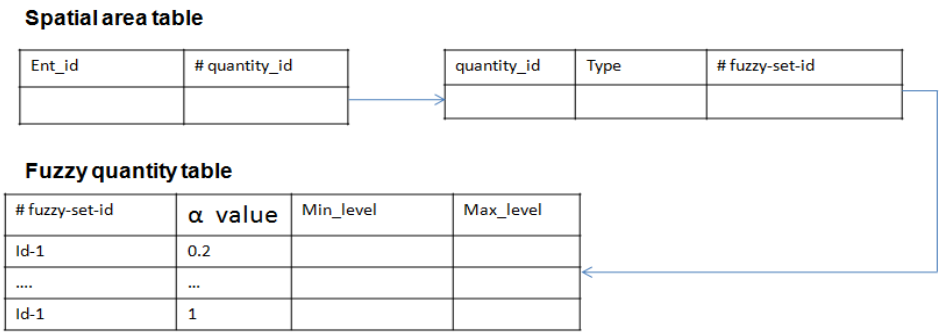


Figure 8. Illustration of the storage of fuzzy quantities as multivalent set of values.

4 Conclusion

In this paper, we started from the PERCEPTORY conceptual data model to handle spatial, temporal and descriptive data imperfections through the introduction of new

visual symbols. At the database level, a multilayer structure is implemented. Thus, a communication between a multi valued layer, a data layer and a meta layer is established.

In perspective, an application of our approach will be done in the building of an agronomical practice observatory in France. An application in archaeology is also planned.

Acknowledgments

We would thank the Champagne-Ardenne Region Council, France and European Union (FEDER) for their funding of the AQUAL Project.

References

Akdag H., Truck I. (2008), "Uncertainty Operators in a Many-valued Logic", *Encyclopedia Of DataWarehousing And Mining*, 2nd edition, Vol 4, Idea Group Publishing, pp. 1997-2003

Bédard Y, Larrivée S., Proulx M.J., Nadeau M. (2004), "Modeling Geospatial Databases with Plug-Ins for Visual Languages: A Pragmatic Approach and the Impacts of 16 Years of Research and Experimentations on Perceptory", *In* S. Wang *et al*. (eds.) *Proceedings of COMOGIS Workshops ER2004*, LNCS 3289, Springer, pp. 17–30.

Brodeur J., Bédard Y., Proulx MJ. (2000), "Modelling Geospatial Application Database using UML-based Repositories Aligned with International Standards in Geomatics", *In ACMGIS 2000*, Washington DC, USA

De Runz, C., Desjardin, E., Akdag, H. (2010), "Study of spatial fusion of geographical entities and quantitative information in accordance with their imprecision: application to agricultural information in Observox", *In* Tate, N.J., Fisher, P.F. (eds.) *Proceedings of Accuracy 2010,* Leicester, UK, pp. 33-36

De Runz, C., Desjardin, E. (2009), "Issues about qualified spatialization of agricultural practices and data for water contamination study in the Vesle basin", *In Proceedings of Spatial Analysis and GEOmatics*, Paris, France.

Devillers, R., Jeansoulin, R. (2006), *Fundamentals of Spatial Data Quality*, ISTE, London, UK, 312p

Dubois, D., Prade, H. (2009), "Formal representations of uncertainty", *In* Bouyssou, D., Dubois, D., Pirlot, M., Prade, H. (eds.) *Decision-making - Concepts and Methods*, Wiley, 3, pp. 85-156

Fisher, P.F. (1999), "Models of uncertainty in spatial data", *In Geographical Information Systems*, John Wiley & Sons, New-York, pp. 191-205.

Goodchild, M.F. (2006), "Foreword". *In* Devillers, R., Jeansoulin, R. (eds.), *Fundamentals of Spatial Data Quality*. ISTE, London, pp. 13–16.

Jeansoulin, R., Papini, O., Prade, H., Schockaert, S. (2010), *Methods for Handling Imperfect Spatial Information*, Studies in Fuzziness and Soft Computing, Springer, 400p.

Ma, Z. M., Yan, L. (2010), "A Literature Overview of Fuzzy Conceptual Data Modeling", *J. Inf. Sci. Eng*, vol. 26, n° 2, pp. 427-441

Ma, Z. M. (2008), "Fuzzy conceptual information modeling in UML data model", *In Proceedings of the International Symposium on Computer Science and Computational Technology*, vol. 2, IEEE, pp. 331-334

Quality Control of Fieldwork for Estonia's Topographic Mapping

Kiira Mõisja, Tõnu Oja

Department of Geography, University of Tartu, Vanemuise 46, Tartu, Estonia
kiiram@ut.ee, tonu.oja@ut.ee

Abstract

The quality of topographical datasets, as a reference framework for other spatial datasets, is very important. To improve data quality and better inform the users, outcomes of quality control are analysed. In this paper we deal with the results of empirical quality control of fieldworkers. It appears that the structure of errors in the general nonconformity database and the split of errors between fieldworkers differ in geometry and types of quality elements but are quite similar in most critical feature classes and misclassifications. The reasons for this difference need further investigation, solving the problems related to feature classes assumes revision of definitions. Also, expedience of mapping certain features needs revision.

Keywords: thematic accuracy, completeness, topographic mapping, field verification.

1 Introduction

Topographical datasets, collected by the National Mapping Agencies, provide a reference framework for other spatial datasets and for many spatial data services (Jakobsson, Giversen, 2007). The quality of these data sets is very important.

In the field of geomatics two groups of definitions of quality are used: internal quality (products that are exempt from error) and external quality (products that meet user needs) (Devillers, Jeansoulin, 2006). Antti Jakobsson (2006) discusses, how Lillrank's approach of geographic information quality can be categorized using the quality management viewpoints. Hence a production-centred approach is using quality control as a tool, based on ISO 19113 and ISO 19114, while a customer centred approach concentrates on uncertainty analysis, quality visualisation, user requirements and customer satisfaction.

Data quality is a concept related to uncertainty (Fisher *et al*, 2006), the nature of uncertainty is depending whether the feature class to be described is well or poorly defined and is caused by errors, vagueness and ambiguity.

Jakobsson and Marttinen (2003) introduce a model for data quality management including a quality inspection by the producer or the user. Data quality evaluation results are usually recorded and used for analysis for the continuous improvement of the Quality Management System (ISO 9001:2001) or for metadata.

This paper is the first stage in a wider research to find out the relationships between map quality, landscape diversity and fieldworker, which would allow to evaluate quality of map sheets also without direct field inspection of each sheet. The paper concerns the thematic accuracy and completeness of Estonian Basic Map. For the quality control field

inspection of selected sheets was provided. The results of quality check are analysed. Regularities in nonconformities are searched and based on them the reasons producing errors are analysed. To understand the reasons, the spread of nonconformities between geometry and quality elements is provided, most erroneous features and most typical misclassifications are determined.

2 Data and methods

The Estonian Basic Map is a national topographic database, the aim of which is to serve as basis for national thematic maps and registers containing spatial information (Mõisja, 2003). Since 1999 the producers of Basic Map have been chosen through public procurement procedures. For that purpose, the Guidelines for production of Basic Map were established together with quality requirements. The production scheme of the Estonian Basic Map consists of several stages, including stereoplotting and 100% of field verification. That large volume of fieldwork is needed due to the poor quality of aerial imagery and lack of earlier trustworthy mappings and state registers. Also, topographic maps from the Soviet period could not be trusted because of their age or large distortions (Mardiste, 2009).

The only option to determine the quality of thematic accuracy and completeness of fieldwork was field inspection. Starting from 2003, this was done using common methods and homogenous team of six employees of Estonian Land Board (Mõisja, 2003). The time-difference between the fieldwork and the quality control was usually two months. The areas for field inspection were selected trying to control as many fieldworkers as possible. The inspector pervaded and mapped a route in the region inspected. The inspected area (sample) was considered a buffer of 50 m (forests, bushes, and yards) or 100 m (all the rest) to both sides of the route. Altogether while inspection 1140 km was pervaded and the inspected area totaled in 126 km^2. The thematic accuracy subelement classification correctness, omission and commission defined in ISO19113 were inspected, and also, objects displaced or having wrong size (happened with buildings and small line objects) were marked as nonconformities. All nonconformities found were mapped and reported. For this study a geodatabase of nonconformities was created from the reports accepted by fieldworkers. This was the first time such inspection was carried out related to the Estonian Basic Map.

The database includes 4342 errors found in 42 inspected map sheets from 2003 – 2006 (15% of all sheets mapped during this period). The fieldwork on these sheets was provided by 19 fieldworkers, 10 of whom had carried out 67% of all fieldwork. Database analyzed includes all inspected fieldworkers and all errors found.

The following quality measures: error count (number of incorrect items), error sum (total length of incorrect line items, total area of incorrect area items, total number of incorrect point items), error rate (% of erroneous items with respect to the total number of items that should have been present) were used in analysis. The structure of errors in the frames of the total database of errors was analysed and also, the same by fieldworkers was made.

3 Error structure

3.1 Geometry

The errors were analyzed by geometry types as different measurement units are used for different geometry. Table 1 presents spread of errors by geometry types and by fieldworkers. The share of objects in the sample area is somewhat more uniform – 47% of lines, 30% of points and 23% of polygons.

The fieldworkers are split into three groups – those who make mistakes in line objects, those erroneous in points and those erroneous in both. The geometry of the errors of a particular fieldworker does not depend on the dominating geometry of the objects in the personal sample area (correlation r = 0.2-0.4)

Table 1. The structure of errors in the database and between fieldworkers.

Error structure	Database	Fieldworker
Geometry	45% lines, 41% points, 14% areas	**42%** of fieldworkers have more lines, **21%** have more point objects, **37%** have point and line errors equally
Quality element	49% omission 32% misclassification 12% commission	53% of fieldworkers have mostly objects missing, **21%** have dominating misclassification, **26%** have almost equally objects missing and misclassified
Most critical feature class (ranked)	Lines: 1. path, 2. ride, 3. ditch (<2m), 4. hedge Areas: 1. grassland, 2. open land, 3. young stand, 4. field, 5. forest Points: 1. heap of stones, 2. subsidiary building or production facility, 3. deciduous grove, 4. foundation	Lines: 1. path, 2. ride, **3**. hedge, **4**.ditch (<2, 2-4m) Areas: **1**. forest, **2** .grassland, **3**.young stand, 4. field, **5**. open land Points: 1. heap of stones; 2. subsidiary building or production facility, **3**. foundation, **4**. deciduous grove
Classification	Lines: 1.ditch 2-4m/ditch 4-6m, 2.track/path, 3.track/ride Areas: 1.open land/grassland, 2. grassland/field, 3. forest/young stand Points: 1. deciduous tree/grove, 2. dwelling house/subsidiary building, 3. ruins/ subsidiary building	Lines: **1**.track/path, **2.**ditch >2m/ditch 2-4m, 3.track/ride Areas: 1: **1.open land/forest, 2.open land/yard**, 3.forest/young stand Points: **1**. dwelling house/subsidiary building, 2. deciduous tree/grove, 3. ruins/ subsidiary building

3.2 Type of quality element

The biggest share of all errors comes from omission and misclassification (Figure 1, column Database; Table 1).

The share of quality elements for each fieldworker was determined from the total of his/her errors. Also here, the fieldworkers are split into three groups, the largest one is formed by those with omission dominating. Figure 1 ranks the fieldworkers according to difference in the share of omission and misclassification.

Due to the data model of the Basic Map the polygons may have only misclassification errors (polygons total in 100% coverage). Still, the domination of misclassification errors of fieldworkers No 1, 3, 5 and 7 is not due to numerous areal misclassifications.

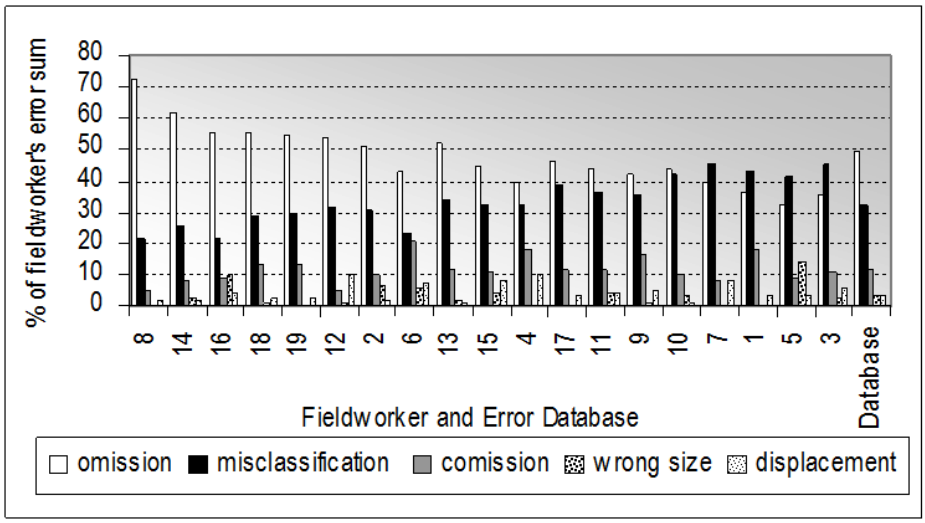


Figure 1. The structure of errors by type.

3.3 Feature class

Out of 104 feature classes that appeared in the sample, 20 were absolutely correct (19%). These are features mostly very clearly recognized in stereo (lake, railway, radio-tower, high voltage power-line), there are reliable data available from other databases (1. and 2. class highway) or, these feature classes appear in nature rarely (impenetrable marsh, ruins of windmill, light tower) and the fieldworker cannot be mistaken.

It is much harder to find out in which classes more errors appear as different quality parameters provide us different rankings. Like error count puts forests in the first place among areal objects whereas according to error rate they are only 13-th.

The most critical feature classes were determined using scatter plots. Features placed clearly above the imaginable diagonal line are more problematic and need thorough analysis (upper left quarter). Also, critical should be considered features close to the diagonal line in the right-hand quarters. The scatter plots were drawn separately for each geometry type (for example, Figure 2), most critical feature classes are listed in Table 1.

Figure 3 presents the share of the most erroneous areal features (% of total area of the same feature class) by fieldworkers. As seen in the figure, the variability is very high. Fieldworkers, not able to determine buildings (No 19) or marshes (No 1) are met.

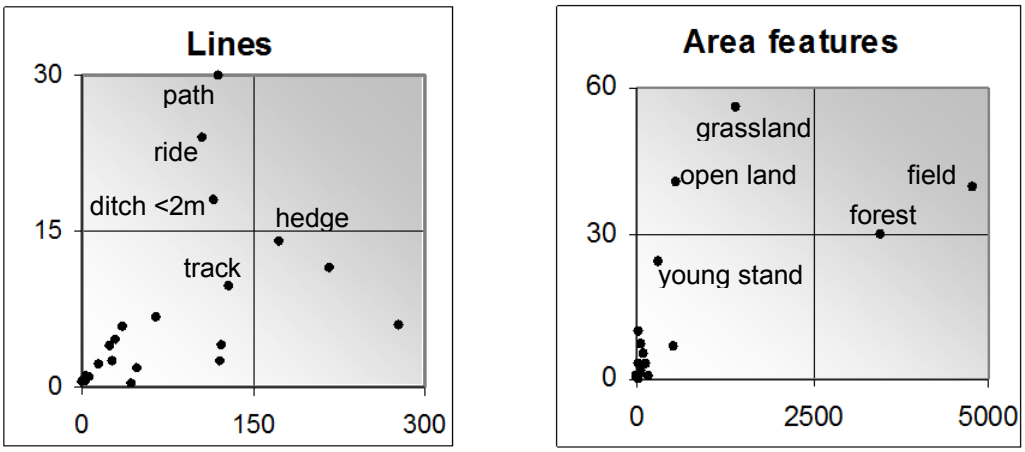


Figure 2. Scatter plots of line and areal features. Horizontal axis presents total length (m) or area (ha) of the feature class in sample area, vertical axis shows error sum of the feature class (m or ha).

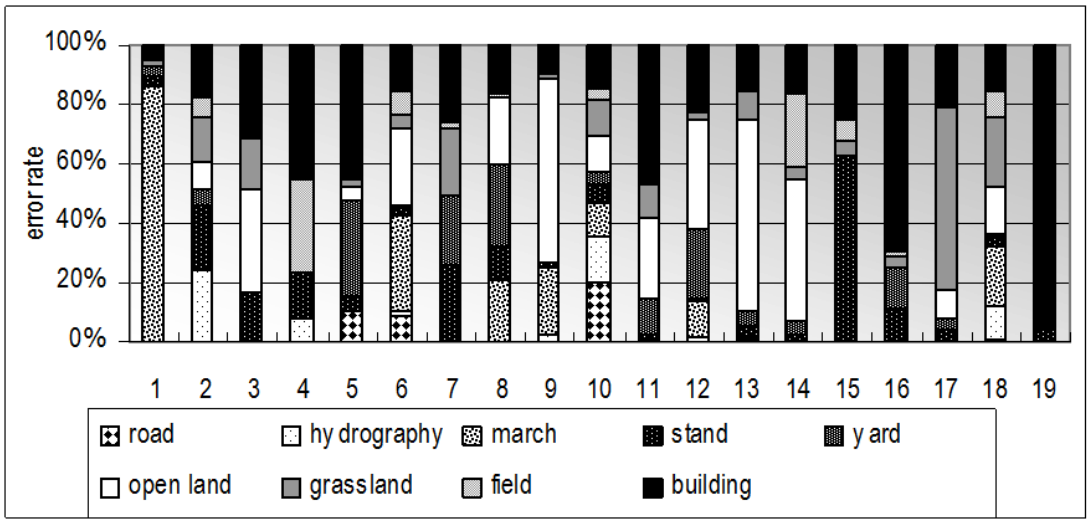


Figure 3. The most erroneous areal feature classes of the fieldworkers.

To determine the most critical feature classes among the fieldworkers, the features were ranked according to number of errors made by each fieldworker. The feature classes with no errors were given ranking number 20. Further, the ranking numbers were summarized over all 19 fieldworkers by feature classes that provides us a value which is the smaller the more often that particular feature class has been mapped with error. The results (column Fieldworker, Table 1) show that the same feature classes were most critical that were problematic within the whole database. There are minor differences in ranking order.

More than half of the errors in feature classes mentioned above are omissions. An exclusion appears for area features where due to the data model only misclassification is possible and track, subsidiary building and deciduous grove with also misclassification dominating.

3.4 Classification

Analyzing the misclassifications it appears that alike to habitat mapping (Cherrill, McClean, 1999; Stevens *et al.*, 2004) features often confused were "neighboring" each other (like grassland-field-open land, types of buildings and trees, types of small roads - ride-path-track, width class of ditch, ruins-foundation) or rapidly changing in time (forest-clear-cut, buildings under construction). The classes are distinguished more clearly while topographic mapping compared to habitat mapping.

By fieldworkers mostly the same classes appear while using ranking. Differences appear in areal features (Table 1).

4 Conclusion

General analysis of nonconformities and analysis by fieldworkers provide sometimes similar, sometimes different results. Notable is the difference in geometry structure – only 26% fieldworkers' errors match the same error pattern met in the general database. To explain the differences further analysis is needed to study the relations between the errors made by fieldworkers and the landscape.

The confusions with field-grassland-other open area are partly due to many fields being abandoned. The problem with the forests and clear-cut areas will probably remain but

decreases if the time-lag between taking aerial imagery and mapping is shortened. The sensibility of mapping some feature classes (heap of stones, foundation, and hedge) needs to be revised. Definitions should be revisited for all feature classes with more nonconformities, in particular misclassifications.

Acknowledgments

Support of the Estonian Land Board while field inspections is acknowledged. Authors are affiliated to the project SF0180049s9 target funded by the Ministry of Education and Science.

References

Cherrill, A., McClean, C., (1999), Between-observer variation in the application of a standard method of habitat mapping by environmental consultants in the UK. *Journal of Applied Ecology,* Vol. 36(6):989–1008.

Devillers, R., Jeansoulin, R., (2006), Spatial Data Quality: Consepts. *In:* Devillers, R., and Jeansoulin, R. (eds). *Fundamentals of Spatial Data Quality* ISTE, London, U.K.,pp.31-42

Fisher, P., Comber, A., Wadsworth, R., (2006), Approches to Uncerainty in Spatial Data. *In:* Devillers, R., and Jeansoulin, R. (eds). *Fundamentals of Spatial Data Quality* Publisher, London, U.K.,pp.43-59.

Jakobsson, A., (2006), *On the Future of Topographic Base Information Management in Finland and Europe*. PhD thesis, Helsinki University of Technology, Finland.

Jakobsson, A. and J. Marttinen, (2003). Data Quality Management of Reference Datasets – Present Practice in European National Mapping Agencies and a Proposal for a New Approach. In: *Proceedings of the 21st International Cartographic Conference (ICC 2003)*, Durban, South Africa, pp. 1748 – 1756.

Jakobsson, A., Giversen, J. (editors) (2007). Guidelines for Implementing the ISO 19100 Geographic Information Quality Standards in National Mapping and Cadastral Agencies , (http://www.eurogeographics.org/eng/documents/Guidelines_ISO19100_Quality.pdf).

Mardiste, H. (2009) Consequences of the Soviet map secrecy to national cartography in Estonia. *Geheimhaltung und Staatssicherheit. Zur Kartographie des Kaltes Krieges. Archiv zur DDR-Staatssicherheit* Bd. 9.1, S. 107, Bd. 9.2 (Abbildungen), Fig. 5.1 ;5.5.

Mõisja, K. (2003) Estonian Basic Map and Its Quality Management, , In *Transactions of the Estonian Agricultural University, 216. Baltic Surveying'03*.. Tartu: Estonian Agricultural University, pp. 135-142.

Stevens, J.P., Blackstock, T.H., Howe, E.A., Stevens, D.P. (2004), Repeatability of Phase 1 habitat survey. *Journal of Environmental Management*, Vol. 73(1):53-59.

Automated verification of road database in digital images

Aluir Porfírio Dal Poz[1] & Marco Aurélio Oliveira da Silva[2]

[1] São Paulo State University – Dept. of Cartography, Presidente Prudente-SP - Brazil
aluir@fct.unesp.br
[2] AMS Kepler Engenharia de Sistemas, São José dos Campos-SP, Brazil
mac_aurelio@yahoo.com.br

Abstract

This paper proposes a solution for the problem of automated verification of road network database in digital aerial images. The proposed method is based on two basic steps. In the first step, a road extraction method based on the dynamic programming algorithm is used to automatically extract roads from a digital aerial image. The road extraction method is initialized by using polylines obtained by projecting roads in database onto the image space. Projection errors along each projected road are estimated to establish the search space of the dynamic programming optimization algorithm. In the second step, a consistency analysis aims at checking the compatibility between the extracted roads and the corresponding projected roads is carried out. The checking criterion is based on an error model that basically embodies the uncertainties of the projected roads. The obtained results have shown that the proposed method is promising, as it can automatize the most part of arduous and time-consuming road network verification task. It was also shown that the method has the potential to be used in applications like the systematic error correction and geometric refinement of road database.

Keywords: Dynamic Programming, Road Extraction, Road Database, Consistence Analysis.

1 Introduction

The update of spatial information data is an important task for ensuring high data quality in Geographic Information Systems (GIS). Road database update using images consists of two subtasks: verification and change detection of road data. The main goal of the verification process is to identify in the image where parts of the road network do not exist anymore. The aim of the change detection consists in detecting and delineating new roads to be added to the database. This paper addresses the subtask of road verification.

Existing road database verification methods can be categorized into two classes. One class includes methods that analyze regions along roads projected onto the image space. In order to verifying roads in the image, some parameters are necessary, as e. g. contrast, collinearity, parallelism, and proximity. An example of method based on this principle is found in Baumgartner *et al*. (1996).

In the other class, a road extraction method extracts roads along road regions predicted in the image. This prediction uses polylines obtained by projecting the corresponding

roads in database onto the image space. The advantage of this verification principle is that it allows the refinement of the original road database, as the retraced roads in the image space can be transformed into the object space and integrated into the road database. Several verification approaches have been proposed, but in essence they compare roads extracted in the image with corresponding projected roads. This comparison is usually based on the displacement between extracted and projected roads. Roads or parts of roads in a database are considered verified if the displacement between corresponding projected and extracted roads is below a threshold. Basically, existing methods differentiate to one another in accordance to the employed road extraction method. For example, snakes are used in approaches proposed in Klang (1998), Fortier *et al*. (2001), and Agouris *et al*. (2000). In Dal Poz and Agouris (2000) a dynamic programming algorithm is used in the verification process. In Baltsavias (2004) and Zhang and Baltsavias (2002) the topographic database VEC25, consisting of road objects digitized from maps 1:25,000, is geometrically refined by extracting roads automatically from aerial imagery. Gerke *et al*. (2004) presented a graph-based method for verifying the road database.

This paper presents a method for verifying road database in aerial digital images. The extraction method is based on a road model that embodies geometric and radiometric road properties, which is optimized by the dynamic programming algorithm. The paper is organized as follow: Section 2 presents our method; Section 3 presents the experimental results; and Section 3 finalized the paper with main conclusions.

2 Method

Our road verification method is based on two basic steps: 1) Automated extraction of projected road by optimizing a road model with the dynamic programing algorithm; 2) Verification of road database by statically comparing corresponding extracted and projected roads.

2.1 Road extraction method

Photometric and geometric road properties are used to formulate a generic road model by considering that the road can be represented by an image-space polyline $P^i = \{p_1, ..., p_n\}$, where p_i is its ith vertex. The generic road model can be formulated by the merit function (Equation 1) and an inequality constraint (Equation 2), as follows (Gruen and Li, 1997),

$$E = \sum_{i=1}^{n-1} ((E_{p_1}(p_i) - \beta . E_{p_2}(p_i, p_{i+1}) + \gamma . E_{p_3}(p_i)).[1 + \cos(\alpha_i - \alpha_{i+1})]/ | \Delta S_i |) \quad (1)$$

$$C_i = | \alpha_i - \alpha_{i+1} | < T \quad (2)$$

where $E_{p_1}(p_i)$ is a function depending on the polyline point p_i and expresses the fact that road pixels are lighter than their neighbors on both road sides; $E_{p_2}(p_i, p_{i+1})$ is a function depending on two consecutive polyline points (p_{i-1} and p_i) and expresses the fact that road gray values along a road usually do not change very much within a short distance; $E_{p_3}(p_i)$ is a function depending on a polyline point p_i and expresses the fact that

a road is a lighter linear feature; α_i is the direction of the vector defined by points p_{i-1} and p_i; β and γ are positive constants; $|\Delta S_i|$ is the distance between points p_{i-1} and p_i; and T is a user-defined threshold for direction change between two adjacent vectors.

Equation 1 shows that only three consecutive points (p_{i-1}, p_i, p_{i+1}) of the polyline P^i are interrelated simultaneously, and that it can be decomposed into a sum of n-1 sub-functions $E_i(p_{i-1}, p_i, p_{i+1})$, i.e.:

$$E = \sum_{i=1}^{n-1} E_i(p_{i-1}, p_i, p_{i+1}) \tag{3}$$

Equation 3 shows that only three consecutives vertices of the road polyline are interrelated simultaneously. In this case, the dynamic programming algorithm is an efficient algorithm for optimizing the merit function given by Equation 1 (or 3) (Ballard and Brown, 1982). As the method requires an initial solution, an operator needs to supply a few seed points along the road. Optimization details of the road merit function can be found in Dal Poz and Agouris (2000).

In the context of road verification, the optimization process is automatically initialized as roads in database can be projected onto the image space. Projected roads substitute the user-supplied seed points, which are required whenever no initial predictions of roads are known, as e.g. in Gruen and Li (1997) and Dal Poz and Agouris (2000). As roads are extracted in the digital reference coordinate system, some steps are necessary to transform a road from the map projection coordinate system to the line (L) and Column (C) image coordinate system. Basic principles are given in the following. Detailed discussion on related geometric transformations can be found in e. g. Wolf and Dewitt (2000). Let $P(E_i, N_i, h_i)$ be a vertex of a road polyline in the database, where (E_i, N_i) are map projection coordinates and h_i is the ellipsoidal height. In order to transform the road vertex point $P(E_i, N_i, h_i)$ into the corresponding one ($p(L_i, C_i)$) in the image space, many parameters, such as the interior and exterior orientation parameters of the sensor, the ellipsoidal and map projection parameters, need to be known. Assuming that the mathematical transformation from object to image space can be represented by two equations, f_1 and f_2, the image coordinates L_i and C_i of a point can be expressed as a function of an object-space point $V_i = (E_i, N_i, h_i)$ and a vector (P_{ar}) of known parameters, as follows,

$$\begin{aligned} L_i &= f_1(P_{ar}, V_i) \\ C_i &= f_2(P_{ar}, V_i) \end{aligned} \tag{4}$$

2.1 Verification method

Our verification principle is based on a tolerance region constructed along both sides of the projected road (Figure 1).

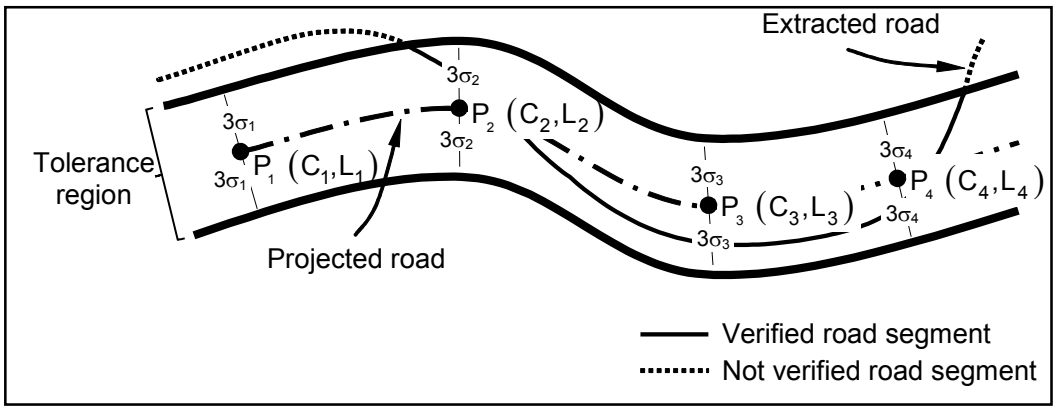


Figure 1. Verification principle.

Figure 1 shows that the tolerance region is established symmetrically around the projected road. The distance of any vertex of the projected road to the border point of the tolerance region is fixed as $3\,\sigma_i$, where $\sigma_i = \sqrt{(\sigma_{l_i})^2 + (\sigma_{c_i})^2}$. σ_{l_i} and σ_{c_i} are respectively the standard deviations of L_i and C_i image coordinates at the vertex point p_i, which are estimated by applying the error propagation principle (Wolf and Guilani, 1997) to the projection equation (Equation 4). The bounding value of $3\,\sigma_i$ takes into consideration the fact that the distance between corresponding points in the projected and extracted roads are 99,7% of times below $3\,\sigma_i$. A road or a segment of road in database is considered as verified if its tolerance region in the image contains the corresponding extracted road.

3 Experimental results

The proposed method was implemented using Borland C++ Builder 5 compiler for Windows XP. One aerial image at an approximate scale of 1:9,200 is used in the experiment. This image shows a region of Switzerland and is available in the LPS (Leica Photogrammetry Suite®) system, along with interior and exterior orientation parameters. A road map in UTM (Universe Transverse Mercator) was used in the experiment, having planimetric accuracy of 0.6 m and altimetric accuracy of 0.8 m. In order to enforce a low rate of road verification, we introduced 3.5 cm and -3.5 cm systematic errors in E and N map projection coordinates, respectively. The road model parameters α and β are empirically assigned to be 0.60 and 0.70, respectively.

Figure 2 presents the extraction results for 10 roads. The initial approximation for each road was obtained as described in Subsection 2.1. Table 1 shows the verification rate (VR) for each road.

Table 1 - Verification rate for each road

Road	1	2	3	4	5	6	7	8	9	10
VR (%)	0	8	6	2	5	78	100	47	100	70

As shown in Table 1, the VR indexes vary over [0%; 8%] for 50% of roads. Two short roads (20%) present verification of 100%. Remaining roads (30%) present verification varying over [47%; 78%]. The average VR is 37%. This low VR was expected, as the road map was contaminated with systematic errors.

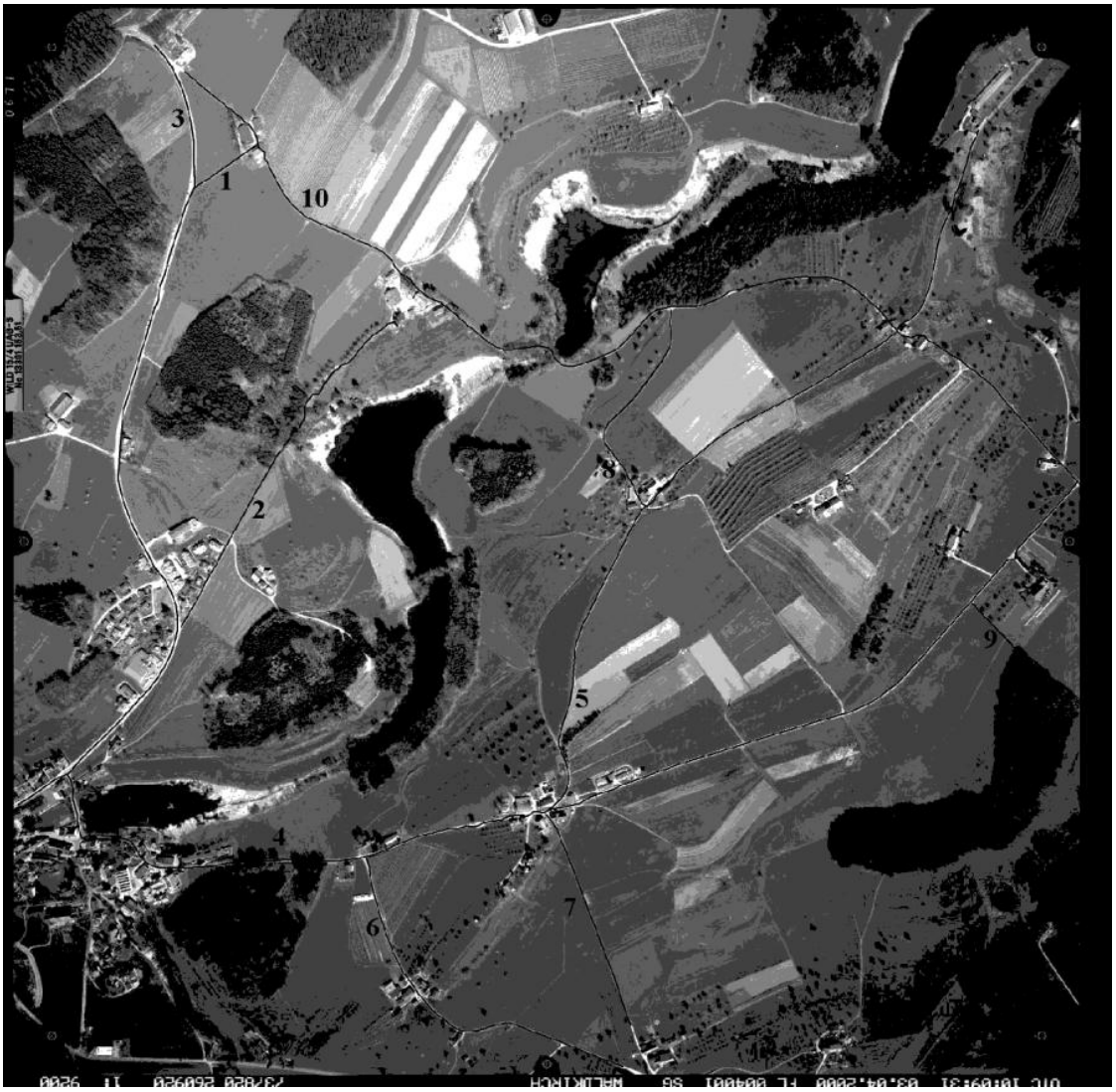


Figure 2. Extracted roads.

4 Conclusion

This paper presented a method for automatic road database verification in digital images. In general, the low verification rate in the presented experiment proved that the proposed method can be used to check geometric changes in the existing roads in a database or even to detect and eliminate systematic errors in the road database. In practical application in the context of large road database updating, the method can be potentially applied in the identification of road regions that require the inspection of a human operator, avoiding the tedious task of checking visually all over the database.

References

Agouris, P., Gyftakis, S., Stefanidis, A. (2000), "Uncertainty in Image-Based Change Detection". In: *Proceedings of Accuracy 2000*, Amsterdam, The Netherlands.

Ballard, D. H., Brown, C. M. (1982), *Computer Vision*. New Jersey: Pratice Hall.

Baltasavias, E. P. (2004), "Object extraction and revision by image analysis using existing geodata and knowledge: current status and steps towards operational systems". *ISPRS Journal of Photogrammetry and Remote Sensing*, 58: 129-151.

Baumgartner, A., Steger, C., Mayer, H., Eckstein, W., Ebner, H. (1996), "Update of road in GIS from aerial imagery: verification and multi-resolution extraction". In: *International Archives of Photogrammetry and Remote Sensing*, Vienna, Austria, pp. 53-58.

Dal Poz, A. P., Agouris, P. (2000), "Georeferenced road extraction and formulation of hypotheses for new road segments". In: Proceedings of SPIE 2000, Orlando-FL, EUA.

Fortier, M. F. A., Ziou, D., Armenakis, C., Wang, S. (2001), "Automated correction and updating of roads databases from high-resolution imagery". *Canadian Journal of Remote Sensing*, 27(1): 76-89.

Gerke, M., Butenuth, M., Heipke, C., Willrich. F. (2004), "Graph-supported verification of road databases". *ISPRS Journal of Photogrammetru and Remote Sensing*, 58: 152-165.

Gruen, A., Li, H. (1997). "Semi-automatic linear feature extraction by dynamic programming and LSB-Snakes". *Photogrammetric Engineering and Remote Sensing*, 63(8): 985-995.

Wolf, P. R., Guilani, C. D. (1997), *Adjustment Computations: Statistics and Least Squares in Surveying and GIS*, New York, 564 p.

Wolf, P. R., Dewitt, B. A. (2000), *Elements of Photogrammetry with Applications in GIS*, Boston: McGraw-Hill.

Zhang, C., Baltsavias, E.P. (2002), "Improving cartographic road databases by image analysis". In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Zürich, Switzerland, pp. 400– 405.

Construction of an elevation model for the evaluation of estuarine flooding frequency: the case of Lima River, Portugal

Ana Paula Falcão[1,3],Alexandre B. Gonçalves[1,3], Nuno Silva[1] , Nádia Braz[2], José Nuno Lima[2], Maria Amélia V. C. Araújo[1,4] & António Trigo Teixeira[1,4]

[1] Instituto Superior Técnico, Techn. University Lisbon, Av. Rovisco Pais, Lisboa, Portugal
afalcao@civil.ist.utl.pt, alexg@civil.ist.utl.pt, nunors@gmail.com, amelia.araujo@ist.utl.pt, trigo.teixeira@civil.ist.utl.pt
[2] Laboratório Nacional de Engenharia Civil, Av. do Brasil, 101, Lisboa, Portugal
nbraz@lnec.pt, jnplima@lnec.pt
[3] ICIST – Instituto de Engenharia de Estruturas, Território e Construção
[4] CEHIDRO – Centro de Estudos de Hidrossistemas

Abstract

Floods in urban areas are natural phenomena with severe impact on property and population. The probability of occurrence and the impacts may be reduced if preventive action is taken. The EU directive 2007/60/CE, concerning the evaluation and management of flood risks, stipulates that member states should evaluate the flooding threat, create extension maps and take adequate action to mitigate the risk. The use of hydrodynamic models in event simulation requires the previous construction of an elevation model that reflects geomorphological characteristics of the entire studied domain. It is then necessary to combine bathymetric data, relative to a hydrographic reference, with height data, relative to the mean sea level. The studied zone corresponds to a part of the Lima River estuary, for which datasets from topographic maps and bathymetric data were used. The main limitation to the construction of a single model was the inexistence of data in some parts of the range between the low and high tides. As such, it was necessary to acquire it with GNSS equipment and perform a transformation of ellipsoidal into orthometric heights through a local geoid undulation model. After the information adjustment, all data were transformed into the same height reference. The final elevation model was obtained by spatial prediction techniques and resampled in a regular 5-m resolution grid.

Keywords: Local geoid undulation model, estuarine flooding, combined elevation model, Lima River.

1 Introduction

Floods in estuarine zones are natural phenomena which occur mainly due to sea level uprising caused by meteorological factors (Araújo *et al.*, 2010; Weaver and Slinn, 2010). The production of flood vulnerability maps corresponds to the application of a European directive 2007/60/CE, and constitutes an essential tool to decision support in soil use

politics. The hydrodynamic modelling of extreme weather events needs a digital elevation model (DEM) valid for all the studied area, and modelling results are influenced by DEM characteristics (Ali *et al.*, 2009; Picado *et al.*, 2010). Within the study framework, when setting up such a model a combination of height and bathymetric data should be considered, and distinct reference systems must be taken into account (Casaca *et al.*, 2008). Data obsolescence, the existence of gaps and the diversity of sources, different acquisition dates and techniques frequently turn the direct application of hydrodynamic models impossible in the coastal engineering context. It is then necessary to combine all the available data into a single model.

2 Description of the studied area and data

The Lima River estuary was chosen and the analysed area was its final section (10 km). The river mouth is in Viana do Castelo, north Portugal, where a harbour and two bridges exist.

The following datasets were used in the study: a contour/spot height dataset acquired from 1:25 000 and 1:10 000 topographic maps (contour intervals of 10 m and 5 m, respectively) produced by the Portuguese Army Geographical Institute (IGeoE) and the Portuguese Geographical Institute (IGP) in 1997 and 1996; and a bathymetry dataset, surveyed by Hidrodata enterprise and divided into:

- A bathymetry dataset acquired in 2006 with multibeam sonar. This dataset is composed by 99 399 points, presented in regular grid format with 5 m spacing. This dataset was named "*Estuary*" in this paper;
- A 5120 point set, surveyed in 2006 near and around the Eiffel bridge, in regular grid format with 4 m spacing (here named "*Eiffel*");
- A 65 535 point set, surveyed in 2004 upstream the Eiffel bridge, with 25 m per 10 m spacing, called "*Lima*".

The vertical reference for height datasets in Portugal is the tide gauge mean sea level at Cascais, a reference value defined by an average of the local tide gauge records from 1882 until 1938. Bathymetric data are relative to the hidrographic zero, defined 2.0 m below the vertical datum for the Viana do Castelo harbour, as referred in the Portuguese Hydrographic Institute tide tables (IH, n/d).

In planimetry the coordinate system for all datasets is Hayford-Gauss Datum Lisboa (SHGDtLx). Spatial data acquired with Global Navigation Satellite System (GNSS) was matched through Bursa-Wolfe transformations, following the official parameters (IGP, n/d). The hidrographic zero was chosed as the reference level.

Figure 1 displays the height and bathymetric datasets described above.

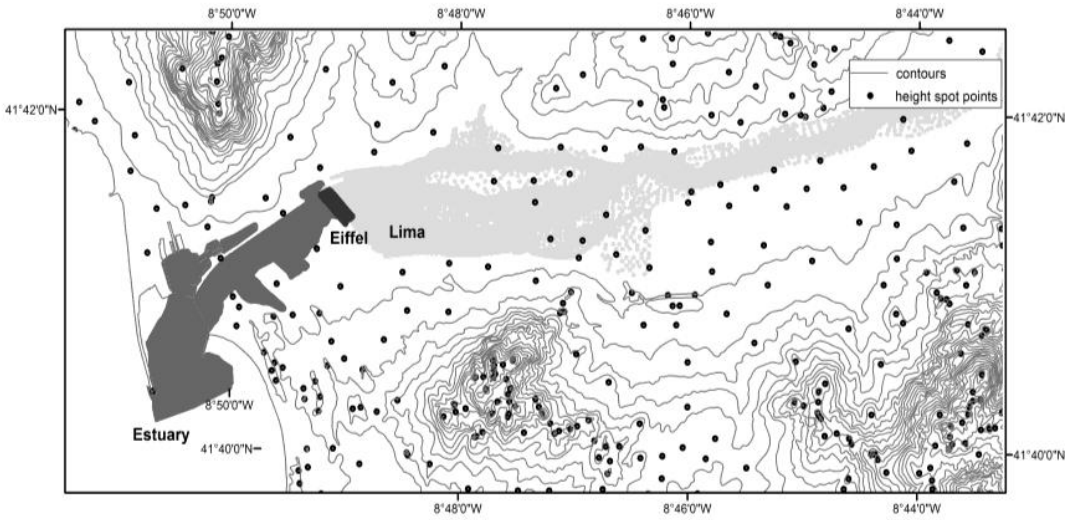


Figure 1. Height datasets (contours and height spots) and bathymetric datasets (Estuary, Eiffel and Lima).

3 Production of the model

3.1 Data collection

Data for the intertidal zone of the Lima River included several areas where height information was not available (Figure 2). A selection of these areas was surveyed with GNSS equipment: a total of 182 points was collected in Sept. 8-10, 2010. Dates were chosen to match the equinox high water with the spring tide. A Leica GS20 GPS and a AT501 monofrequency antenna were used to record 360 positions per point (5 s interval). Considering the objectives of the work and the unavailability of real-time positioning equipment, data was post-processed using Leica Geo Office software. The permanent station of Paredes de Coura (30 km base) enabled the fixed point for the post-processing correction.

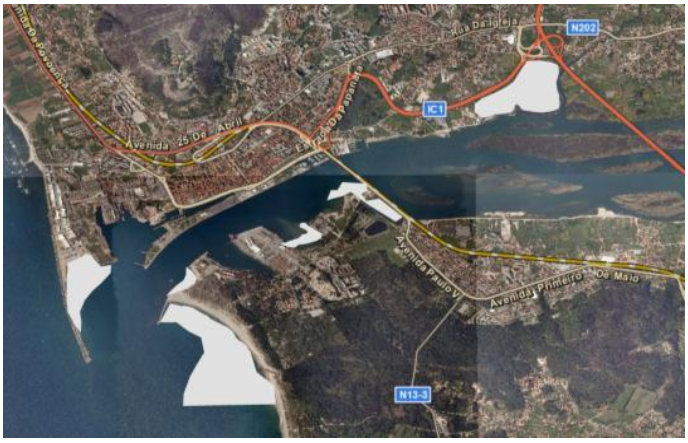


Figure 2. Areas in the Lima River estuary lacking height information (white polygons) over Google Maps imagery.

Figure 3. Acquisition of altimetry in the intertidal zone by GNSS acquisition.

Figure 3 shows the data acquisition operations in the intertidal zone. Table 1 lists the positional quality values obtained after post-processing for a sub-sample composed by geodetic points and benchmarks.

Table 1. Statistics for positional uncertainty of GNSS surveyed points.

Statistic	Planimetry (cm)	Altimetry (cm)
Range	14.5	20.5
Average	2.5	3.3
Standard deviation	1.9	2.8

As the collected heights are ellipsoidal and WGS84-based, it was necessary to transform it into orthometric heights using a local model of geoid undulation, according to (1):

$$h(P) \cong H(P) + N(P) \tag{1}$$

where H(P) denotes the orthometric height, h(P) the ellipsoidal heights and N(P) the geoid undulation relative to a generic point P.

3.2 Geoid undulation local model

To support the geoid undulation local model construction three geodetic marks and seven level benchmarks were identified. The orthometric heights of this set were given by IGP, and the ellipsoidal heights collected with GNSS field work. Records were registered with a Topcon GB-1000 receiver (5 s interval, 720 positions per point) and a Topcon choke ring CR3 antenna, and the used software was Topcon's Pinnacle. The geoid undulation local model was obtained by spatial prediction using the kriging technique according to Falcão Flôr (2010). The exponential model was chosed to express the spatial autocovariance, according with:

$$C(\rho) = \sigma^2 \exp\left(-\frac{3\rho}{a}\right) \tag{2}$$

where ρ denotes the distance between pairs of points in the sample, a the effective range, and σ^2 the variance at the origin. Parameters a and σ^2 were extracted based on the analysis of the experimental variogram presented in Figure 4.

Figure 5 (l) shows the predicted geoid undulation model for the studied area. Ellipsoidal heights were transformed into orthometric heights using (1) and later converted to hidrographic zero, as described above.

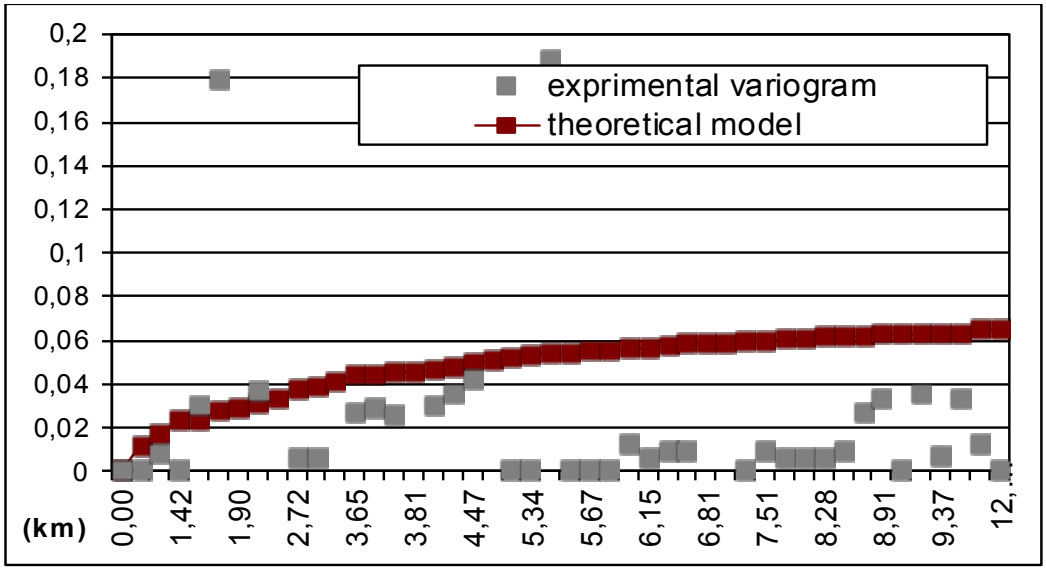


Figure 4. Theoretical model and experimental variogram.

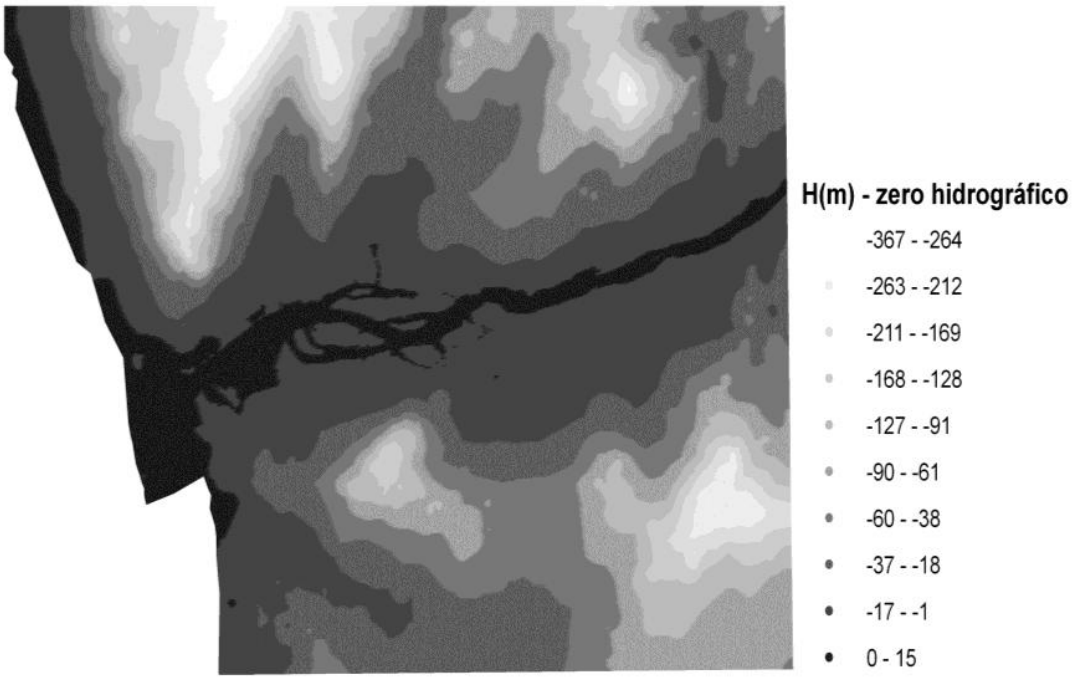


Figure 5. Combined elevation model (combining height and bathymetric datasets)

3.2 Combined elevation model

Due to the diversity of details (distinct sources and acquisition periods) of the datasets, these were firstly analysed to evaluate the compatibility between these datasets. Results indicate that orthometric heights relative to scales 1:25 000 and 1:10 000 have a non-systematic differences between -5.97 m and 15.24 m, which impossibilities a merge. Hence, only the 1:25 000 scale altimetry was used.

As the three bathymetry datasets overlap, the same procedure was conducted. For the *Estuary-Lima* pair (243 points), average and deviation values were 24 cm and 20 cm, respectively. For the *Eiffel-Lima* pair (5120 points), the average, standard deviation and range of the differences were 21 cm, 36 cm and 113.3 cm. This justifies the choice of not including the *Eiffel* dataset in the final elevation model. This combined model was pro-

duced from the 1:25 000 data, the remaining two bathymetric datasets, and the points surveyed with GNSS.

Finally, the spatial prediction was performed with kriging techniques. Data trends were previously modelled by a bilinear function, subtracted to the original values. An ordinary kriging was then applied to this difference. The combined elevation model was finally converted to a regular 5 m grid (Figure 5, r).

4 Conclusion

The compatibilization of heights and bathymetric data is frequently a requirement in coastal engineering studies. The methodology presented in this study allows the construction of an elevation model combining both types of information, with distinct levels of detail and potentially obtained by distinct acquisition techniques, in a precise and prompt manner.

In areas where generically there is lack of data, namely in the intertidal zone, data acquisition with GNSS techniques is a very efficient way but requires a local geoid undulation model to convert ellipsoidal heights into orthometric heights. In this construction it is necessary to select a set with known orthometric height values (geodetic marks and/or benchmarks), acquire the corresponding ellipsoidal height, and, based on these samples, estimate the value of the geoid undulation through a spatial predictor.

References

Ali, A., Zhang, H., Lemckert, C .J. (2009), "Numerical of hydrodynamics of a very shallow estuarine systems – Coombabah Lake, Gold Coast, Australia". *Journal of Coastal Research*, Special Issue 56: 922-926.

Araújo, M. A. V. C., Trigo Teixeira, A., Mazzolari, A. (2010), "Simulation of storm surge events at the Portuguese coast (Viana do Castelo). *Proceedings of International Conference Littoral 2010: Adapting to Global Change at the Coast*, London, UK, 2010.

Casaca, J., Falcão, A.P. (2008), "A revolução altimétrica do século XXI". *Revista Engenharia e Vida*, no. 45, Lisbon, Portugal.

Falcão, A. P., Matos, J., Casaca, J., Sousa, A. J., Gonçalves, A. (2008), "Preliminary Results of Spatial Modelling of GPS/Levelling Heights: A Local Quasi-Geoid/Geoid for the Lisbon Area". *Proceedings of International Symposium on Gravity, Geoid and Earth Observation 2008*. Chania, Crete, Greece.

Falcão Flôr, A. (2010), "A transformação de altitudes geométricas GNSS em altitudes gravíticas no contexto dos trabalhos topográficos". PhD thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisbon, Portugal.

IGP (n/d), webpage of Instituto Geográfico Português, URL: www.igeo.pt, access in Sept. 2010

Picado, A., Dias, J., Fortunato, A. (2010), "Tidal changes in estuarine systems induced by local geomorphologic modifications". *Continental Shelf Research*, Vol. 30: 1854-1864.

Weaver, R. J., Slinn, D. N. (2010), "Influence of bathymetric fluctuations on coastal storm surge". *Coastal Engineering*, Vol. 57: 62-70.

Partitioning of cadastre features based on straight skeletons for uncertain boundaries

Sunghwan Cho[1], Jonggun Gim[2], Gyoungju Lee[3]

[1] Department of Civil & Environmental Engineering, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul, Republic of Korea
hallem@snu.ac.kr
[2] Department of Information and Communication Engineering, KAIST, 291 Daehak-ro, YouSung Gu, Daejon, Republic of Korea
jonggun.gim@kaist.ac.kr
[3] Department of Urban Engineering, Chungju National University, 50 Daehak-ro, Chung-ju-si, Chungbuk-do, Republic of Korea
lgjracer@gmail.com

Abstract

In land-cadastre datasets, gaps or overlaps among parcels are frequently found due to non-abutting edges with adjacent parcels. These non-abutting edges are also called uncertain edges, and polygons containing at least one uncertain edge are called uncertain polygons. This paper discusses a new algorithm for efficiently searching parcels of uncertain polygons in given land cadastre dataset, and rationally partitioning them to allocate each to adjacent parcels. We used Constrained Delaunay Triangulation (CDT) and labeled the outputs for searching for uncertain polygons to produce polygons containing gaps or overlaps. Then, these polygons are partitioned using the straight skeleton based method; finally we allocated each partitioned gaps or overlaps to adjacent parcels to improve topological consistency of cadastre datasets. We have performed experimental application of this automated derivation of partitioned boundary from a real land-cadastral dataset.

Keywords: GIS, Uncertain boundaries, Land cadastre, Spatial Data Quality, Straight Skeleton.

1 Introduction

Land-cadastre presents boundaries of land ownership; each piece of land constitutes a parcel whose geometry is represented with a polygon. One of the most essential properties of these polygons is the topological consistency; polygons should share common edge(s) with adjacent parcels without gaps or overlaps (Laurini and Milleret-Raffort, 1994). Nonetheless, in real world cadastre datasets, topological inconsistencies, due to various causes, are rather frequently observed. These errors can be caused and propagated, for example, due to vectorization errors during a database conversion from a blue print cadastre data into a digital form (Chrisman, 1987; Ubeda and Egenhofer, 1997). Moreover, if the digitized cadastral data is not centrally collected and managed, topological inconsistencies are more commonly found, especially along the jurisdictional boundaries of surveying or administrative authorities. Usually the topological consistencies are

relatively well maintained in the central parts of administrative areas, whereas much more gaps and overlaps of various sizes are found along the administrative boundaries.

GAP-Tree based method (Tinghua and Peter, 2002) and other methods have been proposed to solve this problem. However, they are shown to be applicable only to limited cases or produce unsatisfactory results.

The ideal solution to this problem is developing central and seamless cadastral database by performing additional surveys on parcels with uncertain boundaries. This will guarantee error-free, consistent topology across datasets. But this, obviously, requires significant investment of time and budget. Practical alternatives to this approach are automated adjustment of parcel boundaries via various algorithms. However, these existing algorithms produce results within specified tolerances, usually based on specified tolerances from merging datasets (Beard and Chrisman, 1988). Our algorithm differs from them in two aspects: first, we do not need user-defined tolerances for searching for uncertain areas, second, we maintain history of vertices modification that occur during allocation of uncertain areas to adjacent polygons. We used the technique of tagging the triangles produced by *constrained Delaunay triangulation* (CDT) with the numbers of overlapping areas with other triangles. Then, we identified gaps and overlaps by checking the tag and merged them as necessary. Finally, we performed the conflation by partitioning the gaps/overlaps via straight skeleton method and allocating the segments to adjacent polygons.

2 Skeleton Operators

Figure 1 illustrates varying results of different vector operators stemmed from the skeleton operator: Triangulation based skeleton allows assigning varying weights to different polygon edges, consequently shifting the center line closer to the polygon that shares more semantic similarities. But it is difficult to apply this method to polygons with spikes (e.g., zigzag shape). Medial axis method is less problematic to be applied to such conditions. However, this does not allow unequal weight assignment which is advantageous in some cases. The straight skeleton method, with augmentation of the original algorithm, can produce better results for all the above mentioned situations. The straight skeleton extension method (Tanase and VeltKamp, 2004) adds additional edges to vertices to make the result geometrically similar to the results of Medial Axis Method.

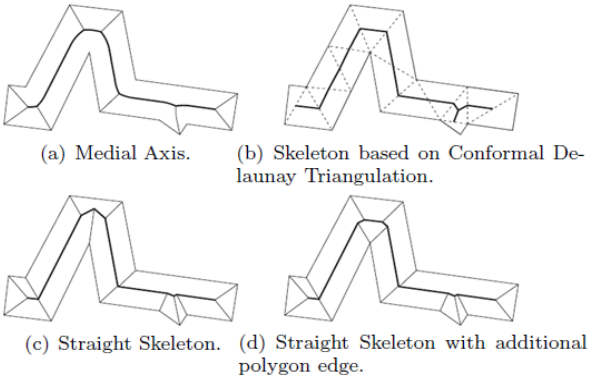


Figure 1. Comparison of skeleton operators (with emphasized centerline).

3 Algorithm for partitioning polygons with uncertain boundaries

The algorithm consists of three steps:

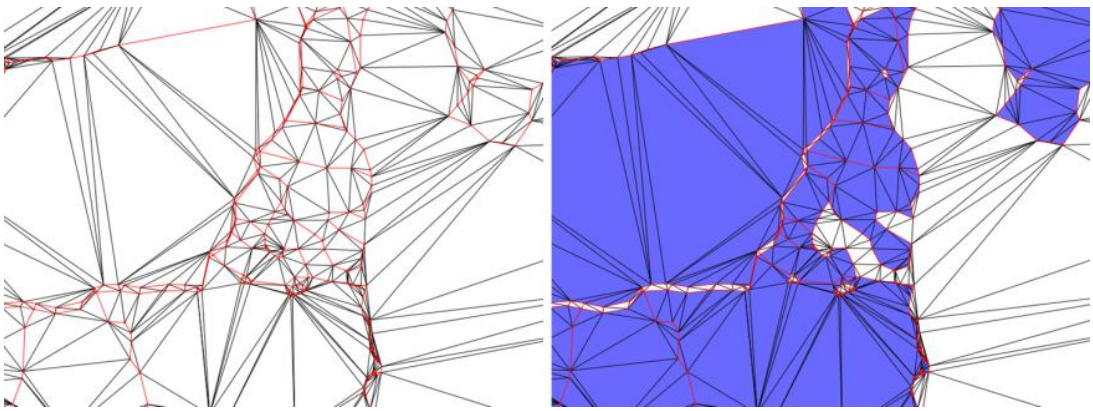
1. Search for uncertain areas via CDT based tagging.
2. Produce straight skeletons for uncertain areas.
3. Adjust boundaries by allocating segments to adjacent polygons.

3.1 Search for uncertain areas

We used *constrained Delaunay triangulation* (CDT) for searching areas of topological inconsistency among polygons. CDT fills entire area of polygons with triangles, which allows detection of overlaps or gaps within datasets. We tagged each triangle by the number of crossing edges of other triangles (for gap, the value is 0, for overlap the value is greater than 1).

CDT decomposes polygons into non-overlapping triangles, none of which cross the boundary of the polygon. This means that all the polygons, including ones with holes, can be triangulated without adding extra vertices (de Berg *et al.*, 1997). We created CDT inside and outside of each polygon boundary. If two polygons are adjacent by an edge e, e is an edge that is drawn twice. Since most of triangulation programming libraries discard one point at duplicated points, this does not require further manipulation. Likewise, when edges are found to intersect, they are split with a new vertex created at the intersection point.

Identifying uncertain area: If all polygons in a dataset form a planar partition, all triangles will be tagged as 1. Gaps or overlaps are easily recognized by checking if the tag value is not equal to 1. Figure 2 (a) shows a CDT produced by constraints. Figure 2 (b) shows the gaps and overlaps detected via tag counting.



(a) Produced CDT (b) Triangles with tagging value = 1

Figure 2. Detecting uncertain areas based on CDT.

3.2 Skeleton based partition operator

The straight skeleton algorithm was used for partitioning uncertain areas. Since individual vertices of parcel data are of great monetary significance, we ensured that the history of vertex change is traceable.

3.3 Assignment of segments to adjacent polygons

This step assigns the partitioned segments of uncertain areas to adjacent polygons. Two different procedures can serve the purpose: we could merge the segments with adjacent polygons or we could shift the boundaries of adjacent polygons to the centerline created by straight skeletons. We adopted the latter procedure since it allowed us to get the matching pairs of vertices before and after the conflation.

4. Experiment

We have implemented the algorithm described in this paper with the C++ programming language, which allows input and output from a large variety of data formats common in GIS. The program used Google API so that the realistic background images of interested areas can be displayed.

We used the cadastre datasets of Suwon city (perimeter: 69,960 meters, area: 121.01㎢) that are used for KLIS (Korea Land Information System). They include 4,704 meters of uniform data resolution area without weight value and 7,513 meters of varying data resolution area with weight value. Since the target area is covered by datasets independently developed by multiple local governments, uncertain areas are densely populated along the administrative boundaries. Figure 3 shows exemplar uncertain areas, and results of the algorithm application along with mash-up with Google maps backdrop.

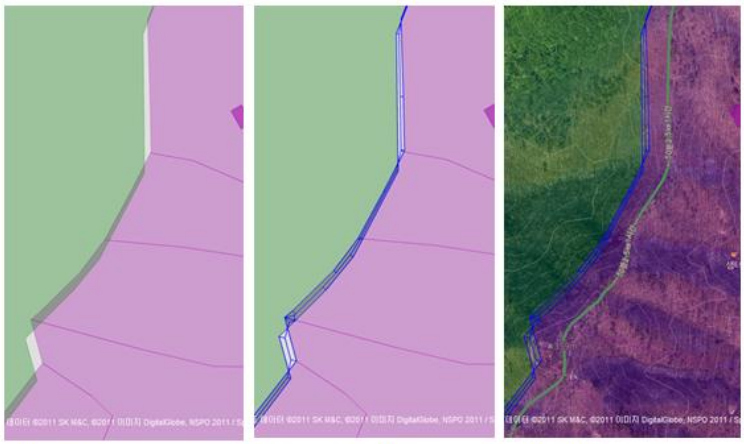


Figure 3. Exemplar Result of the algorithm

For evaluation of the results, we have compared the result with manual conflation of the same area performed by MLTM (Ministry of Land Transportation and Maritime affairs). We have thoroughly studied accuracy of location of linear features compared to the work done by high-skill experts as suggest by Goodchild and Hunter (Goodchild and Hunter, 1997). The comparison is carried out by using buffers to determine the percentage of line from one dataset that is within a certain distance of the same feature in another dataset of manual work (Figure 4).

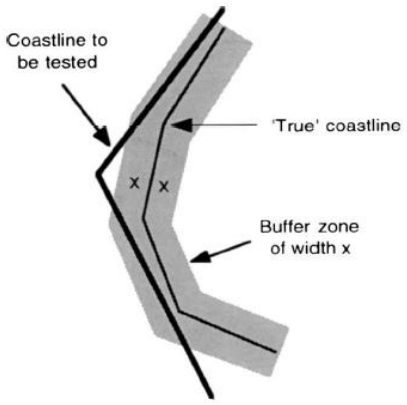


Figure 4. Goodchild and Hunter buffer comparison method. The buffer of width x is created around the reference source, and the percentage of the tested source that falls within the buffer is evaluated (Goodchild & Hunter, 1997).

The result of comparison is shown in Table 1: In areas without weight value, 90 percentile is approximately 0.8 meter which proves the algorithm to be quite effective. On the other hand, in the areas of weight values, the comparison with manual conflation yields around 90% percentile at buffer width of 12 meters. In the areas of weight values, polygons adjacent to uncertain areas are from datasets of varying data resolutions. In this case, manual conflation was performed by adjusting the lower data resolution boundaries to the higher data resolution boundaries. The weighted straight skeleton will produce better result in this case.

Table 1. Comparison with manual conflation

Same data resolution area (no weight factor)		Different data resolution area (with weight factor)	
Buffer width (m)	**Length of centerline within buffer(m)**	**Buffer width (m)**	**Length of centerline within buffer(m)**
0.1	631.5(13.4%)	**1**	601.4(7.9%)
0.5	3523.5(74.9%)	**2**	1571.9(20.9%)
1	4480.7(95.3%)	**3**	2382.8(31.7%)
2	4637.5(98.6%)	**5**	4041(53.8%)
3	4692.9(99.8%)	**10**	6245.4(83.1%)
5	4702.3(99.9%)	**15**	7512.6(99.9%)

5. Conclusions

This paper presents an algorithm for partitioning polygons with uncertain boundaries. It works in three steps: CDT based uncertain area searching, straight skeleton operation to the uncertain polygons, and allocation of partitioned segments to adjacent polygons. We showed that the algorithm is efficient and effective. The algorithm provides two advantages over existing methods: it does not require user-defined tolerance for the uncertain area search, and it maintains the vertices shift history throughout the conflation

This algorithm produces highly accurate results for areas with uniform resolution. However, for areas with varying data resolutions, methods for automated weight assignment should be further studied. We leave this as a future work.

References

Beard, M.K., Chrisman, N.R.(1988), Zipper: a localized approach to edge matching. *The American Cartographer* 15, pp. 163-172.

Chrisman, R. (1987), Efficient digitizing through the combination of appropriate hardware and software for error detection and editing. *Journal of Geographical Information Systems* 1 (3), pp. 265–277.

de Berg, M., van Kreveld, M., Overmars, M., Schwarzkopf, O.(1997), *Computational Geometry Algorithms and Applications*. Springer, Berlin, pp. 365.

Goodchild, M. F. and Hunter, G. J. (1997), A simple positional accuracy measure for linear features, *International Journal of Geographical Information Science*, 11(3): pp. 299-306.

Klajnsek, G. and Zalik, B. (2005) Merging polygons with uncertain boundaries. *Computers & Geosciences* 31: pp. 353-359

Laurini, R., Milleret-Raffort, F.(1994), Topological reorganisation of inconsistent geographical databases: a step towards their certification. *Computers & Graphics* 18 (6), pp. 803–813.

Mirela Tanase and Remco C. Veltkamp(2004), A straight skeleton approximating the medial axis. In Susanne Albers and Tomasz Radzik, editors, *Proc. 12th Eur. Symp. Algorithms(ESA 2004),* number 3221 in Lecture Notes in Computer Science, pp 809-821, Springer-Verlag.

Tinghua Ai and Peter van Oosterom (2002). Gap-tree extensions based on skeletons. In D. Richardson and P.J.M. van Oosterom, editors, Advances in Spatial Data Handling, *10th International Symposium on Spatial Data handling*, pp. 501–513.

Ubeda, T., Egenhofer, M.(1997), Advances in Spatial Databases, *Fifth International Symposium on Large Spatial Databases, SSD'97*, Lecture Notes in Computer Sciences.Berlin, Springer, pp. 283–287.

Preliminary assessment of the positional accuracy of a QuickBird ortho image

Nuno Afonso[1], Ana Fonseca[1], José Nuno Lima[1], Teresa Santos[2], Sérgio Freire[2], Ana Navarro[3] & José António Tenedório[2]

[1] Laboratório Nacional de Engenharia Civil, Av. do Brasil, 101, 1700-066 Lisboa
nafonso@lnec.pt, anafonseca@lnec.pt, jnplima@lnec.pt
[2] e-GEO, Faculdade de Ciências Sociais e Humanas (FCSH), Universidade Nova de Lisboa
teresasantos@fcsh.unl.pt, sfreire@fcsh.unl.pt, ja.tenedorio@fcsh.unl.pt
[3] LATTEX-IDL, Faculdade de Ciências (FCUL), Universidade de Lisboa
acferreira@fc.ul.pt

Abstract

This work, accomplished in the framework of the GeoSAT research project, had as main goal the definition of methodologies to extract geographic information from high-resolution satellite images in an urban environment. This paper presents and describes the positional accuracy assessment processes of a QuickBird satellite ortho image of the city of Lisbon.

To evaluate the possible applications of an ortho image for updating the Lisbon municipality base maps, some geometric pre-processing operations were previously performed, using the 1/1000 municipality base map as a reference source.

In order to obtain a significant sample of independent ground control points (GCP), a set of GNSS (Global Navigation Satellite System(s)) observations were used to determine its ground coordinates. These GCP were then used as an information reference, considering the specifications of the final product.

Results have shown that was not possible to reach the specifications of the 1/1000 scale positional accuracy, this was only verified for scales smaller than 1/5000. Alternatively, the ortho image produced in this study might be used to perform other tasks, like the extraction of features relevant for updating municipal master plan (PDM), at 1/1000 scale for urban municipalities and at 1/25000 scale for rural municipalities.

Keywords: Positional Accuracy, Ortho, Lisbon, QuickBird, GNSS.

1 Introduction

High-resolution satellite images, namely the ones from *Ikonos* and *QuickBird* satellites, can fulfil several needs of geographic information in a variety of applications, mainly in the urban environments. The GeoSAT research project has contributed for the development of methodologies of information extraction from high-resolution images, as well as, for quality assessment of those images.

The required quality for geographic information depends on its final purpose, thus, the existence of a quality control system is needed, to impose quality assessment parame-

ters and maximal errors thresholds that determine the acceptance or rejection of the data depending on their use (Santos *et al.*, 2009).

The evaluation of the quality of georeferenced information might be done considering various components, such as positional, thematic and temporal accuracy, logic consistence and completeness. In particular, the positional accuracy consists on assessing the difference between the map coordinates and the correspondent ground coordinates measured in the field using a high precision technique, like the spatial positioning systems (GNSS observations) or topographic methods. These quality components are properly identified and documented, in the ISO 19113 and ISO 19114 standards, which defines, the magnitudes to use and the quality assessment procedures respectively (Santos *et al.*, 2009; Freire *et al.*, 2010).

This work is focused on the positional accuracy assessment of a *QuickBird* ortho image, using as reference for the orthorectification process the municipality base map at scale 1/1000. The quality control of the ortho image, namely in terms of positional accuracy, was performed using as reference data, a regular sample of ground control points coordinated in the field, through GNSS observations (Casaca, 1999; Casaca *et al.*, 2005).

2 Study area and information available

According to the GeoSAT project purposes, the main study area was the city of Lisbon municipality, in Portugal. A *QuickBird* image (Figure 1) collected on March 11, 2007, including five bands, one panchromatic and four multispectral (visible spectra and infrared), with spatial resolutions of 0.60m and 2.40m, respectively, was used. The radiometric resolution of the image is 11 bits.

Regarding the vector information, the Lisbon municipality base map at scale 1/1000, in the ETRS89-TM06 reference system, was used. The altimetric data used in this work, was extracted from this base map, namely, elevation points and contours with 1m of equidistance.

Figure 1. Extract of the *QuickBird* pansharp image of Lisbon city (collected in 2007).

3 Methodology

3.1 Pre-processing

During this stage, several tasks were performed over the spectral bands and altimetric information, as well as, image geometric correction, in order to register all the information in the same reference system. All the operations were achieved with the image processing software *PCI Geomatics 10*.

The multispectral bands were merged with the panchromatic image using the PANSHARP algorithm. This operation is performed to combine the multispectral information with the panchromatic band spatial resolution, retrieving an RGB composition (visible spectra and infrared bands) with the spatial resolution of the panchromatic band (0.60m). This sort of information benefits the control points identification and posterior image segmentation and feature extraction operations.

A Digital Elevation Model (DEM) with a pixel size of 0.5m was generated using the elevation points and contours extracted from the 1/1000 base map. For this DEM generation, it was considered only elevation data on the soil, therefore the building's roofs were not corrected and had maintained folded due the distortion introduced by the relief.

The pansharpened image was then geometrically corrected to assign the image to a known reference system (in this case, PT-TM06 ETRS89) and also to decrease the geometric distortions introduced by the relief. Then, the image was orthorectified based on the RPC (*Rational Polynomial Coefficients*) model, where the polynomial coefficients extracted from the original image are used, which allows the calculation of the function that describes the geometry of the image. To perform this operation, a set of 29 ground control points distributed over the municipality, were collected, and a set of 22 checkpoints for the internal quality assessment of the model applied in this process. The results of orthorectification process and internal positional accuracy of the image (with the 22 checkpoints) are presented in Table 1. It can be verified that the variations of RMSE (*Root Mean Square Error*) in X (east-west) and in Y (north-south) are approximate half of the pixel size of the image.

Table 1. Results of orthorectification and internal positional accuracy of the pansharp image.

Points	#	RMSE X (m)	RMSE Y (m)	RMSE (m)
Control	29	0.42	0.36	0.55
Check	22	0.42	0.47	0.63

3.2 GNSS observations campaign

The availability of rigorous and independent data is essential to promote a good quality assessment process, namely an external assessment of the image. Therefore, a set of GNSS (simultaneous GPS and GLONASS signals receiver) observations was acquired, around the Lisbon city during 3 years, in order to have an acceptable sample of rigorous ground control points (GCP) coordinates. Those points were previously selected on the pansharpened *QuickBird* image (Figure 3), to obtain a regular distributed sample of control points along the city of Lisbon municipality.

In Figure 2, is illustrated an acquisition of GNSS observations for a ground control point.

Figure 2. Collection of a GCP coordinates through GNSS observations.

Figure 3. Extract of the *QuickBird* pansharpened image with the reference GCP distribution.

During this campaign, the regular distributed set of points selected in the image, was observed with a dual frequency receiver, in a differential mode (rapid-static), with 10 degrees of elevation mask, and at a sampling rate of 5 seconds for five minutes. Coordinates were obtained after post-processing.

From the 400 GCP that were collected, only the ones with a standard deviation of less than 5 cm in planimetry and less than 10 cm in altimetry were considered as valid points. Thus, the coordinates of 348 GCP were considered as reference information for the assessment of the external positional accuracy.

3.3 External quality assessment

To check the geometric quality of the image, the level of agreement between the coordinates collected in the ortho image and the ones observed with the GNSS observations were evaluated. For the external positional accuracy assessment of the ortho image, the planimetric coordinates of 70 control points were manually selected in the image, using for reference data the ground control points of the GNSS observations. Therefore the global RMSE retrieved was of 0.76m.

Results of the positional accuracy assessment performed with the *QuickBird* ortho image using external points to the orthorectification process are presented in Table 2. Using an external quality assessment, it could be verified that the RMSE achieved is slightly higher than when comparing with the internal quality assessment (0.63m), which was more focused in the behaviour of the orthorectification model.

Table 2. External positional accuracy assessment of the pansharp ortho image.

Points	**#**	**RMSE (m)**	**90% sampling points with deviation < 1.517*RMSE (m)**
GNSS	70	0.76	64 points with deviation < 1.15m

4 Analysis of results

The results obtained in the internal and external quality assessment shows that, as it was expected, the results of the internal assessment are more favourable (normally used by the producers), than the results of the external quality assessment (normally used by the users), which is the most accurate methodology to evaluate the quality of the cartography products.

Regarding the GeoSAT objectives, it has been considered to test the positional accuracy on scales for maps frequently used in the municipal and urban level, like 1/1000 and 1/5000, despite these scales, are beyond the mapping scales suitable for the pansharpened *QuickBird* image (Freire *et al.*, 2010).

Table 3 shows the cartographic constraints from the technical specifications adopted by the Portuguese Geographic Institute as the National Authority for Geodesy, Cartography and Register (NA) for orthorectified images and the Topographic Numeric Model (TNM) at scales 1/1000 and 1/5000.

Table 3. Specifications for the planimetric tolerance for selected scales of TNM (IGP, 2009).

	Tolerance	
Scales	**RMSE (m)**	**90% sampling points with deviation < 1.517*RMSE (m)**
1/1000	< 0.18	0.27
1/5000	< 0.75	1.25

Results show that the extraction of cartographic features (buildings, roads, etc.), from the ortho image, are consistent with the specifications of positional accuracy for orthos and the TNM for the 1/5000 scale and lower. The use of 1/1000 base map scale, of the Lisbon municipality, will lead to a final product with less quality specifications than the

base product, therefore not allowing its use for updates at this scale. Thus, data could be used by the municipality to fulfill other needs for cartographic purposes at smaller scales.

5 Conclusions

From this preliminary analysis, we can wrap up that the final product, using this orthorectified image, do not meet the specifications of the positional accuracy for large scale cartography, such as scale 1/1000, not allowing its use for cartography updates. Only for scales equal or inferior than 1/5000, it can be assured that the final product could have the compliance to reach the quality specifications for the positional accuracy. Examples of applications, consistent with the positional accuracy obtained, are the ones related with the features extraction for analytical purposes, such as, the calculation of indexes for urban soils imperviousness, the extraction of green areas and the creation of monitoring systems for land soil use transformation (conversion of non-urban land into urban land). The use of this type of satellite ortho images might be considered to generate indexes to quantify changes between updates of topographic base maps. This information might then be used as an argument to justify the need of more periodical cartographic updates.

The presented specifications, from the NA, are more consistent (and reached), when the production of data is obtained from image stereo pairs, where the constraints for the geometry image acquisition of aerial photographs are very strict and the information extraction is performed in a 3D environment.

References

Casaca, J. (1999), A Avaliação da Qualidade Posicional de Cartografia Topográfica a Escalas Grandes, Série Comunicações, LNEC, Lisboa.

Casaca, J., Matos, J., Baio, M. (2005), Topografia Geral, LIDEL, Lisboa.

Freire, S., Santos, T., Navarro, A., Soares, F., Dinis, J., Afonso, N., Fonseca, A., Tenedório, J.A. (2010), Extraction of buildings from QuickBird imagery for municipal planning purposes: quality assessment considering existing mapping standards, Proceedings of the 3rd International Conference on Geographic Object-Based Image Analysis (GEOBIA 2010), Ghent, Belgium.

IGP (2009), Exactidão e precisão posicionais para a cartografia nas escalas 1:1000, 1:2000, 1:5000 e 1:10000.

Santos, T., Freire, S., Portugal, I., Fonseca, A., Tenedório, J.A. (2009), Accuracy Assessment of features extracted from QuickBird Imagery for urban management purposes, Proceedings of the 33rd International Symposium on Remote Sensing of Environment (ISRSE), Stresa, Italy.

http://www.igeo.pt/servicos/Autoridade_Nacional/INTERNET_precisoes_para_1k_2k_5k_10k.pdf

Colour Coded Traffic Light Labeling: An Approach to Assist Users in Judging Data Credibility in Map Mashup Applications

Nurul Hawani Idris[1,2,3], Mike J. Jackson[1] & Robert J. Abrahart[2]

[1] Centre for Geospatial Science, The Nottingham Geospatial Building, University of Nottingham, Triumph Road, Nottingham NG72TU
[2] School of Geography, Sir Clive Granger Building, University Park, Nottingham NG7 2RD
[3] Department of Geoinformatics, Faculty of Geoinformation and Real Estate, University Teknologi Malaysia, 81310 UTM Skudai Johor Malaysia
lgxhi1@nottingham.ac.uk; mike.jackson@nottingham.ac.uk; bob.abrahart@nottingham.ac.uk

Abstract

Nowadays, the capture of location referenced data and the development of web mapping by the general public, who do not have formal training or remit, has become commonplace. However, little empirically based guidance exists in the literature to assist amateur and professional data producers design a map mash-up so as to convey credibility and quality of the information communicated. Likewise, there is also little material accessible to the mass consumer to guide them in judging the quality of presented map information. The present study examines the impact of textual metadata and graphical quality indicators in users' judgement about the credibility of map mash up information. Experimental self-completed questionnaires to a large number of respondents have been conducted. An approach of using a simple traffic light scheme has been tested. The findings demonstrate the low influence of textual metadata compared to visual based indicators in users' judgement about the credibility of a map-mash up. The authority element that is a prominent criterion in judging content credibility in traditional mapping was not consistently perceived as important in the map-mash-up experiments carried out in this research. The paper's findings are an important step in understanding how the mass of people choose and evaluate credibility in web map information, particularly in map mash ups. The findings support the growing research emphasis on promoting quality awareness among web map users.

Keywords: mashups, metadata, credibility, trust, web mapping, quality

1 Introduction

The Web 2.0 revolution, 'Digital Earth' vision and recent location technology advancement have had a big impact on trends with regard to mapping culture, and given rise to the so-called sub-discipline of neogeography (Haklay *et al.* (2008)). Until the appearance of neogeography, mapping activities were mainly dominated by professional developers. The data used in these applications are typically supplied by government or com-

mercial organisations that have standardised procedures for its capture, quality control and dissemination. However, the ease of capture of geographic and location-referenced data and of map construction without the need for high cost software or programming skills, experience or prior knowledge on mapping design means many citizens now can and do produce their own maps (Haklay *et al.* (2008)).

This current trend however, raises the concern about the truth, quality and accuracy of information being conveyed. In fact, this issue have been widely debated in other domains that use the World Wide Web as a medium for dissemination. For example, the ecommerce and health information ethics domain have implemented third party seal programs such as TRUSTe (Cheskin, 1999) and HONcode certificates (Fallis and Fricke, 2002). These programs provide a 'stamped trust label' for websites that adhere to the ethics guidelines in their respective fields. Rating systems that are calculated, based on peer reviews, such as used in eBay, and communicated using visual labelling, such as colour coded traffic light (CCTL), have been applied to certain product reviews.

Visual quality indicators on GIS maps were first suggested in Devillers *et al.* (2002:700) as a response to the difficulty of communicating metadata to professional and lay users. The challenges of using the typical textual form of metadata to assist users analysing the data's fitness for purposes have been stated in several publications. Devillers *et al.* (2007) has posited a practical model to implement a quality rating system in GIS for use by experts to give advice about the quality of a dataset. The approach of using visual indicators as an aid to judging the quality of a map has a potential to be further tested in a Web medium, particularly in map mash up environment. This study therefore, proceeds to examine the potential of a CCTL rating label in map mash-ups to assist users in making informed judgment about the information credibility based on the critical values.

2 Methodology

2.1 Sample

Two series of experiments using online map based questionnaires were conducted. Both experiments were distributed to respondents on the basis of convenience (opportunity) sampling (Black, 2009) via students mailing lists and the university internet portal. There were 208 respondents aged between 18 and 35 involved in these two experiments. The sample comprised of members of the University of Nottingham and was split into two groups of respondents: geoliterate and non-geoliterate users. The groups were classified based on the background information given by each respondent in the user demographics form. Geoliterate users were grouped based on those who stated that they had current or previous backgrounds (i.e. had attended academic or/and professional courses) in geography, cartography, remote sensing, land surveying or geographic information science: the remainder were classified as non-geoliterate users. Table 1 shows the demographic data for the two experiments.

Table 1: Demographic data for each of experiments

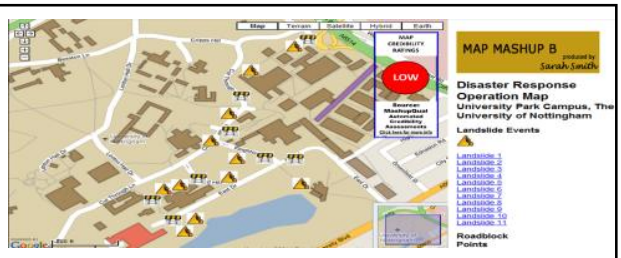
	Exp 1 (n = 133)	**Exp2 (n = 75)**
Geo-literate respondents	31	28
Non geo-literate respondents	102	47

2.2 Experimental Datasets and Response Measure

Both experiments used similar datasets as in Figure 1. Dataset 1 was to simulate a high credibility map by indicating 'The University of Nottingham' as the mash-up producer. Dataset 2 was to simulate a low credibility mash-up produced by an individual (viz. *Sarah Smith*) without presenting any credential information about the author. The information about the author/creator of the mash-up was displayed on the right top of the side bar. The differences between the two experiments were on the experimental tasks and the presence of the CCTL label. The CCTL label was tested in Experiment 2 but not in Experiment 1.

The developed questionnaires were informed by previous study in Fogg *et al.* (2003) and a pilot survey conducted earlier. The dependent variables included in the questionnaire and referred to as 'elements' in the rest of this paper, were the identity of mash-up producer (to represent authority), and ratings label for user credibility judgment. This study has measured other variables such as map data supplier, website affiliation, colour scheme and currency as demonstrated in Idris *et al.* (2011). Other possible variables suggested in Fogg *et al.* (2003) that might influence the credibility such as the accuracy, functionality, performance, readability, information bias and advertisement were excluded in this study. The simplification applied is believed not invalidate the findings demonstrated in previous studies. The label was designed using the concept of CCTL plus an overall rating (TF+), as demonstrated by Kelly *et al.* (2009). Figure 1 depicts the labels on top of the maps.

Dataset 1: Mash-up A



Dataset 2: Mash-up B

Figure 1: The datasets and labels used in Experiment 2

3 Results

Finding 1: The authority of map mashup information (i.e. the producer of a map) was not perceived as an important element for a high percentage of users determining credibility of an online map mashup

In Experiment 1, respondents were asked two similar questions, but in different contexts. In the first context, respondents have to give their responses purely based on the experimental task, but in the latter context respondents have to rate their responses in purely generic context based. In the first question (Q1), respondents were asked 'how important map producer (map author) in influencing you to choose the map in the experimental task'. In the second question (Q2), respondents were asked to rate 'how important the map producer's (map author's) reputation when in influencing you in assessing a credibility of any online map', which is a generic question, not specific to the experimental task. A four-point scale (0 = do not know, 1= not important, 2= slightly important, 3=important) was used to measure the responses. From the analysis on Q1 shown in

Figure 2, sixty-nine (51.88%) of respondents rated the map producer (author) element as not important in influencing their decisions in the experimental task. Analysis on Q2 indicated thirty (22.56%) of respondents rated map producer element as not important. The responses between two questions were inconsistent and the difference was highly statistical significant in the chi-squared test $\chi^2(1) = 15.364$, probability (p) < 0.001 (in probability there is only one chance in a thousand the different could have happened by coincidence).

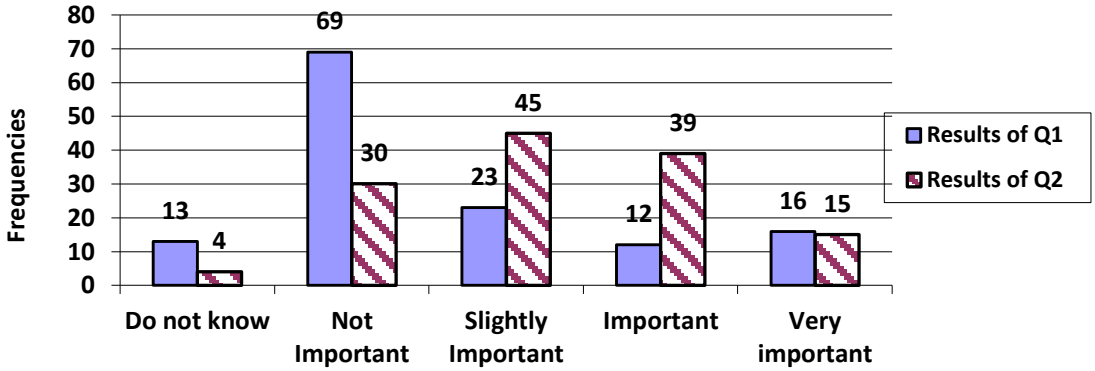


Figure 2: The comparison between the responses in Q1 and Q2

Finding 2: Participants were three times more likely to identify a high credibility map mash up from the CCTL labelling than from a textual label indicating authority of the data.

From the graphs in Figure 3, the number of respondents that chose Mash-up A (high credibility map) in Experiment 2 was increased 27.24 per cent over the responses in Experiment 1.This is highly significant as measured by the chi-squared test $\chi^2(1)= 31.44$, $p < 0.001$. From the test, there was a significant association between the presence of the CCTL label and whether the respondents chose the high credibility map (mash-up A) or low credibility map (mash-up B). From the odds ratio calculation (Field, 2009:700), the odds of respondents choosing the high credibility map (map A) were 3.3 times higher if they were given a map with a CCTL label than a map without the CCTL label. Figure 4 indicates results where respondents have to rate their agreement on the statement of 'I chose the map because I have been influenced by the credibility ratings provided with the map'. 54.67% of respondents rated their responses within the scale of slightly important (5) to very important (7), and 30.67% rated this element as slightly 'not important' (3) to very 'not important' (1); this difference was statistically significant [chi-squared test $\chi^2(1) = 5.06$, $p<0.05$].

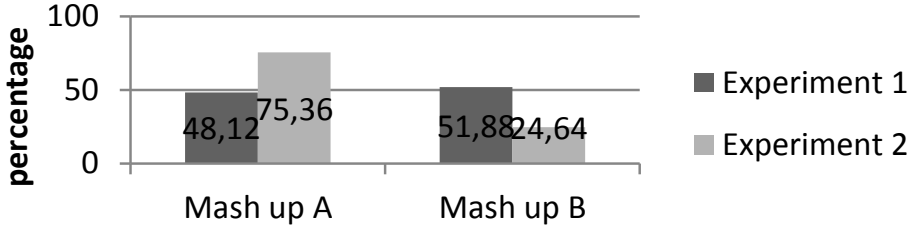


Figure 3: The results for the two experiments of the question 'which of the two maps do you perceive as having more credibility (believability) to assist you in the task - Map A or Map B?'

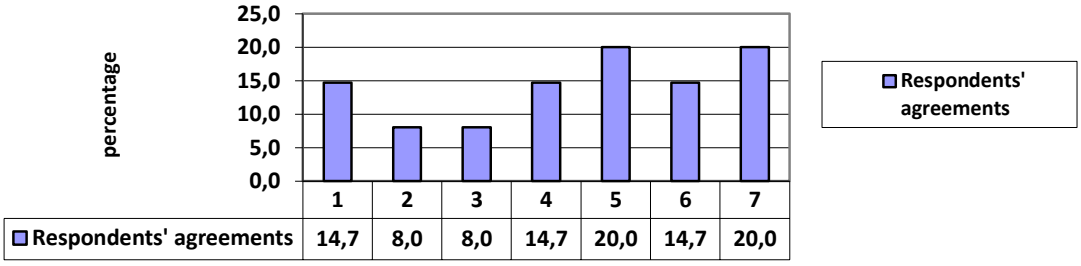


Figure 4: The respondents' agreement on the statement of 'I chose the map because I have been influenced by the credibility rating provided with the map'

4 Discussions and Conclusions

This study demonstrates the significant difference in response when respondents were asked a question about the level of influence of a map producer element in two different questioning approaches – an experimental task-based context and non task-based based context. It indicates respondents were aware of the importance of the identity of the map producer in judging credibility of information, but would not invariably include that element in their assessment. A number of observations may be made with respect to this finding; 1) the element might not have sufficient visual prominence when judging the credibility in the tasks. As in Prominence-Interpretation Theory (Fogg, 2003), users make a judgment pertaining to credibility based on the elements that they noticed. To address this, in Q1, a prominence test approach was used. Q2 used an interpretation test approach where respondents interpreted the statement in a generic context so that the prominence of the element was not relevant. As Fogg's *et al.* (2003) research predicted, the responses between the two-fold questioning approaches were different. 2) The Web is in an author-less environment where information in the web is not subjected to assertion of the identity of the author; hence the emphasis on importance of authority in judging information credibility as in traditional approach has diminished (Warnick, 2004). Our results are consistent with these earlier studies.

This study also demonstrates the impact of a colour coded traffic light ratings (CCTL) label for online mapping and particularly for map mash-ups. The probability of respondents making informed judgments by choosing a high credibility map based on the CCTL rating label is three times higher than the setting without the label. This finding is in line with the study of Kelly *et al.* (2009) which demonstrated that the probability of users identifying the healthier food from the traffic light labelling format was five times higher than with a label using monochrome text information. The knowledge that focusing on visual cues to attract users' attention in low motivation groups is well established in marketing and advertising. As Petty and Cacioppo (1986) identified, users in such groups would rely on the peripheral signal rather than critical element. The Fogg *et al.* (2003) study supports that finding by demonstrating visual related elements as major indicators that users used to determine credibility of online information. The proposed CCTL ratings label has the potential to provide a solution for a group of users who are sometimes in low motivation to scrutinise the critical element when judging the credibility of information. Given the selection criteria used in this study for the experiment participants, however, the validity of the findings might only apply to young adult map users and the

findings might be different if all respondents were required to have a deep engagement with the task.

In conclusion, the development of a CCTL rating label for online maps, and particularly for map mash-ups is showing promise as an approach for assisting users to make informed judgement about the credibility of online map information in a quick and easy manner. Further investigations have to be made on how to increase the influence of the ratings label so that it could become the main element being measured in users' credibility assessments. A model of a credibility ratings index for map mash-ups and the conceptual design of the implementation of automatic CCTL creation in map mash-up environments are the next steps for this current research.

References

Black, T. R. (2009). *Doing Quantitative Research in the Social Sciences*, SAGE, Great Britain., 125p.

Cheskin. (1999). Ecommerce Trust Study [Electronic Version]. Available at: from http://www.cheskin.com/cms/files/i/articles//17_report-eComm%20Trust1999.pdf. [Accessed on 18 June 2009].

Devillers, R., Gervais, M., Bedard, Y., & Jeansoulin, R. (2002). *Spatial Data Quality: From Metadata to Quality Indicators and Contextual End-User Manual*. In: The OEEPE/ISPRS Joint Workshop on Spatial Data Quality Management, Istanbul, Turkey, pp. 45-55.

Devillers, R., Bedard, Y., Jeansoulin, R., & Moulin, B. (2007). "Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data". *International Journal of Geographic Information Science*, Vol. 21(3): 261-282.

Fallis, D., & Fricke, M. (2002). "Indicators of accuracy of consumer health information on the Internet: a study of indicators relating to information for managing fever in children in the home". *Journal of the American Medical Informatics Association: JAMIA,* Vol. *9*(1), 73-79.

Field, A. (2009). *Discovering Statistics Using SPSS* (Third ed.). SAGE, Dubai., 550p.

Fogg, B. J., Cathy, S., David, R. D., Leslie, M., Julianne, S., & Ellen, R. T. (2003). *How do users evaluate the credibility of Web sites? A study with over 2,500 participants*. In: *Proceedings of the Conference on Designing for User Experiences*, California, USA, pp.1-5.

Fogg, B. J. (2003). "Prominence-interpretation theory: explaining how people assess credibility online". *Conference on Human Factor in Computing System (CHI 200*3). Florida, USA, pp. 722-723.

Haklay, M., Singleton, A., & Parker, C. (2008). "Web mapping 2.0: The neogeography of the GeoWeb". *Geography Compass*, Vol. 2(6), 2011-2039.

Idris, N. H., Jackson, M.J., & Abrahart, R. (2011). "Map Mashup: What looks good must be good?". *Conference on GIS Research U* . Portsmouth, UK, pp. 137-143.

Kelly, B., Hughes, C., Chapman, K., Louie, C., Dixon, H., Crawford, J., King, L., Daube, M., & Slevin, T. (2009). Consumer testing of the acceptability and effectiveness of front-of-pack food labelling systems for the Australian grocery market. *Health Promotion International,* Vol. 24(2), 120-129.

Petty, R., Cacioppo, J. (1986), "The elaborating likelihood model of persuasion". *Advances in Experimental Social Psychology,* Vol. 19: 179-190.

Warnick, B. (2004), "Online Ethos: Source Credibility in an 'authorless' environment". *American Behavioral Scientist*, Vol. 48(2): 256-265.

Multi-Scale Analysis Approach of Simulating Urban Growth Pattern using a Land Use Change Model

Amir Hossein Tayyebi[1], Saeid Homayouni[1], Jie Shan[2], Mohammad Javad Yazdanpanah[3] , Bryan Christopher Pijanowski[4] & Amin Tayyebi[4]

[1] Department of Surveying and Geomatics Engineering, School of Engineering, University of Tehran, Iran
amirhossein.tayyebi@gmail.com, saeid.homayouni@gmail.com
[2] Purdue University, School of Civil Engineering, Department of Geomatics Engineering, West Lafayette, IN 47907-2051
jshan@purdue.edu
[3] Control & Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University of Tehran, P.O. Box 14395/515, Tehran, Iran
yazdan@ut.ac.ir
[4] Purdue University, School of Agriculture, Department of Forestry and Natural Resources, West Lafayette, IN 47907-2051
bpijanow@purdue.edu, amin.tayyebi@gmail.com

Abstract

Spatial-explicit models are usually used to study the pattern of land use and land cover (LUCC) changes. It has been widely cited in LUCC field that the scale of observation can influence the outcome of an analysis and impact the model parameters that describe urban land use change processes. Thus, different urban land use patterns can be formed at different spatial and temporal scales. Land Transformation Model (LTM) integrates Artificial Neural Network (ANN) and Geospatial Information system (GIS) to simulate urban land use change pattern with socio-economic and environmental variables. Extent is overall area of region under research while resolution is minimum detectable area that can be measured. The main objectives of this paper are: (1) Analyze the parameters and outputs of LTM in response to varying cell resolutions and (2) Examine the pattern of urban growth of LTM across different cell sizes. This experiment analyzed the objectives of this paper in Muskegon River Watershed of USA. Although the LTM captures the pattern of urban growth across cell sizes, differences between the observed patterns across scales were substantial. Sensitivity analysis of urban growth across cell size indicates that LTM perform better at certain cell sizes than at others. The results of this paper can help urban land use modelers to determine the appropriate spatial scales for the processes being simulated and the importance of spatial scale in urban land use change modeling research.

Keywords: Land use/ land cover change, spatially-explicit models, Uncertainty, Multi-Scale.

1 Introduction

Urban land use change patterns can be different over time and space (Gibson *et al*. 2000, Goetz *et al*. 2004). Spatial scale can influence the model's ability to capture the

patterns that drive urban land use changes. It also becomes more complicated when the patterns of urban land use change are not usually related across different spatial scales in a simple way (Gardner *et al*. 1989, Jenerette and Wu 2001, Kok and Veldkamp 2001). There are recent examples of urban land use modeling frameworks that integrate multiple scales (Verburg and Chen 2000, Walsh *et al*. 2001, Soares-Filho *et al*. 2002). Some Urban Growth Models (UGMs) were designed to simulate urban land use change at fine scale (Bockstael 1996) while other UGMs were developed to simulate the patterns at coarse scale (Johnston and Barra 2000). According to White and Running (1994), scale is "both the limit of resolution where a phenomenon is discernible and the *extent* that the phenomenon is characterized over space and time".

In spatially-explicit models, identification of driving forces of change is of great importance. Besides environmental constraints (such as altitude, rainfall, slope and temperature), demographic and other socioeconomic variables, like location of cities, population density, or level of education, are main spatial determinants of land use (Kok, *et al*., 2001). In several land use models, statistical tools and transition rules are being used to analyze in gridded spatial data (SLEUTH (Clarke *et al*., 1997), LTM (Pijanowski *et al*., 2002), CLUE (Veldkamp *et al*., 2001; Verburg *et al*., 2002), GEOMOD (Pontius *et al*., 2001)). Land use change models are capable of examining the sensitivity of land use patterns in different spatial and temporal dimension and the stability of linked social and ecological systems, through scenario building. LTM can simulate land use change using artificial neural network, from which complex patterns emerge (Pijanowski *et al*., 2002). LTM focuses on grids, simulating the relationship between inputs (as pixels) and an output (as pixels). LTM was incorporated in different forms during the last decade to accomplish spatially urban explicit land use change modeling. For example, LTM has been employed to create land use legacy map (Pijanowski *et al.*, 2007), historical land use maps using back-casting approach (Ray and Pijanowski, 2010), urban boundary change (Tayyebi *et al*., 2011), as well as examining error propagation in coupled LTM and Regional Atmospheric Model System (RAMS) models and error propagation in a coupled land use and ground water model (Pijanowski *et al.*, in review and Ray *et al.*, in review).

It is vital how to represent drivers in the model and how the model does respond to scale analysis. Multiple processes that form land-use are scale dependent, so they act differently on various scales. Extent and resolution are two aspects of scales while they are not practically independent. As resolution reduces and extent increases, identifying major processes becomes more difficult. For each process in land use and land cover change, a range of scales may be defined over which it has a significant influence on the land use pattern (Meentemeyer, 1989; Dovers, 1995). Using two approaches, we can deal with scale dependency: 1) Fixed spatial units (grids), as extent and resolution varies (Walsh *et al*., 2001; Kok and Veldkamp, 2001); 2) Changing spatial units (Nelson, 2001). Although, the LTM has been evolved in an active area of research, few studies have explicitly addressed scale issues in urban land use simulation of LTM. Because LTM is pixel-based approach, the quantity and location of urban land use change patterns can be influenced by varying the spatial cell size of the data. In this paper, we present results from a series of widely used LTM-based urban model (slope, elevation, exclusionary zone, distance to urban, road and stream) in the Muskegon River Watershed, in the Upper Midwest USA to investigate how the parameter and outcome of LTM responds to changes in cell resolution. We also analyze the relationship of urban land use change patterns across different spatial scales.

References

Bockstael, N. E., 1996. Modeling economics and ecology: the importance of a spatial perspective. American Journal of Agriculture and Economics, 78, pp. 1168–1180.

Clarke, K. C., Hoppen, S. and Gaydos, L. 1997. A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. Environment and Planning B: Planning and Design, 24, pp. 247–261.

Dovers, S. R. 1995. A framework for scaling and framing policy problems in sustainability. Ecological Economics 12, 93-106.

Gardner, R. H., O'Neill, R. V., Turner, M. G. and Dale, V. H. 1989. Quantifying scale-dependent effects of animal movement with simple percolation models. Landscape Ecology, 3, pp. 217–227.

Gibson, C. C., Ostrom, E. and Ahn, T. K. 2000. The concept of scale and the human dimensions of global change: a survey. Ecological Economics, 32, pp. 217–239.

Goetz, S.J., Jantz, C. A., Prince, S. D., Smith, A. J., Varlygun, D. and Wright, R., 2004. Integrated analysis of ecosystem interactions with land use change: the Chesapeake Bay watershed. In Ecosystem Interactions with Land Use Change, G.P. Asner, R.S. DeFries and R.A. Houghton, 13, pp. 212–225.

Jenerette, G.D. and Wu, J., 2001, Analysis and simulation of land-use change in the central Arizona–Phoenix region, USA. Landscape Ecology, 16, pp. 611–626.

Johnston, R.A. and Barra, T.D.L., 2000. Comprehensive regional modeling for long-range planning: linking integrated urban models and geographic information systems. Transportation Research A: Policy and Practice, 34, pp. 125–136.

Kok, K., Farrow, A., Veldkamp, A. and Verburg, P. H., 2001. A method and application of multi-scale validation in spatial land use models. Agriculture, Ecosystems and Environment, 85, pp. 223–238.

Kok, K. and Veldkamp, A., 2001. Evaluating the impact of spatial scales on land use pattern analysis in Central America. Agriculture, Ecosystems and Environment, 85, pp. 205–221.

Meentemeyer, V. 1989. Geographical perspectives of space, time, and scale. Landscape Ecology 3, 163-173.

Nelson, A., 2001. Analysing data across geographic scales: detecting levels of organization within systems. Agric. Ecosyst. Environ. 85, 107–131.

Pijanowski, B. C., Daniel G. Brown, Bradley A. Shellito, Gaurav A. Manik. 2002. Using neural networks and GIS to forecast land use changes: a Land Transformation Model. Computers, Environment and Urban Systems. 26, 553–575.

Pijanowski, B.C., D. K. Ray, A. D. Kendall, J. M. Duckles, and D. W. Hyndman. 2007. Using back-cast land-use change and groundwater travel time models to generate land-use legacy maps for watershed management. Ecology and Society 12 (2):25.[online] URL: http://www.ecologyandsociety. org/vol12/iss2/art25/.

Ray, D.K., B.C. Pijanowski, A.D. Kendall and D.W. Hyndman. In review. Coupling land use and groundwater models to map land use legacies: Using GIS to assess model uncertainties relevant to land use planning. Applied Geography.

Pijanowski, B.C., N. Moore, D. Mauree and D. Niyogi. Revised and in review. Evaluating error propagation in coupled land-atmosphere models. Earth Interactions.

Pontius R. G. Jr and L. C. Schneider. 2001. Land-use change model validation by a ROC method for the Ipswich watershed, Massachusetts, USA. Agriculture, Ecosystems & Environment 85(1-3) p. 239-248.

Ray, D. K., and B. C. Pijanowski, 2010. A backcast land use change model to generate past land use maps: application and validation at the Muskegon river watershed of Michigan, USA, journal of land use science, 5: 1, 1-29.

Soares-Filho, B.S., Cerqueira, G.C. and Pennachin, C.L., 2002. A stochastic cellular automata model designed to simulate the landscape dynamics in an Amazonian colonization frontier. Ecological Modeling, 154, pp. 217–235.

Tayyebi, A., Pijanowski, B. C., A. H. Tayyebi, 2011a. An urban growth boundary model using neural networks, GIS and radial parameterization: An application to Tehran, Iran. Landscape and Urban Planning 100, 35-44.

Tayyebi, A., Pijanowski, B. C., B. Pekin, 2011b. Two rule-based urban growth boundary models applied to the Tehran metropolitan area, Iran. Applied Geography 31, 908-918.

Veldkamp, A., Fresco. L. O. 1996. CLUE: a conceptual model to study the conversion of land use and its effects. *Ecological Modeling*, 85:253270.

Verburg, P. H. and Chen, Y., 2000. Multi-scale characterization of land-use patterns in China. Ecosystems, 3, pp. 369–385.

Verburg, P. H., Soepboer, W., Limpiada, R., Espaldon, M. V. O., Sharifa, M. and Veldkamp, A. (2002). Land use change modelling at the regional scale: the CLUE-S model, Environmental Management, vol. 30, pp. 391-405

Walsh, S. J., Crawford, T. W., Welsh, W. F. and Crews-Meyer, K. A., 2001. A multi-scale analysis of LULC and NDVI variation in Nang Rong district, northeast Thailand. Agriculture, Ecosystems and Environment, 85, pp. 47–64.

White, J. D., Running, S. W., 1994. Testing scale dependent assumptions in regional ecosystem simulations. J. Vegetation Sci. 5, 687–702.

Uncertainty Framework in Land Use Change Models: An Application of Data, Model Parameter and Model Outcome Uncertainty in Land Transformation Model

Amir Hossein Tayyebi[1], Saeid Homayouni[1], Jie Shan[2], Mohammad Javad Yazdanpanah[3] , Bryan Christopher Pijanowski[4] & Amin Tayyebi[4]

[1] Department of Surveying and Geomatics Engineering, School of Engineering, University of Tehran, Iran
amirhossein.tayyebi@gmail.com, saeid.homayouni@gmail.com
[2] Purdue University, School of Civil Engineering, Department of Geomatics Engineering, West Lafayette, IN 47907-2051
jshan@purdue.edu
[3] Control & Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University of Tehran, P.O. Box 14395/515, Tehran, Iran
yazdan@ut.ac.ir
[4] Purdue University, School of Agriculture, Department of Forestry and Natural Resources, West Lafayette, IN 47907-2051
bpijanow@purdue.edu, amin.tayyebi@gmail.com

Abstract

As sustainable development is a goal for many urban communities, land use change models have drawn more public attention because it provides local land use planners and regional resource managers with information about the potential effects of urban growth. Uncertainty is important issue for decision maker and urban planners because they should be careful in communicating the uncertainties within the urban land use maps. Having knowledge about the origin and impacts of uncertainty is needed to make the goals of reliability in urban planning. Walker et al. (2003) developed new frameworks and typologies of uncertainties for decision support fields. This study attempts to use Walker et al, (2003) framework to address the importance of assessing various dimension of uncertainty in urban growth simulation of Land Transformation Model (LTM) for more efficient urban planning. Muskegon River Watershed of USA was considered as study area to meet the objectives of this paper. We assess the uncertainties associated with LTM such as data, model parameters and model outcome uncertainties using quantity and location metrics to compare the outputs of different versions of LTM with each other as well as actual land use map. We also discuss how different sources of errors can affect the quantity and location of urban simulated maps of LTM.

Keywords: Land use/ Land cover change, Land Transformation Model, Uncertainty, Urban planning

1 Introduction

Many researchers have focused on developing Urban Growth Models (UGMs) that successfully link the spatial predictor variables to urban land use change (Veldkamp and

Fresco 1996, Mertens and Lambin 2000, Schneider and Pontius 2001, Serneels and Lambin 2001, Overmars and Verburg 2005; Pijanowski *et al.*, 2002). The reliability of these UGMs highly depends on the uncertainty in data, model structure, model parameter, model outcome and other sources (Walker *et al.*, 2003). Some research programs in the field of LUCC have quantified the uncertainty associated with urban land use change models. Many plausible scenarios exist for the uncertainty assessment and error propagation investigation in the Land Use Land Cover Change (LULCC) models. There are different sources of uncertainties that affect the UGMs simulations, but the most important ones including the data uncertainty, model structure uncertainty, model parameter uncertainty and model outcome uncertainty (Walker *et al.*, 2003). The uncertainty is accumulative in UGMs and uncertainty in each step would affect amount of uncertainty in next step. Sometimes this would cause to increase uncertainty and sometimes may lead to reduce amount of uncertainty. During the past decade, many studies have focused on assessing various sources of uncertainty associated with hydrological forecasts (Beven and Binley, 1992; Kuczera and Parent, 1998; Vrugt *et al.*, 2003; Maier and Ascough, 2006; Ajami *et al.*, 2007) and hydrologic modeling (Vrugt *et al.*, 2003). Uncertainty is increasingly important in different sciences such as environmental science (Van Der Sluijs, 2007), water management (Pahl-Wostl *et al.*, 2007) and transport planning (Marchau *et al.*, 2009). Burnicki *et al.*, (2010) recently examined the impact of varying spatial and temporal patterns of error on a post-classification change analysis.

LTM include many parameters (weight and biases) describing the urban growth patterns which need to be estimated through calibration with historical data. The parameters of LTM can take wide variety of values in each cycle and they are different according to structure of neural network and type of spatial driving forces. Investigating different dimensions of uncertainty (such as data, model structure, model parameters and model outcome uncertainties) in LTM have often been ignored or addressed indirectly. Furthermore, the links between LTM uncertainties and urban planning uncertainties have not been extensively explored. Reliable LTM urban growth simulations can provide decision makers with information that allows them to incorporate risk in decision making and therefore decrease the social, economic and environmental impact of land use change.

This study intends to build an LTM to address the three main dimensions of uncertainties in urban growth simulation associated with data, model parameters and model output uncertainty. To accomplish this objective, the paper is divided into three major parts. First, all inputs and output spatial layers perturb with different amount of quantity and location errors (0% (error-free), 5%, 10%, 15% and 20%) which refer to data uncertainty. Second, the derived data (error and error-free data) were used to build and train five versions of LTMs which refer to model parameter uncertainty. Third, the outcomes of four versions of LTM that include error in data for training were compared with outcomes of one version of LTM that did not include error in data as well as actual map which refers to model outcome uncertainty. We will illustrate that the data, model parameter and model outcome uncertainty incorporated within the LTM are major uncertainty sources in the urban growth simulation of LTM. We will also show that how different sources of errors would affect the quantity and location of final simulated maps of LTM. This allows decision makers to appreciate with the uncertainties in the LTM which respond to urban land use changes.

References

Ajami, N. K., Q. Duan, and S. Sorooshian. (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, Water Resource. Res., 43, W01403.

Beven , K., and A., Binley. (1992). The future of distributed models: model calibration and uncertainty prediction, *Hydrological Processes*, *6*, 279-298.

Kuczera, G., and E. Parent. (1998). Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The metropolis algorithm, J. Hydrol., 211, 69– 85.

Maier, H. R., and J. C. Ascough II, (2006). Uncertainty in environmental decision-making: Issues, challenges and future directions, in Proceedings of the iEMSs Third Biennial Meeting: Summit on Environmental Modeling and Software [CD-ROM], edited by A. Voinov, A. Jakeman, and A. Rizzoli, Int. Environ. Modell. and Software Soc., Burlington, Vt. (Available at http://www.iemss.org/iemss2006/sessions/all.html).

Mertens, B., and Lambin, E. F (2000). Land-cover change trajectories in southern Cameroon. Annals of the Association of the American Geographers, 90 (3).

Overmars, K. P. and P. H. Verburg. (2005). Analysis of land use drivers at the watershed and household level: Linking two paradigms at the Philippine forest fringe. International Journal of Geographical Information Science**.** 19(2): 125-152.

Pijanowski B. C., Brown D., Shellito, B. and Manik, G. (2002). Using neural networks and GIS to forecast land use changes: a Land Transformation Model. *Comp. Environ. Urban Syst.* 26, 553–575.

Schneider, L., and Pontius Jr, R. G. (2001). Modeling land-use change in the Ipswich watershed, Massachusetts, USA. *Agriculture, Ecosystems & Environment* 85, 83-94.

Serneels, S., Lambin, E. F. (2001). Proximate causes of land use change in Narok district Kenya: a spatial statistical model. Agric. Ecosyst. Environ. 85, 65–81.

Van der Sluijs. (2007). Uncertainty and precaution in environmental management: insights from the UPEM conference. Environ Model Softw, 22(5):590–8.

Veldkamp, A. and Fresco. L. O. (1996). CLUE: a conceptual model to study the conversion of land use and its effects. *Ecological Modeling*, 85:253270.

Vrugt, J. A., H. V. Gupta, W. Bouten, and S. Sorooshian, (2003). A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, Water Resour. Res., 39(8), 1201.

Walker, W., Harremo¨es, P., Rotmans, J., van der Sluijs, J., van Asselt, M., Janssen, P. & Krayer von Krauss, M. (2003). Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. Integrated Assessment 4(1), 5–18.