Spatial data quality: from description to application

Spatial data quality: from description to application

Pepijn van Oort

Publications on Geodesy 60

NCG Nederlandse Commissie voor Geodesie Netherlands Geodetic Commission

Delft, December 2005

Spatial data quality: from description to application
Pepijn van Oort
Publications on Geodesy 60
ISBN 90 6132 295 2
ISSN 0165 1706

Published by: NCG, Nederlandse Commissie voor Geodesie, Netherlands Geodetic Commission,
Delft, The Netherlands
Printed by: Optima Grafische Communicatie, Optima Graphic Communication, Rotterdam,
The Netherlands
Cover illustration: Pepijn van Oort

NCG, Nederlandse Commissie voor Geodesie, Netherlands Geodetic Commission
P.O. Box 5058, 2600 GB Delft, The Netherlands
T: +31 (0)15 278 28 19
F: +31 (0)15 278 17 75
E: ncg@lr.tudelft.nl
W: www.ncg.knaw.nl

The NCG, Nederlandse Commissie voor Geodesie, Netherlands Geodetic Commission is part of
the Royal Netherlands Academy of Arts and Sciences (KNAW)

Acknowledgements

For four years I have been working with pleasure on this thesis and it is now a pleasure to thank all those who contributed to its completion. First of all, I would like to thank my supervisors. Arnold, Sytze and Alfred, most of all I should thank you three for all your efforts in learning me to write a scientific article. Arnold, you were stimulating throughout the project, provided a great network when looking for case-studies and showed great engagement in the testing phase of the questionnaire for chapter 7. Sytze, you were the best reviewer I could have wished for. At the start of the PhD, I would get upset by your reviews on draft documents, only later to learn that your remarks were always to the point and always led to significant improvement. Alfred, your enthusiasm throughout, but especially on chapter 5, was much appreciated. Both Sytze's and your statistical support were of the ideal combination of scientific soundness and pragmatism.

The anonymous reviews on submitted articles were always to the point, they always showed great understanding of the topics addressed and always led to improvement of the submitted work. So thanks to you, anonymous reviewers. I hope you will one day also be reading other chapters of this thesis.

Perhaps I should thank the two PhD's who started on this project but quit, after 1 week and after 6 months respectively. Thanks to them I got the chance to start and complete the project. But more than these two, I should thank the agencies for funding the project and for continuing to do so after these two false starts. The PhD was funded by:

- The Netherlands Geodetic Commission (KNAW-NCG);
- The Geo-Information and ICT Department of the Ministry of Transport, Public Works and Water Management (RWS-AGI), formerly the Survey Department (RWS-MD);
- The Cadastre, Land Registry and Mapping Agency (TDK), formerly the Netherlands Topographical Service (TDN);
- The Netherlands Council for Geographical Information (Ravi);
- TNO Built Environment and Geosciences (TNO-BenO), formerly the National Geological Survey (TNO-NITG); and
- Wageningen University and Research Centre - Alterra (WUR-Alterra).

At the initial stage of the project, Bert Kolk (TDK), Co van Oogen (Ravi), Ruben Dood (RWS-AGI), Herman Thunnissen (Alterra) and Pieter Meijers (TNO-BenO) were involved. When I started some names had already changed. Twice a year we held a progress meeting with the supervising committee in which Adri den Boer (RWS-AGI), Gerard Vrijkotte (TDK), Ko van Raamsdonk (Ravi), Marcel Reuvers (Ravi), Jan-Diederik van Wees (TNO-BenO), Allard de Wit (Alterra), Sytze de Bruin, Alfred Stein and Arnold Bregt participated. I much appreciated the feedback and support received during and in between these meetings. Frans Schröder invited me to annually contribute in the KNAW-NCG annual report, and helped in preparing the thesis for publication in the KNAW-NCG series "Publications on Geodesy".

None of the chapters of this thesis could have been written without help from outside. For chapter 2, a lengthy discussion with Henri Aalders and on spatial data quality standards was much appreciated. Chapter 3 would have taken so much longer if

reference data had not already been collected and pre-processed by Allard de Wit. As the first submitted article of the thesis it required a lot of effort from supervisors, so they deserve special thanks for this chapter. Chapter 4 built on earlier work by Marthijn Sonneveld, Jeroen van den Brink, Kees de Zeeuw and Arnold Bregt. They are acknowledged for giving access to their data and model. For chapter 5, data and information about the case-study were supplied by Jasper Kuipers of the National Forest Service of the Netherlands. For chapter 6, a discussion with Bernard Cornélis and Sébastien Brunet helped shape thoughts on the approach to be taken. Floris de Bree, Arnold Bregt and friends and family of Arnold Bregt spent much time on testing of trail questionnaires. After that, a trial questionnaire was distributed at the ESRI GISTech conference (Rotterdam, 2003). This would have been impossible without permission and support from Jan Willem van Eck, organiser of the conference. For the final questionnaire, respondents were personally contacted. Anne Schmidt, Gerard Nieuwenhuis, Ludolph Wentholt, Wim Cofino and Arnold Bregt are acknowledged for providing lists of potential respondents. For chapter 7 I would like to thank all the persons interviewed: Jan de Kruif, Peter Wynnia, Frans Plantinga, Titus Tiel Groenestege, Hero Dijkema, Eke Elzinga, Klaas Mulder, Theo de Boer, Dammie van der Poel, Bert van Wijlen, Albert Timmerman, Jasper Brouwers, Wilma Winkelhorst, Han Nap, Paul Rietjens and Erik de Graaf. Parameterisation of the model presented in chapter 7 would have been impossible without help from Hero Dijkema and Titus Tiel Groenstege who distributed a questionnaire among companies represented by CUMELA and impossible without data supplied by two large contractors.

In a number of cases it was decided after initial interviews not to continue work on a case-study. Two case-studies almost made it into the thesis and I would like to thank the following persons for their contribution. Firstly, on the calculation of variance in polygon area of agricultural parcels for which European subsidies are received, I would like to thank Theo van der Heijden, Auke de Bruin, Jandirk Bulens and Marc Hoogewerf. Secondly, on uncertainty in road length, thereby assessing both effects of positional errors and generalisation effects, I would like to thank Adri den Boer, Jacqueline van Rooij, Daan Snaathorst, Piet ten Haaf, Luc Heres and Jeroen Vrielink for their time.

I wish to thank my colleagues at the Centre for Geo-Information for four years of collegiality and for technical and administrative assistance provided. I would like to thank the C.T. de Wit graduate school for Production Ecology and Resource Conservation (PE&RC) for providing excellent courses and for facilitating in PhD discussion groups. For four years I have participated with great pleasure in the PE&RC discussion group on statistics, mathematics and modelling. The group was a success thanks to active participation of my fellow PhDs, so thanks to you all.

Finally, most theses end with thanking family, for patience, understanding and support. Dear Barbara and Ida, indeed this thesis could have been written without you so I will not thank you for that. Thank you for all other things that make life beautiful.

Pepijn van Oort
Wageningen, september 2005

Table of contents

1. Introduction

1. Introduction

1.1 Examples

I will start with two examples to illustrate what this thesis is about:

According to the cadastre's records, the area of a parcel is 10.044 hectares. The parcel is sold and the new owner measures the area by moving around the parcel with a global positioning system (GPS) device, according to which the area is 9.890 ha, 0.154 ha less than according to the cadastre. At a price of 35,700 euro per hectare (data from the Netherlands Central Bureau of Statistics, CBS), he claims from the vendor 5,498 euros. The vendor argues that the difference of 0.154 ha may well be caused by errors in the GPS coordinates and rejects the claim. Using the equations presented in chapter 5 of this thesis, the buyer argues that a difference as large as 0.154 ha cannot be explained from errors in GPS coordinates. The vendor then argues that the difference could further be explained by misinterpretation of the parcel boundary definition, or due to inherent fuzziness of the boundary. Did the buyer move along the inner edge or the centre of the ditch that surrounds the parcel?

As a consequence of being a party to the United Nations Framework Convention on Climate Change, a country has designed a national, spatially explicit, system for monitoring land use, land use change and changes in the forestry sector. Through an overlay of maps of two dates, changes are recorded. But are all recorded changes real changes? Probably not. Probably part of the changes shown are actually misclassifications at one or both dates. Probably part of the changes shown can also be attributed to changes in definitions of land use classes. Then, knowing the frequencies of misclassifications and knowing which changes in definitions have occurred, can these uncertainties be quantified and eliminated?

What do these examples tell us? Firstly, that spatial data quality is an issue in many decisions and analyses. This will be elaborated in section 1.2. Secondly, the examples show the importance of definitions. Section 1.3 elaborates on the issue of definitions and the topic is briefly touched upon in each of the chapters of the thesis. Thirdly, the examples identify a need of information on spatial data quality. Section 1.4 explains how descriptions of spatial data quality can be applied to quantify their implications for applications. Fourthly, the examples show that spatial data quality is not an issue for scientists only. Spatial data are used in decision-making and spatial data quality has implications for decision-making. In chapters four to six these implications are quantified based on spatial data quality descriptions, chapter seven is about how users cope with spatial data quality.

1.2 Why spatial data quality is an issue

For centuries cartographers, geographers, surveyors and geodesists have been involved in collection, storage, analysis and visualisation of spatial data. They have also been studying spatial data quality (for example see Haasbroek 1968, Balchut et al. 1979, Maling 1989 and Polman and Salzmann 1996). Since the 1980s, concerns about spatial data quality have increased, as a result of a two developments: (1) the emergence of Geographical Information Systems (GIS) in the 1960s and (2) from the 1970s onwards, a strong increase of available spatial data from satellites. With the large-scale adoption of GIS, the number of users from non-spatial disciplines has grown. Also with GIS, opportunities to use and combine data have grown tremendously. It is now much easier to use spatial data in all sorts of applications, even if inappropriate considering the quality of the data. Brimicombe (2003) wrote: "Historically, a detailed consideration of data quality issues in GIS lagged considerably behind the mainstream GIS development and application". Based on Aronoff (1989), Morrison (1995) and Longley et al. (1999) the five main reasons for current concerns about spatial data quality issues were identified as:

1. There is an increasing availability, exchange and use of spatial data;
2. There is a growing group of users less aware of spatial data quality;
3. GIS enable the use of spatial data in all sorts of applications, regardless of the appropriateness with regard to data quality;
4. Current GIS offer hardly any tools for handling spatial quality;
5. There is an increasing distance between those who use the spatial data (the end users) and those who are best informed about the quality of the spatial data (the producers).

Early warnings on the potential implications of spatial data quality were given by amongst others Chrisman (1984) who would later become chair of the group defining the spatial data quality part of the United States of America spatial data transfer standard (Chrisman 1987) and author of a chapter on spatial data quality (Chrisman 1991) in one of the most influential books on GIS. Brimicombe (2003) showed that the annual number of articles on spatial data quality has strongly grown since 1987, with less than five articles per year published before 1987. From the end of the 1990s onwards, there is a boom of international conferences on the topic: Symposia on Spatial Accuracy Assessment (1996, 1998, 2000, 2002, 2004) and Symposia on Spatial Data Quality (1999, 2003, 2004, 2005). Articles presented at these symposia have been published in Goodchild and Jeansoulin (1998), Lowell and Jaton (1999), Heuvelink and Lemmens (2000), Mowrer and Congalton (2000), Foody and Atkinson (2002), Heuvelink and Burrough (2002), Hunter and Lowell (2002) and Shi, Fisher and Goodchild (2002). Before this boom of conferences few books were published on spatial data quality, an exception being the book edited by Goodchild and Gopal (1989). Nowadays, no book about GIS or Geographical Information Science (GISc) goes without a chapter on spatial data quality. With the adoption of GIS in other disciplines, the attention for spatial data quality and its propagation in models has also grown, for example in ecology (Hunsaker et al. 2001) and in environmental modelling (Heuvelink 1998, Brimicombe 2003). Apart from these developments in the field of GIS, a branch of statistics called geostatistics emerged (Matheron 1971, Deutsch and Journel 1998, Goovaerts 1997, Isaaks and Srivastava 1990, Cressie 1991). The field is of great importance in spatial data quality research, because almost always variables are spatially correlated. Often, they are also correlated temporally. Geostatistics offers the theory to model both the implications of temporal and spatial correlation (see Pebesma et al. 2005 and Bogaert et al. 2005 and many, many more). GIS and geostatistics have different historical roots; the two became seriously engaged in the 1990s (Goodchild

2002) and they are now generally recognised as essential partners (Burrough 2001), increasingly also by GIS software producers.

1.3 From the real world to a spatial data set

An important part of spatial data quality research concerns the description of error and uncertainty in spatial data. Fundamental to the understanding of these concepts is an understanding of the process of how spatial data are derived from the real world. Aalders (2002) partitioned this process into two steps:

- Conceptualisation = the specification of what should be considered the real world and the abstraction of the selected objects;
- Measurement[1] = the specification of the measuring methods and the measurement requirements for capturing the data.

I will not discuss alternative terms, nor go into more detailed descriptions of these two steps and their sub-steps. The interested reader may consult Laurini and Thompson (1992), Burrough and Frank (1996), Burrough and McDonnell (1998), Molenaar (1998), Fisher (1999), Raper (1999), Frank (2001), Uitermark (2001), Kresse and Fadaie (2004) and Leyk (2005). Information on conceptualisation rules and measurement rules can be found under the heading of the terms lineage and ontology. Lineage is generally more focussed on measurement (including transformations after data acquisition), while ontologies are generally more focussed on conceptualisation. In this section, I will use the term ontology to refer to a set of conceptualisation and measurement rules. In other words ontology is here defined as the definition of the objects in a spatial data set.

Fisher (1999) distinguished three forms of uncertainty that arise during in the process of deriving a spatial data set from the real world: error, vagueness and ambiguity:

- Error is the difference between the value of a property of an object measured with unknown error (in the test data set) and the true value of the same property of the same object measured without error (in the reference data set). Error can be measured if a clear definition (ontology) is available;
- Vagueness arises due to poor definitions. Vagueness can be caused by poor documentation, or more fundamentally, if objects are fuzzy;
- Ambiguity arises due to disagreement on the definition of objects in a spatial data set. Such disagreement can arise because the definition was not specific, and it can arise due to fundamental differences in opinion.

According to Fisher (2000) many books on spatial data quality are limited to the treatment of error, thereby ignoring the two other forms of uncertainty. The same observation can be made from the overview of spatial data quality definitions in chapter 2 of this thesis. Clearly, the measurement of error requires agreement on a clear definition of what is perceived to be reality. This requirement cannot always be met, causing problems in the measurement and interpretation of error. Measurement of error is further complicated by the fact that no instrument exists with which error-free measurements can be made. Three solutions have been proposed. In recognition of the

[1] Aalders (2002) called it ‘mensuration’

fact that these solutions do not completely solve the problem the data in a reference data set are often referred to as 'accepted as true' rather than 'true' values. The solutions that have been proposed are:

- The solution to the interference of vagueness is to agree on a set of well defined rules and to apply these to both the test and the reference data. After doing this, error can be measured free from interference of vagueness and will be relevant to those who share the same ontology. This solution cannot always be applied. Firstly, the test data may already be collected according to poorly defined rules. Secondly, some concepts are inherently vague and it may be more rewarding to formalise their vagueness using fuzzy set theory than to try to eliminate it (Burrough and Frank 1996, Burrough et al. 1997, Fisher 1999, 2000);
- In case of ambiguity, there is disagreement on what constitutes the truth. Ambiguity is caused by the difference between the data set ontology (used in the collection of the test data) and the user ontology (of a certain user with a specific application of the test data). To satisfy the needs of users, producers may decide to collect reference data according to the user's ontology. For example Boschetti et al. (2004) dealt with the case where the data set ontology contained a rule stating that only one hard class is assigned to a pixel, while the user's ontology contained a rule stating that one or more hard classes may be assigned to a pixel. With error defined as the difference between test and true values, and disagreement on what constitutes the truth, one may question the appropriateness of the term "error" when reference data have a different ontology;
- The generally accepted solution to the problem that no instrument exists with which error-free measurements can be made is to collect reference data with instruments that a have a far greater accuracy than the instruments used in the collection of test data.

Consider the case where a test data set has been collected using a clear definition of what is perceived to be reality. Even then, vagueness can hamper the measurement of error. Ideally to eliminate vagueness, the same definitions are applied to both test and reference data. In practice, the definitions will often be slightly different. The highest consistency is obtained when reference data are collected through field measurements and when these measurements occur at the same time as the measurement of the test data. An often chosen alternative is to select a candidate reference data set with an ontology that does not differ too much from the test data set's ontology. This candidate reference data is then transformed to minimise the ontological difference between test and reference data. Sometimes this alternative is chosen because it is cheaper or faster, on other occasions it may be the only option. It is the only option when collection of reference data starts at a point in time where the real world phenomena represented in the test data have changed since the test data set's acquisition date. For example, there is increasing attention for digitisation of historical data (Petit and Lambin 2002, Kramer and Knol 2004, Leyk 2005) also in the context of climate change (Vitousek 1994, Houghton et al. 1999). As a result of this practice, test and reference data definitions will be similar but not exactly the same. This will result in vagueness and will interfere with measurement and interpretation of error descriptions.

In conclusion, the reader should realise after reading this section that measurement of error is not at all that simple and straightforward as one might expect. To interpret correctly the errors reported in spatial data quality reports, it is important to be aware of the possible interference of vagueness and ambiguity, which may be caused both by test and reference data. It is of importance to be informed about how both test and reference data were collected. This information should be provided under the heading of the spatial data quality element called lineage (see next chapter). Finally, is important to choose appropriate models for describing the uncertainty and its propagation, appropriate considering the different forms of uncertainty discussed in this section.

1.4 Fitness for use assessment

Sections 1.1 and 1.2 showed why concerns about spatial data quality exist. Section 1.3 briefly explained the uncertainties that arise during the process of deriving spatial data from the real world. There is an increasing call for formal methods to describe these uncertainties and to apply these descriptions in fitness-for-use assessment. It is generally accepted that spatial data quality descriptions serve to allow the user to evaluate the fitness of the data for his particular application (Moellering 1988, Morrison 1995).

The process of assessing fitness-for-use can be partitioned into three steps:

1. To search for a spatial data set that contains the information needed for the intended application (Brassel et al. 1995 called this the assessment of model completeness);
2. To explore whether there are legal or financial constraints to access or particular use of the spatial data (Aronoff 1989 called this the usage component);
3. Finding out if, given the spatial data quality, risks are acceptable (see Agumya and Hunter 2002).

All necessary information for going through these 3 steps is found in the meta-data report (FGDC 1998a, ISO 2003b), but not necessarily in the spatial data quality section of the meta-data report. The first two steps are extensively addressed in research on meta-data, spatial data infrastructures and interoperability. This thesis deals exclusively with the third step. Different approaches exist to finding out if risks are acceptable, given the spatial data quality. Extensive discussions are found in Agumya and Hunter (1999a, 1999b, 2002), Openshaw (1989) and chapter 7 of this thesis. They are often based on theoretical considerations from risk analysis (e.g. Fischhoff et al. 1981, Morgan et al. 1990, Jaeger et al. 2001). The most rigorous approach, advocated by Agumya and Hunter, is to apply error propagation analysis to quantify how errors in input data sets propagate into the outcomes of the analysis or decision-making process. The approach is gaining popularity in research (e.g. see Stein et al. 1995, Broos et al. 1999, Li et al. 2000, Crosetto and Tarantola 2001, de Bruin et al. 2001, de Bruin and Hunter 2003, chapters four to six of this thesis). The approach is often called risk analysis, although in many cases quantitative risk analysis would be a more appropriate term. More qualitative forms of risk analysis make an inventory of which events may occur and their consequences, without quantification of the probabilities and consequences of events.

Fischhof et al. (1981) provided a typology and discussion on different approaches to assessing risks and their acceptability. Within research on risks caused by limited spatial data quality, frequently approaches other than quantitative risk analysis are adopted to assess the fitness for use of spatial data. Agumya and Hunter (1999a), Openshaw (1989) and Longley et al. (1999) suggest that often the third step of the fitness-for-use assessment is omitted. For step 3, approaches less rigorous than quantitative risk-analysis are often applied. These less rigorous alternatives include:

- Use a spatial data set if acceptable results were obtained with it in the past;
- Rely on feedback mechanisms that are expected to reduce the risk;
- Consult experts;
- Consult information on usage and purpose;
- Apply trial and error as a means of finding out the amount and acceptability of risk;
- Insurance and other forms of risk absorption.

These approaches have in common that in the short run they are faster and cheaper than quantitative risk analysis. They are less rigorous because risks are not quantified. As a result, it is possible that unknowingly, the spatial data input to a decision or analysis produce larger risks than acceptable.

1.5 Knowledge gaps and scope

This thesis is about the description of spatial data quality and the application of these descriptions in fitness-for-use assessment. The previous sections showed that already a large body of literature exists. The following two subsections identify knowledge gaps and scope with regard to the description and application of spatial data quality as dealt with in this thesis.

1.5.1 Description of spatial data quality

From an early stage, researchers like Chrisman and Aalders have been involved in committees that developed standards for the description of spatial data quality. Spatial data quality standards were accepted in the USA in 1992 and by the International Standardisation Organisation ISO in 2002; final meta-data standards were accepted in the USA in 1998 and by ISO in 2003. With these standards accepted and used by the GI user community (Crompvoets et al. 2004), there seems not much left of the knowledge gap with regard to the definition of spatial data quality. However, research on the definition of spatial data quality is not yet complete. Three knowledge gaps can be identified:

- After the acceptance of standards comes the need to adopt and implement them. Research into the problems that occur and the questions that arise when implementing the standards is outside the scope of this thesis;
- Spatial data quality is defined by its elements. In literature somewhere between five and eleven elements are distinguished (Aalders 2002, Devillers et al. 2005). Chapter 2 presents an overview. Based on the overview and the chapters thereafter, recommendations are given in chapter 8, section 8.2;
- There is increasing attention for modelling the spatial variability in spatial data quality, especially the spatial variability in the accuracy of attributes with a nominal measurement scale (Bierkens and Burrough 1993, Steele et al. 1998, McGwire and

Fisher 2001, Bogaert 2002, Smith et al. 2003, de Bruin et al. 2004, Foody 2005). Chapter 3 contributes to the research on this topic.

1.5.2 Application of spatial data quality

According to Longley et al. (1999), "Despite what appear obvious arguments in favour of explicit treatment of data quality in GIS, and despite substantial research into appropriate methods, much GIS practice continues to proceed as if data were perfect". As argued by Veregin (1999) and Fisher (1997) in a review of Guptill and Morrison (1995), much research has focussed on defining and reporting spatial data quality, while fitness-for-use assessment has received relatively little attention. Error propagation analysis is the most scientific way to assess fitness for use. Pleas for incorporating error propagation analysis tools in GIS appeared in the 1990s (Burrough 1992) and have since then only gained in strength (Burrough 2001). Many tools have been developed within the GISc research community (Heuvelink 1993, Forier and Canters 1996, van der Wel 2000, Crosetto and Tarantola 2001 and Duckham 2002). However, such tools have to date still only very scarcely been implemented in commercial GIS software. Despite the availability of tools, methods and knowledge in the GISc community, researchers like Openshaw (1989), Agumya and Hunter (1999a) and Longley et al. (1999) noted that these are scarcely utilised by users outside the GISc community. One way to increase their use could be, as advocated by Agumya and Hunter (1999a, 1999b, 2002), to make the results of error propagation analysis more tangible by continuing the analysis not to the stage of the end product of a spatial analysis, but further to the stage of expressing the consequences of errors in spatial data in terms of risks. The number of publications following this approach is growing (a.o. Stein et al. 1995, Broos et al. 1999, Li et al. 2000, Crosetto and Tarantola 2001, de Bruin et al. 2001, de Bruin and Hunter 2003, chapters four to six of this thesis). Another way to increase the use of tools, methods and knowledge about error propagation analysis would be to increase the usability of the tools. However, to my knowledge no scientific publications exist in which the use of these tools (or the lack of it) has been properly evaluated. With regard to the knowledge gap in the application of spatial data quality descriptions, the contributions of this thesis are:

- Chapter 4 shows how knowledge on classification errors can be used to improve estimates of land cover change. It also shows that classification errors between different dates are often correlated and that this correlation needs to be taken into account when improving change estimates;
- Chapter 5 presents theory and an illustration of the calculation of the financial uncertainty in the value of a set of polygons, where uncertainty exists on the true position of the polygon boundaries;
- Chapter 6 discusses the elements of spatial data quality that are relevant in digging activities around underground cables and pipelines, and proposes a model to quantify the implications of over- and incompleteness. As in chapters 4 and 5, the theory is illustrated in a case-study;
- Chapter 7 is a study of why users of spatial data are in many cases less willing to spend resources on the quantification of implications of spatial data quality. Contrary to existing publications on the same topic, this chapter focuses on the decision-theoretic considerations of users.

1.6 Aim, research questions and outline

This chapter started with two examples of spatial analysis and decision-making in which spatial data quality is an issue. GIS practitioners are continuously confronted with such questions. In the Netherlands, a number of important practitioners recognised the need for research in addressing issues of spatial data quality. The work presented in this thesis was inspired and funded by their organisations. The aims and research questions below were formulated on the basis of literature (in which knowledge gaps were identified) and on the basis of feedback from funding organisations. Feedback was received from representatives during joint biannual progress meetings, and in between during bilateral meetings. The names of the organisations and their representatives are listed in the acknowledgements.

The aim of this thesis is twofold: (1) to enhance the description of spatial data quality and (2) to improve our understanding of the implications of spatial data quality. Six research questions were formulated, which are answered consecutively in chapters 2 to 7. The questions and chapters follow the order already indicated in the thesis title: their contents move from description to application. Chapters 2 and 3 contribute mostly to the first aim (description), chapters 4 to 7 mostly to the second (application). Chapter 8 presents conclusions and recommendations for further research. The six research questions are:

1. What is spatial data quality? (chapter 2);
2. Does land cover classification accuracy depend on landscape heterogeneity? (chapter 3);
3. How can knowledge of classification errors be used to improve land cover change estimates? (chapter 4);
4. How do positional errors in vertex coordinates propagate into errors in area estimates for polygons sharing a boundary? (chapter 5);
5. How can implications of spatial data quality be calculated in the case of digging activities around underground cables and pipelines? (chapter 6);
6. Which decision-theoretic factors determine the willingness of researchers and policy makers to spend resources on the quantification of implications of spatial data quality (chapter 7);

In each chapter, new theory is presented and then applied in case-studies. In the case-studies I analysed the national land cover database LGN, the national topographic database Top10vector, data from Veregin (1989), data on subsidies for management of natural areas of the Forest Service of the Netherlands and data from interviews and questionnaires.

2. Spatial data quality: terms and definitions

2. Spatial data quality: terms and definitions

This chapter gives an overview of definitions of spatial data quality. First, some basic terms related to spatial data quality are discussed. Next, contents of five definitions are compared and discussed. The chapter concludes with a summary. For convenience terms in this chapter are printed in bold when first introduced, to enable the reader to quickly look up these terms.

2.1 Some basic terms related to spatial data quality

Previous chapter (§1.3) showed that the measurement of error requires agreement on a clear definition of what is perceived as reality. Presence of vagueness and/or ambiguity can complicate the measurement of error and hence its interpretation. **Error** was defined as the difference between the value of a property of an object measured with unknown error and the true value of the same property of the same object measured without error. Below, terms related to error are discussed.

The total error consists of **systematic error** and **random error**, the systematic error is often called **bias**. Random error is the distance between a measurement and the mean of measurements, total error is the distance between a measurement and the true value and bias is the distance between the mean of measurements and the true value. Accuracy and precision are summary measures of error. **Accuracy** is calculated from total errors, **precision** from random errors. If the systematic error (bias) equals zero, then precision equals accuracy. For this reason, the two are often used interchangeably. This can be confusing and misleading and the best thing to avoid confusion is to report explicitly the bias, even if it equals zero. The most common way to summarise errors is with the Mean Square Error (MSE) or the Root Mean Square Error (RMSE). Other summary measures are for example the Mean Error (ME), Mean Absolute Error (MAE) and Maximum Error (ME). The error matrix, used in reporting classification accuracy, measures the frequency and type of correct and misclassifications. Summary measures of the error matrix are the users', producers' and overall accuracy. The set of data of which the accuracy or precision is to be calculated is called the **test data set**, and the set of data containing the true values is called the **reference data set**. These two terms will be applied throughout the thesis.

Two types of precision can be distinguished: **statistical precision** and **storage precision**. The two terms have also been called geographical precision and digital precision (Vauglin 2002). Statistical precision is a summary measure of random errors, corresponding with the definition of precision given above. Storage precision refers to the detail (number of digits) with which data are stored in a database. Statistical precision is affected by storage precision, by the precision of the measurements and by all transformations applied thereafter (Burrough and McDonnell 1998). A high storage precision does not guarantee a high statistical precision and a high statistical precision does not guarantee a high accuracy. In this thesis, the term precision refers always to statistical precision. If storage precision is meant then this is explicitly reported so.

The terms **resolution** and **precision** are often used interchangeably (e.g. Worboys 1998, Veregin 1999), because both refer to the level of detail. There is a subtle difference between the two. Resolution is the detail at which data are presented; precision is the amount of detail that can be discerned. It is not justified to present data at a resolution higher than determined by the precision. However, there may be sound reasons to present data at a courser resolution than justified by their precision (Veregin 1999). Resolution is not limited to the spatial domain (pixel size, smallest polygon, smallest distance between vertices); it can also be given for the thematic and temporal domain (Veregin 1999). Related to the term resolution is the term **scale**. Historically, large scale is associated with high resolution and high precision (Maling 1989, Aronoff 1989). In a digital era, this implicit relationship no longer holds (Aronoff 1989). In a GIS it is quite easy to increase the resolution of a data set, thereby suggesting a high precision and large scale. Similarly it is quite easy to increase the digital precision, thereby suggesting a high statistical precision. With GIS, data can be visualised at any scale desired. Consequentially, the use of the terms scale and resolution without further explanation may cause confusion.

2.2 Elements of spatial data quality

Quality is defined as "totality of characteristics of a product that bear on its ability to satisfy stated and implied needs" (ISO 2002, originally in ISO standard 8402). Without a more detailed definition of these characteristics, the definition remains meaningless. The characteristics are traditionally called elements, so we speak of elements of spatial data quality. Definitions of elements of spatial data quality were derived from the following five sources, in historical order of appearance (see also Aalders 2002 and Devillers et al. 2005):

- **Aronoff (1989)** presented an interpretation of the draft USA-SDTS (Chrisman 1987, Moellering 1988) from a management perspective;
- **USA-SDTS (1992)**. The United States of America spatial data transfer standard contains a section on spatial data quality elements. The spatial data transfer standard was accepted in 1992 (Department of Commerce 1992) and was later incorporated in the USA metadata standard (FGDC 1998a,b);
- **ICA (1995)**. On behalf of the International Cartographic Association, Guptill and Morrison (1995) published the book entitled "Elements of Spatial Quality". The book contains contributions by a range of authors;
- **CEN/TC287 (1998)**. Technical committee 287 of the Comité Européen de Normalisation (CEN) developed the European pre-standard ENV 12656. During the process, ISO started standardisation and CEN/TC287 dissolved into ISO/TC211 (Kresse and Fadaie 2004: page 6). This meant that development of ENV 12656 stopped and that after acceptance of standards as international standards by ISO/TC211, CEN/TC287 started considering approval of these standards as European standards. The CEN/TC287 standard referred to in this thesis is the European pre-standard, CEN/TC287 (1998);
- **ISO/TC211 (2002)**. Technical committee 211 of the International Standardisation Organisation (ISO) has developed a number of international standards for geographic information: 19113 (Quality principles), 19114 (Quality evaluation procedures). These elements are described as part of the 19115 (Metadata). The standard 19138, which is still under development at the time of writing this thesis, will define a set of measures for the spatial data quality elements identified in ISO 19113. See ISO (2002, 2003a,b) and Kresse and Fadaie (2004).

The elements distinguished in these five sources are listed in table 1. Table 1 shows that all elements are part of the metadata, yet not all elements are part of the spatial data quality section of the metadata. For this reason, some researchers (Devillers et al. 2005) have criticised spatial data quality standards for providing the user with insufficient information to allow for assessing fitness for use. Also, the explicitness with which elements are recognised in the sources differs. For example, the term resolution is used repeatedly in the "Entity and Attribute Information" section in the USA metadata standard (FGDC 1998a), but is not explicitly distinguished in the spatial data quality section. However, in its data quality section, it states in its description of the element completeness: "geometric thresholds such as minimum area or minimum width must be reported". This could well be interpreted as resolution. The elements in table 1 are discussed in this section in the order in which they are numbered in the table. For details and exact contents the reader is referred to the references cited above.

Table 1. Elements of spatial data quality in five sources

element*	Aronoff (1989)	USA-SDTS (1992)	ICA (1995)	CEN TC287 (1998)	ISO TC211 (2002)
1. Lineage	S	S	S	S	S
2. Positional accuracy	S	S	S	S	S
3. Attribute accuracy	S	S	S	I	S
4. Logical consistency	S	S	S	S	S
5. Completeness	S	S	S	S	S
6. Semantic accuracy			S	S	
7. Usage, purpose, constraints	S	M		S	M
8. Temporal quality	S	M	S	S	S
9. Variation in quality		I	I	S	I
10. Meta-quality		I	I	S	I
11. Resolution	S	I	I	I	M

* the names of the elements are my choice. The five sources in some cases use different names for the same elements, for example thematic instead of attribute accuracy, for example temporal information, temporal accuracy or time instead of temporal quality

S = explicitly recognised as an element, in the spatial data quality section of the metadata. In the USA-SDTS, the CEN/TC287 and the ISO/TC211, the spatial data quality section is part of the larger metadata report. In Aronoff and ICA the distinction between different sections of the metadata report is not made.

M = explicitly recognised as an element, in another section of the metadata

I = implicitly recognised as an element.

1. Lineage

Lineage is in short the history of a geographic data set. A description of the source material from which the data were derived, and the methods of derivation, including all transformations involved in the production process.

2. Positional accuracy

Positional accuracy is the accuracy of coordinate values. A distinction is often made between **relative** and **absolute** positional accuracy and between **vertical** and **horizontal** positional accuracy. Vertical positional accuracy however has also been treated as attribute accuracy (attribute depth or height), for example see the contribution by Goodchild in ICA (1995). Relative positional accuracy is the accuracy relative to other data in the same test data set. Absolute positional accuracy is the accuracy of test coordinate values relative to matching reference coordinate values on the same coordinate reference system. As Chrisman (1991) indicates, relative positional accuracy is sufficient for calculating via error propagation analysis the variance in area and variance in length of lines. More generally speaking, it is sufficient for error propagation analysis on a single spatial data set. If data sets are to be combined then absolute positional accuracy needs to be known.

3. Attribute accuracy

Attribute accuracy is the accuracy of all attributes other than the positional and temporal attributes of a spatial data set. Attributes can be measured on four measurement scales: ratio, interval, ordinal and nominal. The nominal scale is an unordered scale, for example land cover. The accuracy of nominal attributes is normally described in an error matrix, also called misclassification matrix or confusion matrix. The accuracy of ratio attributes is described with the same measures as those used for positional attributes, for example root mean square error (RMSE). In the CEN/TC287 (1998, p. 22), the name semantic accuracy was used for what was in fact attribute accuracy. It is for this reason that table 1 states that attribute accuracy is only implicitly recognised in the CEN/TC287 data quality standard.

4. Logical consistency

Logical consistency is defined in the USA-SDTS as: “the fidelity of relationships encoded in the data structure”. Similar definitions are found in the other references in table 1. Kainz in Guptill and Morrison (1995) elaborates with great rigour the definition of topological consistency, with much less emphasis on other aspects of logical consistency:

- valid values, graphic data, topological, date (USA-SDTS);
- geometric-, semantic- and topological consistency (CEN/TC287);
- conceptual-, domain-, format- and topological consistency (ISO/TC211).

5. Completeness

Completeness is a measure of the absence of data and the presence of excess data. Errors resulting in overcompleteness are called errors of commission, errors resulting in incompleteness are called errors of omission. Crucial to the detection of these errors is to know what does and what does not belong to the complete set that the producer intended to include in his data set. For this reason, the USA-SDTS states: “The quality report must include information about selection criteria, definitions used and other relevant mapping rules. For example geometric thresholds such as minimum area or minimum width must be reported … The report on completeness shall describe the relationship between the object represented and the abstract universe[1] of all such objects. In particular, the report shall describe the exhaustiveness of a set of features. Exhaustiveness concerns spatial and taxonomic (attribute) properties, both of which can be tested"

I will give an example to illustrate. An abstract universe states that a data set contains hospitals and defines a hospital as a building in which one or more doctors are employed and where the average annual number of patients is at least 100. A building with an annual average of 90 patients does not belong to this abstract universe. If this building is present in the data set then that is an error of commission (resulting in overcompleteness). A building with average 110 patients and with doctors employed does belong to the abstract universe. If this building is not present in the data set then the data set is incomplete due to an error of omission.

The definition of completeness by Brassel et al. (1995) from ICA is broader than in the three standards. Brassel et al. distinguish two kinds of completeness: data completeness and model completeness. Data completeness corresponds with the definition given in the standards. Model completeness is a measure of how well the contents of a spatial data set correspond with the information needed for the intended application. In this thesis in section 1.4, assessment of model completeness was identified as the first step of fitness-for-use assessment. A producer cannot provide information on model completeness, because that would require knowledge of the exact information needs of all possible applications. The best a producer can do is provide detailed descriptions of the information content of his product. Often under the heading of "usage, purpose, constraints", the producer documents applications for which other users considered the information content of the spatial data set (model-) complete.

6. Semantic accuracy

As can be seen in table 1, the element semantic accuracy occurred only in the ICA publication and in the European pre-standard (the CEN/TC287). Now that the development of the ISO standards is completed, CEN will adopt the ISO standard so that the element will most likely disappear altogether. An elaborate discussion on the element semantic accuracy is found in the ICA publication, by François Salgé. In his broad definition of the element Salgé stresses the interconnectedness of all data quality elements and he stresses the importance of describing sources of uncertainty other than error (see also §1.3 of this thesis).

7. Usage, purpose, constraints

In section 1.4 I identified 3 steps of fitness-for-use assessment. The element "Usage, purpose, constraints" assists the potential user of a data set in steps 1 and 2 of the assessment. Aronoff (1989) explicitly recognised the importance of this element and provides more details on its contents. ISO/TC211 makes a distinction between usage, purpose and constraints. Its rationale for doing so is that intended use (purpose) is not necessarily the same as actual use (usage). A difference between on the one hand the constraints in Aronoff (1989) and on the other hand the constraints in the USA-SDTS, the CEN/TC287 and the ISO/TC211 is that Aronoff also considers the costs a

[1] the definition of Abstract Universe in the USA-SDTS corresponds closely with the definition of ontology given in §1.3 and with the definition of the nominal ground. See also Aalders (2002)

constraint. The direct cost is the price paid for a data set. As indirect costs, Aronoff considers all the time and material used to make the data ready for use for the buyer. The constraints in the USA-SDTS, the CEN/TC287 and the ISO/TC211 represent only legal or contractual constraints to the access and application of data, while direct costs are listed in yet another section of the metadata report.

8. Temporal quality

CEN/TC287 and ISO/TC211 contain an element called temporal accuracy, in ICA it is called temporal information, Aronoff calls it time and the USA-SDTS (Department of Commerce 1992) contains no such element. I prefer the use of the term temporal quality rather than the term temporal accuracy, because not all sub-elements listed below summarise error (see the definition of accuracy in §2.1). The following sub-elements have been distinguished in CEN/TC287 and in ISO/TC211:

- accuracy of time measurements = summary of errors in time measurements (CEN and ISO);
- temporal validity = the validity in respect of time (CEN and ISO), also sometimes called currency. According to CEN the temporal validity can take on one of the following three values: “out_of_date”, “valid” or “not_yet_valid”;
- temporal consistency = correctness of the order of events (ISO);
- last update (CEN);
- rate of change = an estimate of the rate of change in the phenomenon represented in the data. Together with information on the last update this element can inform the user about the currency (CEN);
- temporal lapse = the average time between change on the nominal ground and its representation in data (CEN).

9. Variation in quality

CEN/TC287 defined the element homogeneity as “a textual and qualitative description of the expected or tested uniformity of quality parameters in a geographic data set”. In this thesis the element is named variation in quality since the element is only relevant if quality varies within the data set. According to table 1, only the CEN/TC287 distinguishes this element as a separate element. Closer reading of the USA-SDTS, ISO/TC211 and ICA learns that these treat variability as a part of the other elements (for that reason they are shown as I’s in table 1). The USA-SDTS states: “Where the spatial variation in quality is known, a quality report must record that variation” and repeats this statement under the elements positional accuracy and attribute accuracy. ISO (2002: 8) incorporates the description of variability using the term **data quality scope**: “At least one data quality scope shall be identified for each applicable data quality sub-element. A data quality scope may be a data set series to which a data set belongs, the data set or a smaller grouping of data located physically within the data set sharing common characteristics. If a data quality scope cannot be identified, the data quality scope shall be the data set”. Annexes in ISO (2002) further show how this variation in quality can be reported: by reporting the data quality scope to which the data quality measure is applicable.

10. Meta-quality

This element provides information on the quality of the quality description. For example if the positional accuracy is estimated from a smaller sample size, then that estimate is of lower quality. The three standards USA-SDTS, CEN/TC287 and ISO/TC211 treat meta-quality as a part of the other elements and require it to be documented if possible. As is in element 9, table 1 shows that the element meta-quality is explicitly recognised in CEN/TC287. The USA-SDTS does not use the term meta-quality, but does describe different tests to assess accuracy and ranks these according to what we may here call meta-quality. It states: "informed assessment of fitness for use is best served by the most rigorous types of tests".

11. Resolution

Being aware of the confusion that can arise about the interpretation of the terms resolution and scale (§2.1), it is also recognised that they can be relevant to the user of a spatial data set (Veregin 1999). Often a decision or analysis requires data at a certain resolution and as such, information on the resolution is important in the first step of fitness-for-use assessment (§1.1.4). As for "variation in quality" and "meta-quality", the element "resolution" is mostly encountered as a sub-element of other spatial data quality elements or other metadata elements.

2.3 Summary

The most important motivation for describing spatial data quality is to provide the potential user of a data set with the necessary information to decide on the fitness for use of a data set for his particular application (§1.4). This chapter presented an overview of definitions of spatial data quality. Five sources of definitions were compared. As can be seen in table 1, there is strong agreement on the contents of the definitions. The treatment of elements of spatial data quality in the five sources differs in two aspects: (1) the location within the meta-data report and (2) the explicitness with which elements are named as individual elements. Leaving these differences aside there appears to be little disagreement on which elements together define spatial data quality. Of the eleven elements listed above, the last three ("variation in quality", "meta-quality" and "resolution") are often encountered not as individual elements but as sub-elements of other individual elements. The element "semantic accuracy" is likely to dissolve in other elements in which case it is no longer recognised as an individual element.

3. Spatial variability in classification accuracy[1]

[1] Based on: P.A.J. van Oort, Bregt, A.K., de Bruin, S., de Wit, A.J.W. and Stein, A., 2004. Spatial variability in classification accuracy of agricultural crops in the Dutch national land-cover database. International Journal of Geographical Information Science 18(6): 611–626. © Taylor & Francis Ltd.

3. Spatial variability in classification accuracy

Abstract

Variability in per cell classification accuracy is predominantly modelled with land cover class as the explanatory variable, i.e. with users' accuracies from the error matrix. We developed logistic regression models to include other explanatory variables: heterogeneity in the 3×3 window around a cell, the size of the patch and the complexity of the landscape in which a cell is located. We found that per cell the probability of correct classification was significantly (α = 0.05) higher for cells with a less heterogeneous neighbourhood, for cells that are part of larger patches and for cells located in regions with a less heterogeneous landscape. To validate the models, a leave-one-out procedure was applied in which the absolute difference between the actual and the model-estimated number of cells correctly classified was summarised over 55 regions in the Netherlands. The sum of differences reduced from 60.9 to 48.1 after adding the variables 'patch size' and 'landscape dominance' to the land cover class model. Spatial variability thus modelled therefore lead to a substantial improvement in per cell classification accuracy estimation.

3.1 Introduction

Land cover data derived from classified satellite images are increasingly used in land use planning and environmental management. As a consequence, concern about the accuracy of these data has grown. Commonly, the classification accuracy is reported by the percentage correctly classified (PCC) and the error matrix (Congalton 1991, Janssen and van der Wel 1994). Information on the spatial variability of these measures is rarely provided (Foody et al. 1992, Goodchild 1995, Smith et al. 2002). Lack of quantitative information on this spatial variability can be a serious problem: usage of a PCC not representative for a region may lead to misleading outcomes in an error propagation analysis and to incorrect assessment of the fitness for use of the data for a specific region. Therefore, McGwire and Fisher (2001) recommend reporting the accuracy for the smallest region size where the producer expects the data set to be used.

Responding to the demands from users, de Wit (2002) has recently published estimates of the accuracy of the Dutch national land cover database (LGN) on a region (62.5 km^2) and province basis. These estimates were computed from reference data located within each of the regions concerned. The disadvantage of such an approach is that proper accuracy assessment may be impossible if there are no or too few reference points within a given region. Model based approaches possibly do not suffer from this disadvantage.

For example, some researchers (Steele et al. 1998) used a spatial interpolator (kriging) for estimating probabilities of correct classification from sample points with known probability of misclassification and models of spatial continuity (variograms). Their method requires that the interpolated points are within the range of influence of sampled points. If this requirement is not met, models calibrated on exhaustively sampled explanatory variables can be considered. The most commonly applied model is based on users' accuracies derived from the error matrix, in which land cover class is the explanatory variable. Smith et al. (2002, 2003) developed logistic regression models to assess the impact of other variables on per cell classification accuracy. They found significant ($\alpha = 0.05$) impacts of both patch size and focal heterogeneity (heterogeneity in the 3×3 window around a cell).

On the basis of visual analysis, de Wit et al. (1999) anticipated that regional differences in accuracy were also associated with complexity of the landscape and that differences in landscape complexity were not entirely captured by the variables land cover class, patch size and focal heterogeneity. To our knowledge, such observation has never been substantiated by quantitative evidence.

In this chapter, our aim is to extend the work of Smith et al. (2002, 2003) by including landscape indices (Forman and Godron 1986, O'Neill et al. 1988, Li and Reynolds 1993, Riiters et al. 1995) as potential explanatory variables of per cell classification accuracy. These indices aggregate the distribution of land cover classes and of focal heterogeneity values within a region to one number representative for a whole region. In the following sections we describe an experiment with 55 regions taken from the Dutch national land cover database. First the data sets, variables and models are introduced and the procedure to validate the models is outlined. Next the results are presented and we end with a discussion of the results.

3.2 Methods

3.2.1 Data

In this study two data sets were used: (1) the national land cover database (LGN) derived from satellite images and (2) a reference data set derived from the agricultural census (REF). LGN has a resolution of 25 m and covers the whole Netherlands. It is produced every 4 years, the most recent update of the database is based on images of 1999 and 2000 and was completed in 2001. The production process involves integration of multi-temporal satellite imagery (from Landsat TM and SPOT), ancillary data and expert knowledge (Thunnissen and Noordman 1996, Thunnissen and de Wit 2000). The classification system comprises 39 land cover classes, 7 of which are agricultural crops. In total, agricultural crops cover 52.9% of the country. The database is widely used by national and regional government agencies for water management, hydrological modelling, land use planning and environmental management (van Soest et al. 2001, de Wit 2002).

The reference data set used in this study was derived from the agricultural census. For the census the government annually sends to all farmers one or more 1:10000 aerial photos, with parcel boundaries derived from the national topographical map and the cadastral map printed on these photos. Farmers return these maps, indicating the areas of crops produced within their parcels. Crop name(s) and area(s) are recorded without reference to their position within the parcel. The maps returned by all individual farmers are combined per region. Figure 1 shows 55 of these regions, their location corresponds with map sheets supplied by the national topographical database which is often used in combination with LGN. The reference data set was derived from this census data set in four steps:

1. Parcels with more than one crop and parcels with a reported crop area exceeding the geometrical area in the census data set by more than 10% were excluded. As a result, the coverage of census data per region ranged between 1% and 42%;

2. The census data set, originally in vector format, was converted to raster to enable overlay with LGN. As in LGN, only one class was assigned to each cell in the reference data set using the majority rule. This rule assigns to a cell the land-cover class with the largest relative area within the cell;
3. We randomly selected 55 of the regions of the census data set (figure 1). Of the selected regions, 47 were sized 6.25 × 10.0 km, the other 8 regions were slightly smaller. The resulting data set contained 955 783 cells with reference land cover;
4. Reference data were sub-sampled to obtain a more or less realistic data density. For that purpose, the data set was overlaid with a point grid with spacing 30 cells (750 m) and a randomly drawn origin was made in each region (de Gruijter 1999). All points coinciding with census data cells were added to the reference data set. The resulting reference data set (REF) contained 1161 cells.

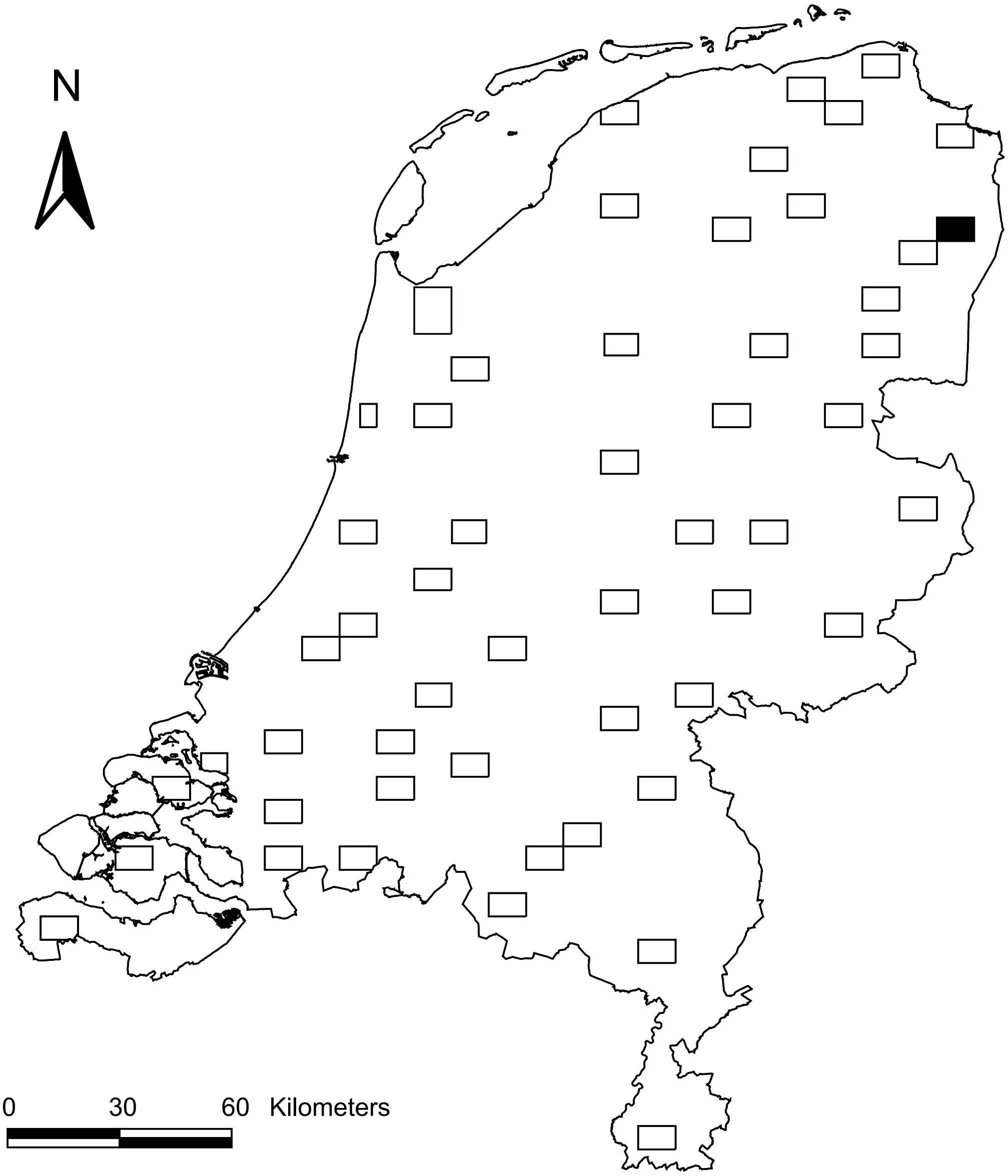


Figure 1. The Netherlands with regions (rectangles) used in this study. Models to estimate the PCC of a validation region (black) were fitted on data of the other 54 regions (white). This procedure was repeated for all regions.

3.2.2 Explanatory variables

We used 4 categories of explanatory variables (see table 1). Category 1, contains a single variable CLASS, which specifies the land cover class of the cell.

Category 2 variables quantify the heterogeneity of the focal (= 3×3) neighbourhood around each cell. Focal heterogeneity, HET, equals the number of different land cover classes in the neighbourhood. Focal homogeneity, HOM, equals the number of neighbouring cells with the same land cover class as the central cell.

The single category 3 variable L10P quantifies the size of the patch in which a cell is located. A patch was defined as a set contiguous cells of the same land cover class. Contiguity was defined as sharing a common boundary, thus each cell had 4 contiguous cells. Patch shape indices express in one number the distribution of patch sizes occurring in a region (Forman and Godron 1986: 189, Baker and Cai 1992). They were not considered here because a preliminary data analysis showed little variation in the values of these indices for the 55 regions assessed in this study (O'Neill et al. 1988 obtained similar results for agricultural areas).

Category 4 includes four landscape indices, which differ in sensitivity to the number of land cover classes and in ability to distinguish between different landscape textures. The indices are known as landscape heterogeneity, landscape dominance, landscape entropy and landscape contagion (O'Neill et al. 1988, Li and Reynolds 1993, Riiters et al. 1995). We will refer to the indices as landscape variables.

Landscape heterogeneity and landscape dominance are derived from marginal probabilities $p(i)$, $i = 1,...,I$, with I the number of land cover classes in the region. Let $A(i)$ be the area of land cover class i in a region. Then $p(i)$, landscape heterogeneity (HTG) and landscape dominance (DMG) are:

$$p(i) = \frac{A(i)}{\sum_{i=1}^{I} A(i)} \tag{1}$$

$$\text{HTG} = -\sum_{i=1}^{I} p(i) \cdot \ln(p(i)) \tag{2}$$

$$\text{DMG} = \ln(I) - \text{HTG} \tag{3}$$

Landscape Entropy (ENT) and Landscape Contagion (CON) are obtained from the probabilities $p_{adj}(i,j)$, that a randomly chosen cell is classified as i and that at least one adjacent cell (in a 3×3 window) is classified as j. Let $N_{adj}(i,j)$ be the total number of adjacencies between cells with class i and class j within a region. Then:

$$p_{adj}(i,j) = \frac{N_{adj}(i,j)}{\sum_{i=1}^{I}\sum_{j=1}^{I} N_{adj}(i,j)} \tag{4}$$

$$\text{ENT} = -\sum_{i=1}^{I}\sum_{j=1}^{I} p_{adj}(i,j) \cdot \ln(p_{adj}(i,j)) \tag{5}$$

$$\text{CON} = 2 \cdot \ln(I) - \text{ENT} \tag{6}$$

Table 2 shows the range of values of the landscape variables. In distinguishing between relatively heterogeneous and finely textured landscapes the variables HTG and ENT are more sensitive to I than the variables DMG and CON.

Table 1. Explanatory variables

Category	Variable	Key aspects of variable
1. Land cover	Land cover class (CLASS)	6 binary variables are used to indicate the presence of one of the 7 land cover classes*
2. Focal	Heterogeneity (HET)	number of different classes in direct (8-cell) neighbourhood of cell
	Homogeneity (HOM)	number of cells with land cover class in direct (8-cell) neighbourhood the same as of centre cell
3. Patch	Patch size (L10P)	contiguous cells with same land cover grouped to patches
4. Landscape	Heterogeneity (HTG)	independent of landscape texture, more sensitive to number of land cover classes
	Dominance (DMG)	independent of landscape texture, less sensitive to number of land cover classes
	Entropy (ENT)	dependent on landscape texture, more sensitive to number of land cover classes
	Contagion (CON)	dependent on landscape texture, less sensitive to number of land cover classes

* if all binaries are zero then the class is LGN = 7

Table 2. Minimum and maximum values of landscape variables

landscape	HTG	DMG	landscape	ENT	CON
heterogeneous	ln(*I*)	0	fine texture	2·ln(*I*)	0
one class dominates	0	ln(*I*)	coarse texture	0	2·ln(*I*)

I is the number of land cover classes in a region

3.2.3 Statistical analysis

Logistic regression

We used logistic regression (Agresti 1990, Collett 1991) to calculate the probability of correct classification $p_{corr}(c)$ of a cell c as a function of the explanatory variables introduced in the previous section. A logistic regression model with intercept b_0 and with $k = 1,\ldots,K$ explanatory variables x_k equals:

$$p_{corr}(c) = \frac{\exp(\beta_0 + \sum_{k=1}^{K} \beta_k \cdot x_k(c))}{1 + \exp(\beta_0 + \sum_{k=1}^{K} \beta_k \cdot x_k(c))} \tag{7}$$

A linear function logit($p_{corr}(c)$) = $\beta_0 + \sum \beta_k \cdot x_k(c)$ is fitted through the data. In some cases the linearity of this function can be improved by transforming a variable x. Preliminary data analysis showed that this was the case for patch size, which we transformed to a logarithmic scale.

Regression coefficients are obtained by minimising the –2 log likelihood (also known as the deviance) of the model. The difference between the deviances of two models follows a χ_l^2 distribution, where l denotes the number of explanatory variables additional to those shared by the two models. A χ^2 test can then be used to test if adding these l variables to the model significantly improves the fit of the model. Ignorance of spatial dependence in the residuals of the fitted regression models will result in models seeming more significant than they actually are (Bio 2000). We visually checked for spatial dependence in the residuals by plotting their variograms.

We applied an exhaustive model selection procedure that would result in finding the model containing the highest number of significant (at $\alpha = 0.05$) explanatory variables. Table 3 lists the evaluated models, table 6 the tests. At each step in the procedure we tested at three significance levels (0.05, 0.01 and 0.001) the significance of the addition of a variable to a model:

1. addition of CLASS to model 0 (creates model 1), addition of HET to model 0 (creates model 2a), …, addition of CON to model 0 (creates model 2g);
2. addition of HET to model 1 (creates model 3a), …, addition of CON to model 1 (creates model 3g);
3. addition of a second explanatory variable to a model containing CLASS and one variable of the same category (table 1). For example addition of HOM to model 3a. Because none of these additions was significant (at $\alpha = 0.05$) we continued the analysis with one variable out of each category: HET, L10P and DMG;
4. addition of one of the variables HET, L10P and DMG to a model containing CLASS and one of these three. For example addition of L10P to model 3a (creates model 4a);
5. addition of HET to model 4c and addition of L10P·DMG to model 4c.

Table 3. Models evaluated

model nr. (m)	Model
0	β_0
1	$\beta_0 + \beta_{1\text{-}6} \cdot \text{CLASS}$
2a	$\beta_0 + \beta_1 \cdot \text{HET}$
⋮	⋮
2g	$\beta_0 + \beta_1 \cdot \text{CON}$
3a	$\beta_0 + \beta_{1\text{-}6} \cdot \text{CLASS} + \beta_7 \cdot \text{HET}$
⋮	⋮
3g	$\beta_0 + \beta_{1\text{-}6} \cdot \text{CLASS} + \beta_7 \cdot \text{CON}$
4a	$\beta_0 + \beta_{1\text{-}6} \cdot \text{CLASS} + \beta_7 \cdot \text{HET} + \beta_8 \cdot \text{L10P}$
4b	$\beta_0 + \beta_{1\text{-}6} \cdot \text{CLASS} + \beta_7 \cdot \text{HET} + \beta_8 \cdot \text{DMG}$
4c	$\beta_0 + \beta_{1\text{-}6} \cdot \text{CLASS} + \beta_7 \cdot \text{L10P} + \beta_8 \cdot \text{DMG}$
5a	$\beta_0 + \beta_{1\text{-}6} \cdot \text{CLASS} + \beta_7 \cdot \text{L10P} + \beta_8 \cdot \text{DMG} + \beta_9 \cdot \text{HET}$
5b	$\beta_0 + \beta_{1\text{-}6} \cdot \text{CLASS} + \beta_7 \cdot \text{L10P} + \beta_8 \cdot \text{DMG} + \beta_9 \cdot \text{L10P} \cdot \text{DMG}$

For each of the six binary variables in CLASS a regression coefficient is estimated: $\beta_1 \ldots \beta_6$.

Validation

A leave-one-out procedure was applied to cross-validate each model. Each time the model was fit on data from 54 regions and one region r was used as the test data set (figure 1). The resulting 55 parameter sets are refered to as versions of a model. For each of the cells c = 1, …, $n(r)$, with $n(r)$ the number of cells in region r, the probability of correct classification $p_{corr,m}(c)$ was calculated with model m. The binary variable $y(c)$ obtained from the test data set indicates if c was actually correctly classified or misclassified. The measure SM_m (eq. 8) summarises over all cells the absolute difference between model estimated and actual correctness of classification. $R0_m$ (eq. 9) is the relative improvement of model m to the model assuming the same probability of correct classification for all cells (m = 0), $R1_m$ (eq. 10) is the relative improvement to the model with CLASS as the explanatory variable (m = 1). SM_m, $R0_m$ and $R1_m$ are calculated as:

$$SM_m = \sum_{r=1}^{55} \sum_{c=1}^{n(r)} \left| p_{corr,m}(c) - y(c) \right| \quad (8)$$

$$R0_m = 100\% \cdot (SM_0 - SM_m) / SM_0 \quad (9)$$

$$R1_m = 100\% \cdot (SM_1 - SM_m) / SM_1 \quad (10)$$

3.3 Results

3.3.1 Error matrix

The error matrix is shown in table 4. Overall LGN has a high classification accuracy: 90.2%. There are large differences in user accuracies between different crops (ranging between 33.3% and 95.8%). An implication of the large differences is that in for example a region with a relatively large area of bulb cultivation, the assumption of a PCC of 90.2% could overestimate this region's PCC.

3.3.2 Model Selection

Table 5 shows the estimated regression coefficients for a selection of models. Model 1 shows the effect of CLASS on classification accuracy, models 2a-3g show the effect of the explanatory variables HET to CON. Models 3a-3g show the effect of these variables when CLASS is accounted for. Model 4c is the model that contained the highest number of significant (α = 0.05) explanatory variables. If $\hat{\beta}_k < 0$, the probability of correct classification decreases with increase of the value of variable k, if $\hat{\beta}_k > 0$ the probability increases. The impact of the variable k decreases as $\hat{\beta}_k$ approaches 0. Of interest are the effect of our explanatory variables on classification accuracy and the influence of accounting for CLASS. We see that classification accuracy is higher for higher values of focal homogeneity, patch size, in regions with landscapes with one dominant class and in regions with a coarser texture.

Table 4. Error matrix

LGN – test data	REF – reference data grass	 maize	 potatoes	 beets	 cereals	 other crops	 bulb cultivation	total	users' accuracy (%)
grass	566	12	4	1	4	13	1	601	94.2
maize	1	182	4	2	0	1	0	190	95.8
potatoes	7	5	99	0	3	2	0	116	85.3
beets	1	3	1	56	0	0	0	61	91.8
cereals	4	0	0	0	122	9	0	135	90.4
other crops	4	4	16	2	6	20	0	52	38.5
bulb cultivation	3	1	0	0	0	0	2	6	33.3
total	586	207	124	61	135	45	3	1 161	
producers' accuracy (%)	96.6	87.9	79.8	91.8	90.4	44.4	66.7		PCC (%) 90.2

Table 5. Estimated regression coefficients

Model	regression coefficients								
nr. (m)	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
0	2.2174								
1	-0.6931	3.4764	3.8177	2.4551	3.1091	2.9322	0.2231		
2a	2.8899	-0.4751							
2b	1.3078	0.1277							
2c	-0.4525	0.9934							
2d	5.7134	-1.8275							
2e	-0.1922	2.0886							
2f	5.8240	-1.3646							
2g	-3.3758	1.6022							
3a	0.0316	3.3821	3.9384	2.4439	3.1197	2.9139	0.2028	-0.4899	
3b	-1.4315	3.3831	3.8174	2.4018	3.0770	2.8734	0.1666	0.1125	
3c	-2.7223	2.1251	3.5502	2.0543	3.0299	2.4440	-0.1030	1.0404	
3d	2.0440	3.4337	4.0300	2.9414	3.5135	3.2936	0.6484	-1.5253	
3e	-2.5261	3.2388	3.7770	2.8400	3.4262	3.2043	0.5814	1.5559	
3f	2.2878	3.4579	4.0815	2.9336	3.5296	3.2677	0.6347	-1.1993	
3g	-4.7307	3.1681	3.6906	2.7595	3.3866	3.1146	0.5129	1.1663	
4c	-3.5752	2.1731	3.4820	2.2875	3.1923	2.6263	0.1395	0.9248	0.9309

See table 3 for model descriptions. For example in model 2c, $\hat{\beta}_1$ is multiplied by L10P, in model 3c $\hat{\beta}_1$ is multiplied by the binary indicating if CLASS = 1 and $\hat{\beta}_7$ is multiplied by L10P.

As an illustration figures 2 to 4 show for the models 2a, 2c and 2e the expected probabilities of correct classification and the 95% confidence intervals of correct classification. The broad confidence interval for HET < 2 is due to the distribution of the values of HET in the observations: HET = 1: 68%, HET = 2: 29% and HET < 2: 3%. If we compare the $\hat{\beta}_1$ ’s of models 2a-3g with the $\hat{\beta}_7$ ’s of models 3a-3g we see that the effect of the focal and patch variables does not change when CLASS is accounted for. The impact of the landscape variables (models 2d-3g and 3d-3g) is lower in a model containing variable CLASS, indicating that a part of the impact of these variables is accounted for by CLASS.

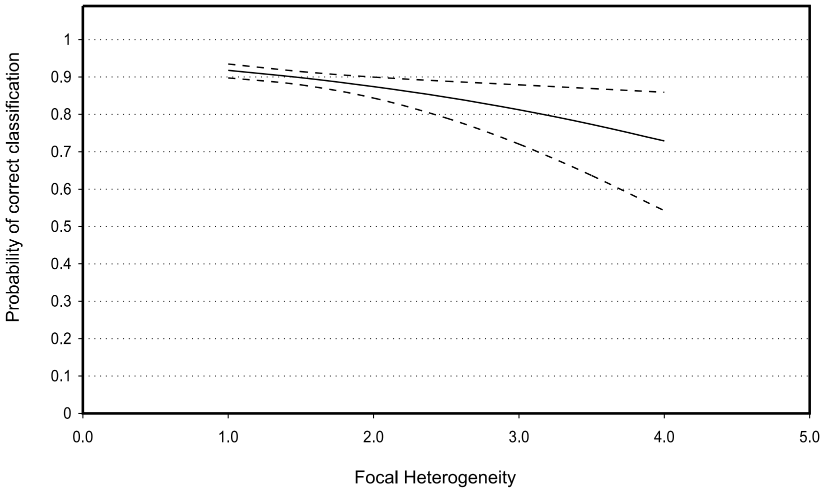


Figure 2. Impact of focal heterogeneity (HET) on probability of correct classification: model 2a (solid line) and 95% confidence interval (dotted line).

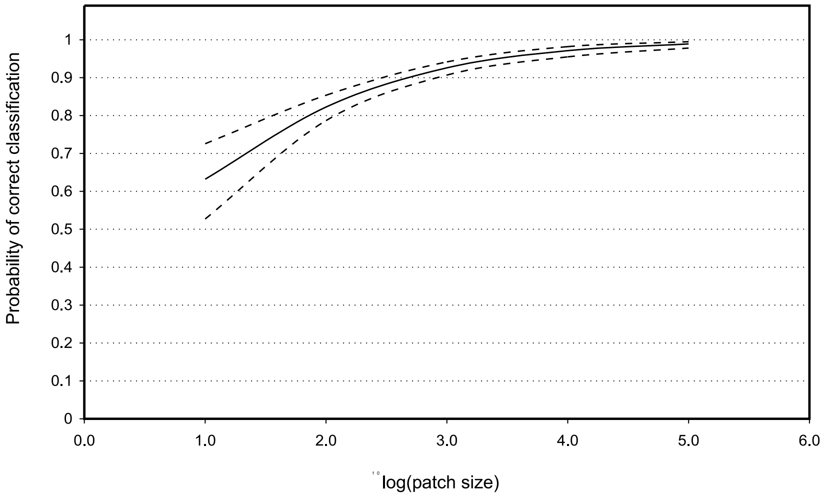


Figure 3. Impact of patch size (L10P) on probability of correct classification: model 2c (solid line) and 95% confidence interval (dotted line).

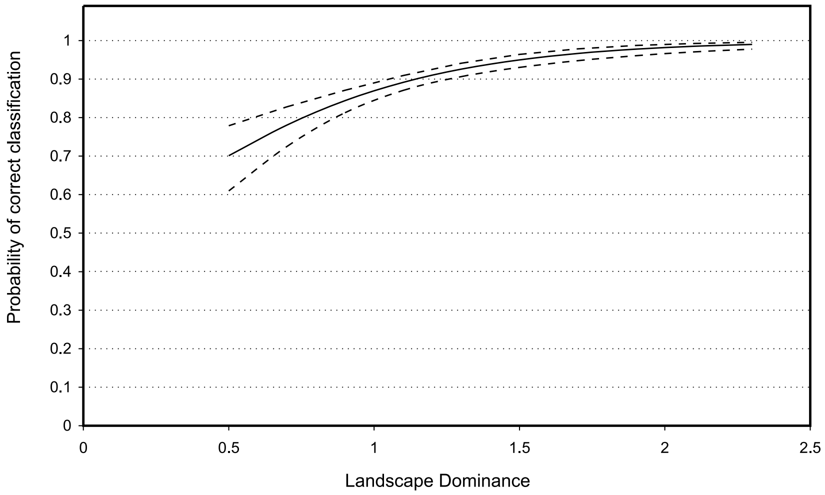


Figure 4. Impact of landscape dominance (DMG) on probability of correct classification: model 2e (solid line) and 95% confidence interval (dotted line).

Figure 5 shows for model 2c that there is no risk of overestimation of significance levels due to spatial dependence in the residuals. Variograms of residuals of the other models also revealed no spatial dependence. Table 6 shows the significance testing of improvement of fit of models by adding an additional explanatory variable x_l. The table shows that of the explanatory variables only HOM was not significant at α = 0.05. The patch and landscape variables were highly significant (all at α = 0.001) also when added to the model already containing CLASS. HET was significant at α = 0.01 in fewer versions if CLASS was already in the model (54 vs. 24 versions). Adding L10P or DMG to a model containing CLASS and HET was in all 55 versions significant at α = 0.001. Adding HET to respectively CLASS & L10P and CLASS & DMG was significant at the α = 0.05 level in 0 and 54 of the 55 versions. We concluded that a model should at least contain CLASS and either L10P or DMG and that it need not contain HET. But would a model containing CLASS and both L10P and DMG still be better? Adding DMG was significant at α = 0.05 in 55 versions and adding L10P was significant at α = 0.001 in 55 versions. Finally we tested if adding to model 4c the variable HET or the interaction L10P·DMG would significantly (at α = 0.05) improve the fit. This was never the case, so model 4c was the model containing the highest number of significant explanatory variables.

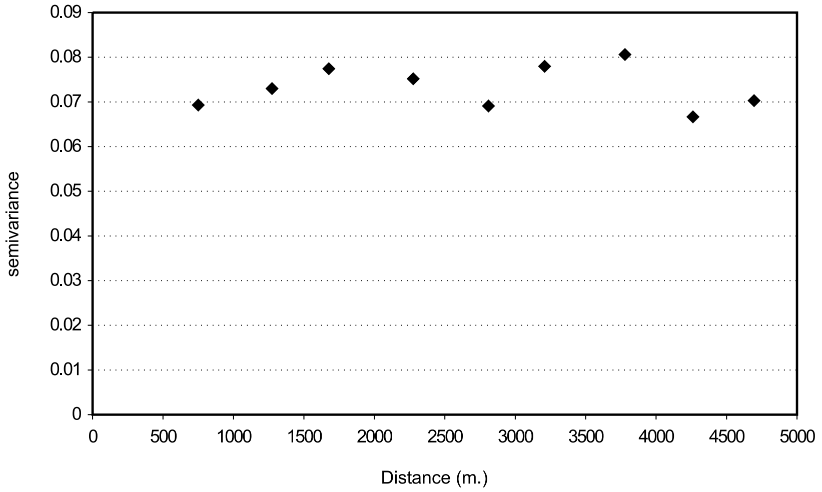


Figure 5. Semivariogram of residuals of model 2c (explanatory variable: L10P).

3.3.3 Model validation

Table 7 shows the absolute differences between actual classification correctness and estimated probabilities of correct classification, summarised over all cells. The table shows that all model estimates were better than model 0 ($R0_m$ > 0 for all m). The models containing focal variables were slightly better than model 0 ($R0_{2a}$ = 1.3%, $R0_{2b}$ = 1.5%), however they produced worse estimates than model 1 ($R1_{2a}$ = -34.4%, $R1_{2b}$ = -34.1%). Their contribution to the model containing variable CLASS was marginal: $R1_{3a}$ = 0.6%, $R1_{3b}$ = 0.2%. The patch and landscape variables provided better estimates than models 0 and 1. Adding a patch or landscape variable to a model containing intercept only (models 2c-2g) improved model 0 estimates by $R0_{2c}$ = 3.5% to $R0_{2g}$ = 19.2%. Adding a patch or landscape variable to a model containing CLASS (models 3c-3g) improved model 1 estimates by 10.6% to 15.7%. Model 4c, the model with the highest number of significant variables, also had the highest $R0_m$ and $R1_m$ values. Relative to models 0 and 1 model 4c improved estimates of probabilities of correct classification by 42.0% and 21.0% respectively.

Table 6. Chi-square tests for selected models

chi-square test (df.)	Description: significance of additional explanatory variable x_l to a model already containing variables x_k		Frequencies of significance at α in the 55 versions* of each model		
	x_k	x_l	α = 0.05	α = 0.01	α = 0.001
$D_0 - D_1$ (6)	1	CLASS	55	55	55
$D_0 - D_{2a}$ (1)	1	HET	55	54	0
$D_0 - D_{2b}$ (1)	1	HOM	5	0	0
$D_0 - D_{2c}$ (1)	1	L10P	55	55	55
$D_0 - D_{2d}$ (1)	1	HTG	55	55	55
$D_0 - D_{2e}$ (1)	1	DMG	55	55	55
$D_0 - D_{2f}$ (1)	1	ENT	55	55	55
$D_0 - D_{2g}$ (1)	1	CON	55	55	55
$D_1 - D_{3a}$ (1)	1 & CLASS	HET	54	24	0
$D_1 - D_{3b}$ (1)	1 & CLASS	HOM	0	0	0
$D_1 - D_{3c}$ (1)	1 & CLASS	L10P	55	55	55
$D_1 - D_{3d}$ (1)	1 & CLASS	HTG	55	55	55
$D_1 - D_{3e}$ (1)	1 & CLASS	DMG	55	55	55
$D_1 - D_{3f}$ (1)	1 & CLASS	ENT	55	55	55
$D_1 - D_{3g}$ (1)	1 & CLASS	CON	55	55	55
$D_{3a} - D_{4a}$ (1)	1 & CLASS & HET	L10P	55	55	55
$D_{3a} - D_{4b}$ (1)	1 & CLASS & HET	DMG	55	55	55
$D_{3c} - D_{4a}$ (1)	1 & CLASS & L10P	HET	0	0	0
$D_{3c} - D_{4c}$ (1)	1 & CLASS & L10P	DMG	55	7	0
$D_{3e} - D_{4a}$ (1)	1 & CLASS & DMG	HET	54	4	0
$D_{3e} - D_{4c}$ (1)	1 & CLASS & DMG	L10P	55	55	55
$D_{4c} - D_{5a}$ (1)	1 & CLASS & L10P & DMG	HET	0	0	0
$D_{4c} - D_{5b}$ (1)	1 & CLASS & L10P & DMG	L10P·DMG	0	0	0

* a version is a model fitted on data from 54 regions, with cells of one region left out for cross-validation

D_m: Deviance of model m; variables are described in table 1, models in table 3; $x_k = 1$ is multiplied by the intercept $\hat{\beta}_0$.

3.3.4 Model validation versus model selection

In general the ranking of models according to significance levels of contributing variables in the model selection corresponded well with the ranking of models according to $R0_m$ and $R1_m$ values in the validation. Landscape variables yielded the greatest improvement in estimates of per cell probabilities of correct classification, the patch variable yielded a greater improvement than the focal variables. Model selection showed that adding CLASS to a model containing DMG or CON was significant at $\alpha = 0.001$ (testing $\chi^2_{l=6} = D_{2e} - D_{3e}$ and $\chi^2_{l=6} = D_{2g} - D_{3g}$), yet validation showed a deterioration when CLASS was added ($R1_{3e} < R1_{2e}$ and $R1_{3g} < R1_{2g}$). Finally, the big differences in significance between HET and HOM observed during model selection were absent in the model validation.

Table 7. Validation: comparison between actual and estimated number of cells correctly classified

Model nr. (*m*)	Model description	SM_m	relative to models 0 and 1 $R0_m$ (%)	 $R1_m$ (%)
0	same PCC in all regions	83.0		
1	CLASS (= error matrix)	60.9	26.6	
2a	HET	81.8	1.3	-34.4
2b	HOM	81.7	1.5	-34.1
2c	L10P	58.8	29.1	3.5
2d	HTG	58.3	29.8	4.3
2e	DMG	51.0	38.5	16.2
2f	ENT	58.5	29.5	3.9
2g	CON	49.2	40.7	19.2
3a	CLASS & HET	60.5	27.0	0.6
3b	CLASS & HOM	60.8	26.8	0.2
3c	CLASS & L10P	52.4	36.8	13.9
3d	CLASS & HTG	54.4	34.4	10.6
3e	CLASS & DMG	51.5	37.9	15.4
3f	CLASS & ENT	54.0	35.0	11.4
3g	CLASS & CON	51.4	38.1	15.7
4c	CLASS & L10P & DMG	48.1	42.0	21.0

3.3.5 Comparison between landscape variables

Table 7 shows that $R1_{3e} - R1_{3d} > R1_{3f} - R1_{3d}$, indicating a stronger increase in *R1* values when HTG was replaced by DMG (10.6% to 15.4%) than when HTG was replaced by ENT (10.6% to 1.4%). Similar comparisons between $R0_m$ and $R1_m$ values of other models, for example $R1_{2g} - R1_{2f} > R1_{2g} - R1_{2e}$, confirmed that PCC estimates improved stronger when a landscape variable less sensitive to the number of land cover classes was included in the model (DMG or CON) than when a variable sensitive to texture was included (ENT or CON). The table further shows that the impact of replacing HTG by DMG or ENT by CON is smaller in models 3d-3g than in models 2d-2g. This indicates that a part of the effect of accounting for sensitivity to the number of land cover classes is accounted for by the inclusion of the variable CLASS in the model.

3.4 Conclusion and discussion

Our aim was to extend the work by Smith et al. (2002, 2003) by evaluating landscape indices (Forman and Godron 1986, O'Neill et al. 1988, Li and Reynolds 1993, Riiters et al. 1995) as potential explanatory variables of variability in the classification accuracy of cells. Four landscape indices were applied, differing in sensitivity to the number of land cover classes and in ability to distinguish between different landscape textures. In this section we will discuss the outcomes of the model selection and validation, and then separately discuss the impact of the landscape variables.

3.4.1 Model selection and validation

Our model selection procedure showed the significance of variables in a series of models with increasing number of variables. The model with the highest number of significant ($\alpha = 0.05$) explanatory variables included the variables 'land cover class', 'patch size' and 'landscape dominance'. We found that per cell classification accuracy was significantly higher for cells with more heterogeneous focal (= 3×3) neighbourhoods, cells located in larger patches, cells located in regions with a less heterogeneous landscape and cells located in regions with a more coarsely textured landscape.

To assess the validity of our statistical analysis we checked spatial dependence in the model residuals (Bio 2000). Variograms of the residuals did not reveal such dependence. A leave-one-out procedure was applied to validate the models. Improvement was measured relative to the model assuming the same probability of correct classification for all cells and relative to the model with 'land cover class' as the single explanatory variable. The model with the highest number of significant explanatory variables also yielded the largest relative improvements: respectively 21% and 42%. In general the ranking of models according to significance levels during model selection corresponded well with the ranking according to relative improvements during the model validation.

3.4.2 Impact of landscape variables

Model selection showed in all stages a significant contribution of the landscape variables HTG, DMG, ENT and CON. The impact was smaller but still significant if the model contained variable 'land cover class', which indicates that a part of the impact of these variables is already accounted for by this variable. Validation results showed that estimates of per cell probabilities of correct classification were better in models with landscape variables less sensitive the number of land cover classes (DMG and CON) and that estimates were slightly better in models with variables sensitive to landscape texture (ENT and CON).

4. Land cover change and classification errors[1]

[1] Based on: van Oort, P.A.J., 2005. Improving land cover change estimates by accounting for classification errors. International Journal of Remote Sensing 26(14): 3009-3024 © 2005 Taylor & Francis Ltd.

4. Land cover change and classification errors

Abstract

In monitoring land cover change by overlay of two maps of different dates the rate of change is frequently overestimated. This is due to three sources of uncertainty: (i) semantic differences in class definitions between two maps, (ii) positional errors and (iii) classification errors. In this study 4 methods are proposed that use the Bayes theorem to update prior estimates of land cover change with information on the probabilities with which land cover classes are mistaken for each other. The methods were illustrated for two case-studies. In the first case-study the real change was 1.4% and by overlay of the two maps 7.4% was predicted. The estimates by the four methods were 6.3%, 15.3%, 6.7% and 1.6%. In the second study these percentages were 48%, 36% and with our 4 methods: 39.2%, 54.1%, 50.8% and 53.0%. Two of the methods account for correlation in classification accuracy between maps of two dates. Where this correlation was high (study area 1), the methods that accounted for correlation yielded change estimates closer to real change than the methods that did not account for this correlation.

4.1 Introduction

4.1.1 Change detection

Land cover change seriously affects our environment (Vitousek 1994, Achard et al. 2002), so it is important to have good estimates of these changes. National land cover databases are often used for monitoring land cover change; research is increasingly focusing on improvements in change detection methods (Lunetta et al. 2002). An important issue is the effect of positional errors and classification errors on change detection.

Different change detection methods exist (Singh 1989, MacLeod and Congalton 1998, Mas 1999). The most obvious method is post-classification comparison of land cover at two dates. The use of post-classification comparison is appropriate when the phenomenon monitored can be partitioned into crisp classes and when data can be stored and collected at a resolution so high that the spatial objects are homogeneous in their land cover class. Otherwise, image differencing, principal component analysis or hierarchical fuzzy pattern matching may be more applicable (Singh 1989, MacLeod and Congalton 1998, Power et al. 2001). The transition matrix is often used as a non-site-specific measure of land cover change. It shows the total area of transitions between different land cover classes and is derived from an overlay of test data sets of two dates.

4.1.2 Sources of uncertainty in change detection

Numerous studies have shown that post-classification comparisons overestimate land cover change, due to uncertainties in the data (MacLeod and Congalton 1998, Mas 1999, de Zeeuw and Hazeu 2001, Achard et al. 2002, Foody 2002). These uncertainties originate from three sources: (1) semantic differences in the definitions of land cover classes between two data sets, (2) misregistration of pixels or object boundaries and (3) misclassifications.

Only when uncertainty due to semantics has been eliminated can the effect of misregistrations and misclassifications be quantified and reduced. Uncertainty due to semantics can be eliminated either by using a consistent classification method (Achard et al. 2002) or by developing a common ontology by generalization of the classes in the available data sets (de Zeeuw and Hazeu 2001, Petit and Lambin 2002). The first solution has the disadvantage of being very costly, requiring restructuring, often of large data sets; the second results in a loss of detail. Most monitoring studies only address this first source of uncertainty.

For satellite images, the most researched method to reduce the effect of misregistration on change detection is by reduction of misregistration (Townshend et al. 1992, Dai and Khorram 1998, Stow and Chen 2002). Just a few of these studies used post-classification comparison as a change detection method. Verbyla and Boles (2000) simulated the effect of positional error in image rectification on land cover change. They found a low correlation between registration root mean square error and the overestimation in land cover change. Carmel et al. (2001) and Kiiveri et al. (2001) showed that misregistration can result in classification errors, which can be reported in the error matrix. For vector data, misregistration refers to the positional accuracy of boundaries of objects (polygons). De Zeeuw et al. (1999), Sonneveld et al. (2000) and de Zeeuw and Hazeu (2001) developed a model called MonGIS for reducing the effect of such misregistrations. MonGIS uses rules based on expert knowledge of the geometry of changing spatial objects to decide if the detected change was real or not. The use of expert knowledge is problematic in that it is difficult to validate, and one cannot be confident about the applicability to other data sets than the calibration data set. The present study is inspired by the work by de Zeeuw et al. (1999); methods are proposed that do not require expert knowledge, but instead require information on classification errors.

Classification errors may arise due to misregistration and due to assignment of the wrong class to spatial objects. The propagation of classification errors depends on the size of the classification errors, not on the source of these errors. (Kiiveri et al. 2001). Veregin (1989) showed how propagation of classification errors at two observation dates in an overlay operation can be calculated. His method addressed calculating the percentage correctly classified of the resulting map. Rosenfield (1982) presented a method of estimating change and variance in land cover change for individual land cover classes as a function of sample design. Kiiveri et al. (2001) reduced mis-estimation of change by using data sets of increased classification accuracy. They increased the classification accuracy by using a priori knowledge of the actual transition probabilities. In this study, I have no prior knowledge of actual transition probabilities, and I will estimate actual transition probabilities by including information on classification accuracy.

4.1.3 Objective

Research into methods developed to reduce overestimation or underestimation of land cover change estimates has predominantly been on increasing the quality of single date data sets and on comparison of change detection methods. Both are often not an option to the user, who is supplied with ‘off the shelf’ data sets and their quality descriptions. The aim of this study is to propose four methods to improve land cover change estimates by accounting for classification errors. For illustration, the methods were applied to two case study areas using data described in Veregin (1989) and in de Zeeuw and Hazeu (2001). In both case studies, exhaustive reference data sets were used to evaluate the methods proposed. A comparison of the proposed four methods with the model MonGIS (de Zeeuw et al. 1999) was made.

4.2 Materials and Methods

4.2.1 Data processing

This section introduces terminology and describes how data were processed. The data sets used in the two case-studies will be described in a subsequent section. In each study, two classified land cover data sets, of dates t1 and t2, were used and also two exhaustive reference data sets with the actual land cover classes at dates t1 and t2. Land cover classes are described by the variable $LC_{a,b}$, where subscripts 'a' refers to a data set or method and 'b' to the date. From five overlays of these data sets the following matrices and table were obtained:

1. Test data set at t1 and test data set at t2 → post classification comparison (PCC) transition matrix, with elements Area($LC_{data,t1}$, $LC_{data,t2}$);
2. Reference data set at t1 and reference data set at t2 → actual transition matrix, with elements Area($LC_{act,t1}$, $LC_{act,t2}$);
3. Test and reference data set at t1 → error matrix of t1, with elements Area($LC_{data,t1}$, $LC_{act,t1}$);
4. Test and reference data set at t2 → error matrix of t2, with elements Area($LC_{data,t2}$, $LC_{act,t2}$);
5. All four data sets → table with Area($LC_{data,t1}$, $LC_{act,t1}$, $LC_{data,t2}$, $LC_{act,t2}$).

Transition probabilities are obtained from the transition matrices. For example from the PCC transition matrix Pr($LC_{data,t2}$ | $LC_{data,t1}$) is the probability that a site is classified as $LC_{data,t2}$ at t2, given that it is classified $LC_{data,t1}$ at t1. Confusion probabilities are derived from the error matrices, indicating the probability with which two classes are mistaken for each other. For example Pr($LC_{data,t1}$ | $LC_{act,t1}$) is the probability that the classified land cover class at t1 is $LC_{data,t1}$, given that the actual class is $LC_{act,t1}$ at t1. Correlation of confusion probabilities at one date with another date, is quantified in confusion probabilities derived from the fifth overlay. From the table one can derive the probability Pr($LC_{act,t2}$ | $LC_{data,t2}$, $LC_{data,t1}$, $LC_{act,t1}$), i.e. the probability that the actual land cover class at t2 is $LC_{act,t2}$ given that a site is classified as $LC_{data,t2}$ at t2, classified as $LC_{data,t1}$ at t1 and has the actual class $LC_{act,t1}$ at t1. For each combination {$LC_{data,t1}$, $LC_{act,t1}$}, the following equality (1) holds:

$$\sum_{LC_{data,t2}} \Pr(LC_{act,t2} \mid LC_{data,t2}, LC_{data,t1}, LC_{act,t1}) = \begin{cases} 0 & \text{if } \mathrm{Area}(LC_{data,t1}, LC_{act,t1}) = 0 \\ 1 & \text{if } \mathrm{Area}(LC_{data,t1}, LC_{act,t1}) > 0 \end{cases} \tag{1}$$

If the sum is 0 then this specific combination does not occur. In that case the probability not accounting for temporal correlation is used: Pr($LC_{act,t2}$ | $LC_{data,t1}$, $LC_{act,t1}$, $LC_{data,t2}$) = Pr($LC_{act,t2}$ | $LC_{data,t2}$).

4.2.2 Proposed methods to reduce errors in change detection

In the Bayes theorem (O'Hagan 1994), prior estimates of a variable of interest are updated with data to obtain posterior estimates of the variable of interest. In this study, the variables of interest are the actual transitions, prior estimates are the PCC transitions, and data are the confusion probabilities. The posterior estimates are the elements Area($LC_{M,t1}$, $LC_{M,t2}$) of the posterior transition matrix, calculated with one of the methods M proposed in this section. Multiplication of the posterior transition probabilities Pr($LC_{M,t2}|LC_{M,t1}$) or Pr($LC_{M,t1}|LC_{M,t2}$) with posterior area estimates (equation (2) or (3)) yields the elements Area($LC_{M,t1}$, $LC_{M,t2}$) of the posterior transition matrix (equation (4) or (5)):

$$\text{Area}(LC_{M,t1}) = \sum_{LC_{data,t1}} \text{Area}(LC_{data,t1}) \cdot \Pr(LC_{act,t1} \mid LC_{data,t1}) \qquad (2)$$

$$\text{Area}(LC_{M,t2}) = \sum_{LC_{data,t2}} \text{Area}(LC_{data,t2}) \cdot \Pr(LC_{act,t2} \mid LC_{data,t2}) \qquad (3)$$

$$\text{Area}(LC_{M,t1}, LC_{M,t2}) = \text{Area}(LC_{M,t1}) \cdot \Pr(LC_{M,t2} \mid LC_{M,t1}) \qquad (4)$$

$$\text{Area}(LC_{M,t1}, LC_{M,t2}) = \text{Area}(LC_{M,t2}) \cdot \Pr(LC_{M,t1} \mid LC_{M,t2}) \qquad (5)$$

The four methods proposed here differ in the calculation of the posterior transition probabilities, Pr($LC_{M,t2} \mid LC_{M,t1}$). In the first method, all elements in the PCC transition matrix are multiplied by the rate of overestimation or underestimation of land cover classes at t1. In the second method, confusion probabilities derived from the 3rd and 4th matrix, are used to update PCC transition probabilities. Specifically, 4 situations can occur:

1. correct classification at t1 and at t2 ($LC_{data,t1} = LC_{act,t1}$, $LC_{data,t2} = LC_{act,t2}$);
2. misclassification at t1, correct at t2 ($LC_{data,t1} \neq LC_{act,t1}$, $LC_{data,t2} = LC_{act,t2}$);
3. correct at t1, misclassification at t2 ($LC_{data,t1} = LC_{act,t1}$, $LC_{data,t2} \neq LC_{act,t2}$);
4. misclassification at t1 and t2 ($LC_{data,t1} \neq LC_{act,t1}$, $LC_{data,t2} \neq LC_{act,t2}$).

The third and fourth methods also update PCC transition probabilities using confusion probabilities at t1 and at t2, but these methods use confusion probabilities from the 5th matrix. In the third method, confusion probabilities at t2 are conditioned on the confusion between classes at t1, {$LC_{data,t1}$, $LC_{act,t1}$}. In the fourth method, confusion probabilities at t1 are conditioned on the confusion between classes at t2, {$LC_{data,t2}$, $LC_{act,t2}$}. The equations for the four methods are listed below:

Method 1:

$$\Pr(LC_{M,t2} \mid LC_{M,t1}) = \frac{\text{Area}(LC_{act,t1})}{\text{Area}(LC_{data,t1})} \cdot \Pr(LC_{data,t2} \mid LC_{data,t1}) \qquad (6)$$

with $LC_{M,t1} = LC_{data,t1} = LC_{act,t1}$ and $LC_{M,t2} = LC_{data,t2} = LC_{act,t2}$

Method 2:

$$\Pr(LC_{M,t2} \mid LC_{M,t1}) = \sum_{LC_{data,t1}} \sum_{LC_{data,t2}} \Pr(LC_{data,t1} \mid LC_{act,t1}) \cdot \Pr(LC_{data,t2} \mid LC_{data,t1}) \cdot \Pr(LC_{act,t2} \mid LC_{data,t2}) \tag{7}$$

with $LC_{M,t1} = LC_{act,t1}$ and $LC_{M,t2} = LC_{act,t2}$

Method 3:

$$\Pr(LC_{M,t2} \mid LC_{M,t1}) = \sum_{LC_{data,t1}} \sum_{LC_{data,t2}} \Pr(LC_{data,t1} \mid LC_{act,t1}) \cdot \Pr(LC_{data,t2} \mid LC_{data,t1}) \cdot \Pr(LC_{act,t2} \mid LC_{data,t2}, LC_{data,t1}, LC_{act,t1}) \tag{8}$$

with $LC_{M,t1} = LC_{act,t1}$ and $LC_{M,t2} = LC_{act,t2}$

Method 4:

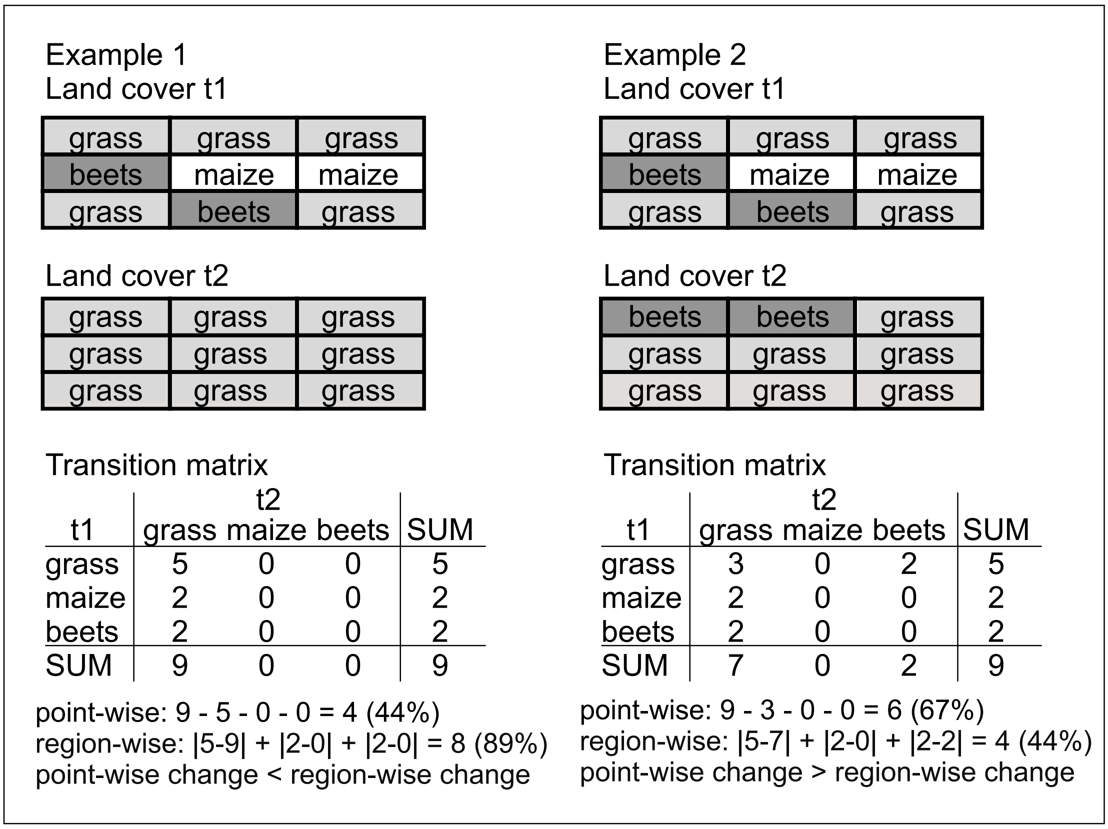
$$\Pr(LC_{M,t1} \mid LC_{M,t2}) = \sum_{LC_{data,t1}} \sum_{LC_{data,t2}} \Pr(LC_{data,t2} \mid LC_{act,t2}) \cdot \Pr(LC_{data,t1} \mid LC_{data,t2}) \cdot \Pr(LC_{act,t1} \mid LC_{data,t1}, LC_{data,t2}, LC_{act,t2}) \tag{9}$$

with $LC_{M,t1} = LC_{act,t1}$ and $LC_{M,t2} = LC_{act,t2}$

From equation (2), it follows that if exact estimates of the confusion probabilities are available (this is the case if an exhaustive reference data set is used), then Area($LC_{M,t1}$) ≡ Area($LC_{act,t1}$) and Area($LC_{M,t2}$) ≡ Area($LC_{act,t2}$). Because $\sum_{LC_{M,t2}} \Pr(LC_{M,t2} \mid LC_{M,t1}) = 1$, it follows from equation (4) that the row total of the transition matrix estimated by methods 1–3 is equal to Area($LC_{act,t1}$) and from equation (5) that the column total of the transition matrix estimated by method 4 is equal to Area($LC_{act,t2}$). Moreover, it can be shown that with exact estimates of the confusion probabilities, the column totals of the transition matrix calculated with method 2 will correspond exactly with Area($LC_{act,t2}$). As a result, method 2 will provide exact estimates for one of the measures of land cover change applied in this study.

4.2.3 Measures of land cover change

MacLeod and Congalton (1998) distinguished four aspects in change detection: (i) detecting that changes have occurred, (ii) identifying the nature of the change, (iii) measuring the areal extent of the change and (iv) assessing the spatial pattern of the change. The transition matrix quantifies the first three aspects. De Zeeuw and Hazeu (2001) defined two summary measures of the transition matrix: point-wise (PC) and region-wise (RC) change (figure 1). As a third measure I defined the sum of absolute differences (SAD). Let the transition matrix be equal to T = $\{t\}_{ij}$. Point-wise change is the total area in which change occurred and is calculated from the diagonal elements of the transition matrix: $PC = \sum_{i,j} t_{ij} - \sum_{i} t_{ii}$. Region-wise land cover change is the sum of absolute net changes in the area of the land cover classes and is calculated from the transition matrix row and column totals: $RC = \sum_{i,j=i} | t_{i+} - t_{+j} |$. The sum of absolute differences is calculated from all elements in the actual transition matrix (TA = $\{ta\}_{ij}$) and the transition matrix of which the accuracy is tested (T): $SAD = \sum_{i,j} | ta_{ij} - t_{ij} |$.



Example 1

Land cover t1

grass	grass	grass
beets	maize	maize
grass	beets	grass

Land cover t2

grass	grass	grass
grass	grass	grass
grass	grass	grass

Transition matrix

	t2			
t1	grass	maize	beets	SUM
grass	5	0	0	5
maize	2	0	0	2
beets	2	0	0	2
SUM	9	0	0	9

point-wise: 9 - 5 - 0 - 0 = 4 (44%)
region-wise: |5-9| + |2-0| + |2-0| = 8 (89%)
point-wise change < region-wise change

Example 2

Land cover t1

grass	grass	grass
beets	maize	maize
grass	beets	grass

Land cover t2

beets	beets	grass
grass	grass	grass
grass	grass	grass

Transition matrix

	t2			
t1	grass	maize	beets	SUM
grass	3	0	2	5
maize	2	0	0	2
beets	2	0	0	2
SUM	7	0	2	9

point-wise: 9 - 3 - 0 - 0 = 6 (67%)
region-wise: |5-7| + |2-0| + |2-2| = 4 (44%)
point-wise change > region-wise change

Figure 1. Illustration of point-wise and region-wise land cover change calculation.

4.2.4 Case-studies

For illustration the methods for error reduction described in previous sections were applied to two case-study areas, using data described in de Zeeuw and Hazeu (2001) and in Veregin (1989). In both cases, exhaustive reference data sets were used to ensure that differences in reduction of mis-estimation of change were due to differences between the methods and not due to sampling design. The two study areas differ in mis-estimation of change, classification accuracy and temporal correlation in classification accuracy (table 1).

Table 1. Differences between the two study areas

	study area 1	study area 2
mis-estimation of change	overestimation	underestimation
accuracy at t1	high	low
accuracy at t2	higher than t1	same as t1
temporal correlation	high	low

The expert knowledge in the model MonGIS is valid for study area 1, so a comparison between this model and the methods proposed will be made for this area. Study area 1 is based on data by de Zeeuw and Hazeu (2001). They document an extensive study of land cover change in the municipality Soest in the Netherlands. Soest was chosen because land cover within this municipality was representative for the Netherlands, it was large enough to be representative, and at the same time small enough to manage the collection of field data for an exhaustive reference data set. The municipality has an area of 3,387 ha. As test data set they used the national topographic database (Top10vector) collected in 1991 and in 1995. Top10vector is a vector-based data set that is updated once in four years. Source data are aerial photographs of scale 1:10,000. These are converted into digital images, classified and then verified by topographers of the national topographic service (van Asperen 1997). The reference data sets for 1991 and 1995 were made using the Top10vector data set as a geometric base. Additional information was obtained by means of aerial photo-interpretation, existing detailed studies and maps and a fieldwork campaign. To compare Top10vector with two other national land cover databases, de Zeeuw and Hazeu (2001) generalised the classes in these databases into a common ontology. The common ontology distinguished four land cover classes: Urban & Infrastructure, Forest & Nature, Agriculture and Water. Figure 2 shows the Soest study area and the areas where transitions were detected using an overlay of the two test data sets and using an overlay of the two the reference data sets. Clearly, land cover change is overestimated by the test data sets.

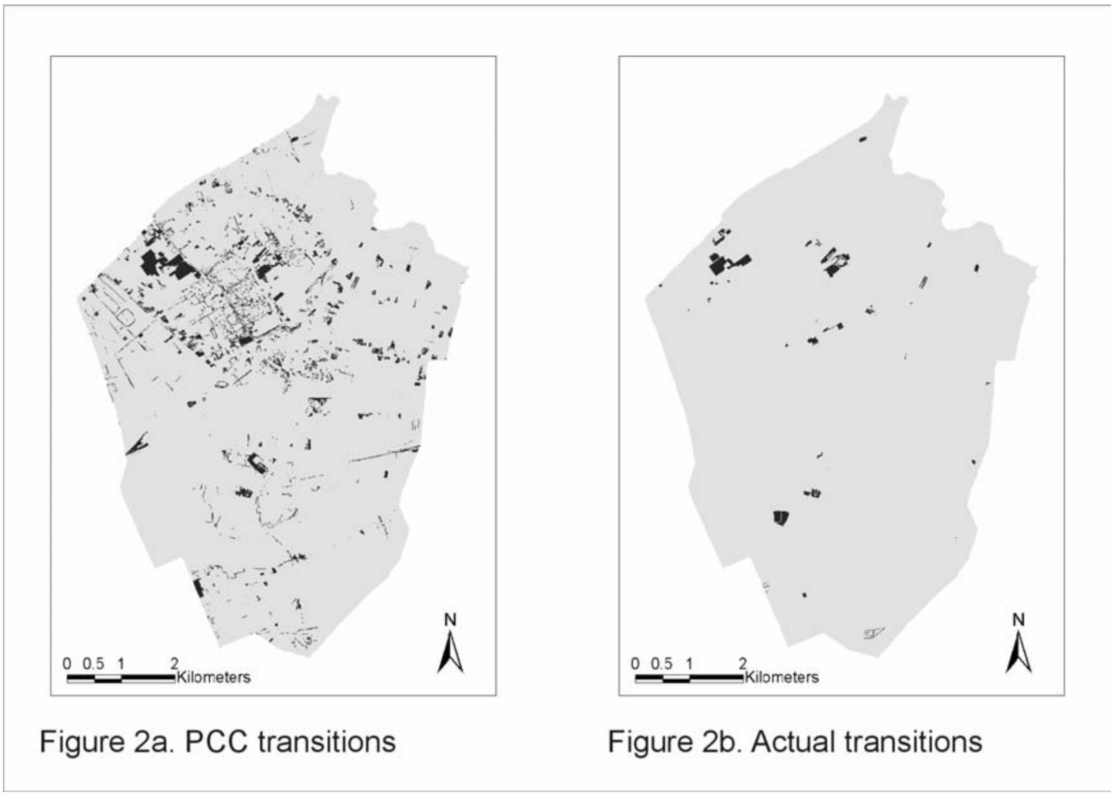
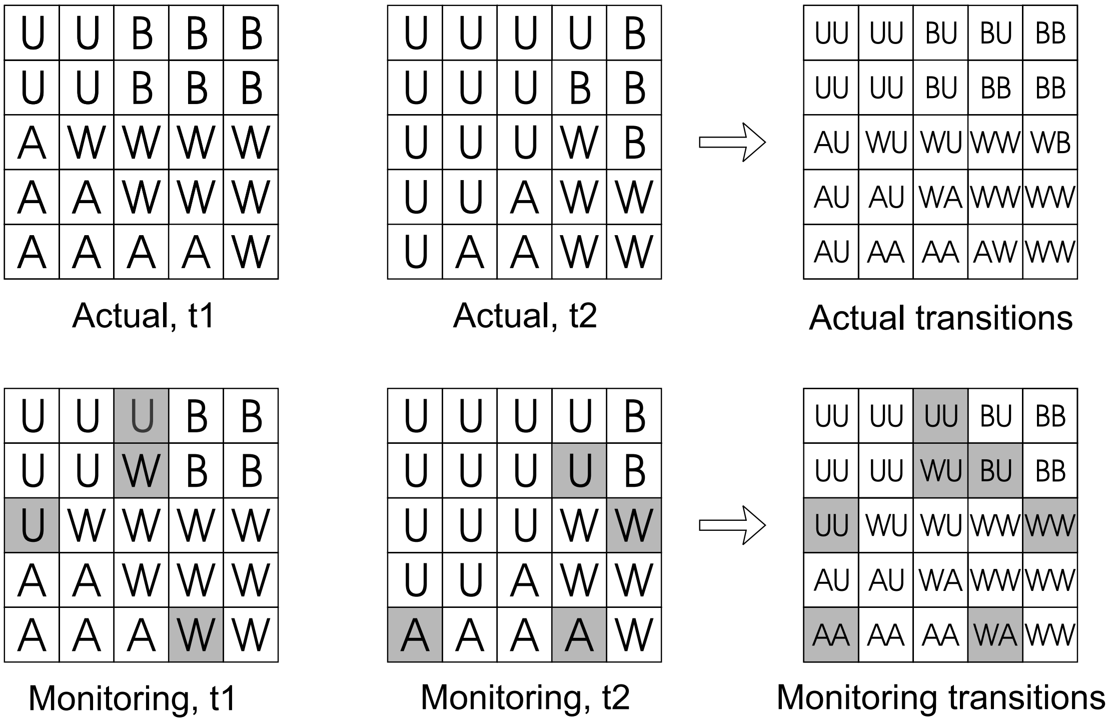


Figure 2a. PCC transitions

Figure 2b. Actual transitions

Figure 2. Study area 2. Municipality Soest (grey), with change between 1991 and 1995 (black). Data derived from de Zeeuw and Hazeu (2001). (a) PCC transitions; (b) actual transitions.

Study area 2 is based on an example by Veregin (1989). Veregin presented four synthetic land cover maps (figure 3) with classes A, B, U and W for illustration of error modelling in an overlay operation. He used these maps to illustrate how for different overlay operations the percentage correctly classified of the resulting map could be calculated. The overlay shown in figure 3 is an AND overlay.



U	U	B	B	B
U	U	B	B	B
A	W	W	W	W
A	A	W	W	W
A	A	A	A	W

Actual, t1

U	U	U	U	B
U	U	U	B	B
U	U	U	W	B
U	U	A	W	W
U	A	A	W	W

Actual, t2

UU	UU	BU	BU	BB
UU	UU	BU	BB	BB
AU	WU	WU	WW	WB
AU	AU	WA	WW	WW
AU	AA	AA	AW	WW

Actual transitions

U	U	U	B	B
U	U	W	B	B
U	W	W	W	W
A	A	W	W	W
A	A	A	W	W

Monitoring, t1

U	U	U	U	B
U	U	U	U	B
U	U	U	W	W
U	U	A	W	W
A	A	A	A	W

Monitoring, t2

UU	UU	UU	BU	BB
UU	UU	WU	BU	BB
UU	WU	WU	WW	WW
AU	AU	WA	WW	WW
AA	AA	AA	WA	WW

Monitoring transitions

Pr(correct)
= 21/25

Pr(correct | t1)
= 18/21

Pr(correct | t1, t2)
= 21/25·18/21=18/25

Figure 3. Study area 2; data derived from Veregin (1989).

4.3 Results and discussion

4.3.1 Classification errors

The error matrices of study area 1 are shown in tables 2 and 3. The percentage correctly classified at t1 (1991) is 90.8% and at t2 (1995) 95.8%. In study area 2 the percentage correctly classified is lower, 84% at both dates. In area 1 the largest error, at both dates, is that Urban & Infra is classified as Agriculture. Table 4 shows the normalised areas of correct and incorrect classification for both study areas. Temporal correlation in classification accuracy is quantified as the probability that classification is correct at both dates. For study area 1 this is 0.93 (0.9002+0.0334), for area 2 it is 0.76 (0.72+0.04). Thus temporal correlation in classification correctness is higher in area 1 than in area 2. Table 4 served to illustrate this correlation. In methods 3 and 4 this temporal correlation was accounted for through conditioning of confusion probabilities. As an example of an unconditioned confusion probability at t2, consider the probability $\Pr(LC_{act,t2} \mid LC_{data,t2})$ = Pr(Agriculture | Urban & Infra) = 0.0079 (=

10/1272, table 3). In method 3 confusion probabilities at t2 are conditioned on the confusion that occurred at t1, $\{LC_{data,t1}, LC_{act,t1}\}$. For example the confusion probability at t2, conditioned on {Agriculture, Agriculture}, is: Pr(Agriculture | Urban & Infra, Agriculture, Agriculture) = 0.2631.

Table 2. Error matrix of study area 1, date t1

	Actual land cover in 1991 (ha.)				
Test land cover in 1991	Urban & Infra	Forest & Nature	Agriculture	Water	Total
Urban & Infra	1104	31	7	0	1141
Forest & Nature	50	1464	1	0	1516
Agriculture	206	14	482	1	703
Water	0	0	0	27	28
Total	1361	1509	489	28	3387

Table 3. Error matrix of study area 1, date t2

	Actual land cover in 1995 (ha.)				
Test land cover in 1995	Urban & Infra	Forest & Nature	Agriculture	Water	Total
Urban & Infra	1260	3	10	0	1272
Forest & Nature	1	1505	0	0	1506
Agriculture	127	1	453	0	580
Water	0	0	0	29	29
Total	1388	1508	462	29	3387

Table 4. Temporal correlation in classification correctness in the two study areas, both normalised by total area

study area 1				study area 2			
correct	correct at t2			correct	correct at t2		
at t1	yes	no	sum	at t1	yes	no	sum
yes	0.9002	0.0082	0.9084	yes	0.7200	0.1200	0.8400
no	0.0582	0.0334	0.0916	no	0.1200	0.0400	0.1600
sum	0.9584	0.0416	1.0000	sum	0.8400	0.1600	1.0000

4.3.2 Land cover change estimates

Post classification comparison and actual transitions

For the Soest study area the PCC and actual transition matrix are shown in tables 5 and 6. The largest error is the 259 ha. underestimation of the number of Urban & Infra to Urban & Infra transitions (1094 ha. vs. 1352 ha.). Analysis of the change detection error matrix (table 7) reveals that of the 259 ha., a total of 102.8 ha. was due to overestimation of the area Agriculture at the expense of Urban & Infra in 1991, and that 102.3 ha. was due to overestimation of the area Agriculture at the expense of Urban & Infra in both 1991 and 1995. Point-wise and region-wise land cover changes are 7.4% and 7.8% according to the PCC transition matrix, while the actual changes are only 1.4% and 1.7% (table 9, first two lines).

For study area 2 the change detection error matrix is shown in table 8. The PCC transition matrix and the actual transition matrix can be derived from this matrix, for example the number of AU transitions according to the test data is 2 cells (row total) and is actually 4 cells (column total). Point-wise and region-wise land cover changes are (table 10, first two lines): 9 cells (36%) and 14 cells (56%) according to the test data sets versus 12 cells (48%) and 18 cells (72%) according to the actual data sets. Thus change is overestimated in study area 1 ($PC_{PCC} > PC_{actual}$) and underestimated in study area 2 ($PC_{PCC} < PC_{actual}$).

Table 5. PCC transition matrix of study area 1, with changes in ha.

	Test land cover in 1995 (ha.)				
Test land cover in 1991	Urban & Infra	Forest & Nature	Agriculture	Water	Total
Urban & Infra	1094	26	22	0	1141
Forest & Nature	39	1467	9	0	1516
Agriculture	139	13	550	1	703
Water	0	0	0	27	28
Total	1272	1506	580	29	3387

Table 6. Actual transition matrix for study area 1, with changes in ha.

	Actual land cover in 1995 (ha.)				
Actual land cover in 1991	Urban & Infra	Forest & Nature	Agriculture	Water	Total
Urban & Infra	1352	7	1	0	1361
Forest & Nature	7	1501	1	0	1509
Agriculture	29	0	459	1	489
Water	0	0	0	28	28
Total	1388	1508	462	29	3387

Table 7. Study area 1 change detection error matrix

Classified	Actual land cover changes (ha.)																
changes	UU	FF	AA	WW	UF	UA	UW	FU	FA	FW	AU	AF	AW	WU	WF	WA	Total
UU	1087.0	2.5	0.1			0.1		2.6			1.5						1093.8
FF	0.5	1460.6			6.0												1467.1
AA	102.8	0.5	444.2			0.5		0.6	0.2		0.5					0.5	549.8
WW				27.1						0.1							27.2
UF		25.1			0.8												25.9
UA	15.2		4.9			0.7		0.1	0.6								21.5
UW				0.2													0.2
FU	38.2	0.1						0.8			0.2						39.3
FA	5.6	0.2	0.7					2.2	0.2								8.9
FW				0.2													0.2
AU	102.3		9.5					0.6			26.6						139.0
AF	0.5	11.9										0.1					12.5
AW				0.2									1.0				1.2
WU	0.2																0.2
WF		0.2															0.2
WA																	0.0
Total	1352.3	1501.1	459.4	27.7	6.8	1.3	0.0	6.9	1.0	0.1	28.8	0.1	1.0	0.0	0.0	0.5	3387.0

U = Urban & Infra; F = Forest & Nature; A = Agriculture; W = Water; UU = transition from Urban & Infra at t1 to Urban & Infra at t2

Table 8. Study area 2 change detection error matrix

Classified	Actual land cover changes (# cells)																
changes	AA	BB	UU	WW	AB	AU	AW	BA	BU	BW	UA	UB	UW	WA	WB	WU	Total
AA	2					1											3
BB		2															2
UU			4			1			1								6
WW				4											1		5
AB																	0
AU						2											2
AW																	0
BA																	0
BU		1							1								2
BW																	0
UA																	0
UB																	0
UW																	0
WA							1							1			2
WB																	0
WU									1							2	3
Total	2	3	4	4	0	4	1	0	3	0	0	0	0	1	1	2	25

Posterior transition matrices

Tables 9 and 10 show, for the two case-studies, land cover change estimates derived from the actual transition matrix (first line), the PCC transition matrix (second line) and the posterior transition matrices (third and following lines). Application of method ‘x’ yielded better estimates of actual change than method ‘y’ when the following three conditions apply:

1. $| PC_{actual} - PC_x | < | PC_{actual} - PC_y |$, i.e. PC_x is closer to PC_{actual} than PC_y is to PC_{actual};
2. $| RC_{actual} - RC_x | < | RC_{actual} - RC_y |$;
3. $SAD_x < SAD_y$;

In all cases land cover change estimates by methods 1 to 4 were closer to PC_{actual}, RC_{actual} and SAD_{actual} than were the estimates by post-classification comparison (PCC). The only exception was method 2 in the Soest study area, where $| PC_{actual} - PC_{method\ 2} | > | PC_{actual} - PC_{PCC} |$ and $SAD_{method\ 2} > SAD_{PCC}$. For the Soest study area a comparison with the model MonGIS was possible. In general the PC, RC and SAD values calculated with the methods 1 to 4 were closer to PC_{actual}, RC_{actual} and SAD_{actual} than were the PC, RC and SAD values calculated with model MonGIS.

In study area 1, $SAD_{method\ 1} < SAD_{method\ 2}$ and $SAD_{method\ 1} < SAD_{method\ 3}$; in study area 2 these relationships were the other way round: $SAD_{method\ 1} > SAD_{method\ 2}$ and $SAD_{method\ 1} > SAD_{method\ 3}$. The use of method 1 is only recommended under two specific conditions. That is, when classification accuracy at t2 is almost 100% and when at t1 the area of one or more of the land cover classes is systematically overestimated or underestimated at t1. These specific conditions applied to study area 1, where classification accuracy at t2 was 95.8% and where at t1 the area Urban & Infra was underestimated and the area Agriculture overestimated (table 2, row total vs. column total).

The two study areas differ in temporal correlation in classification accuracy. In study area 1 this correlation was stronger (0.93, table 4). There, the change estimates by methods 3 and 4 were closer to actual change than were the change estimates by method 2. In study area 2, where the correlation was weaker (0.76, table 4), the differences between change estimates by methods 2, 3 and 4 were smaller. In study area 1, change estimates by method 4 were closer to actual change than were the estimates by method 3: $| PC_{act} - PC_{method\ 4} | < | PC_{act} - PC_{method\ 3} |$, $| RC_{act} - RC_{method\ 4} | < | RC_{act} - RC_{method\ 3} |$ and $SAD_{method\ 4} < SAD_{method\ 3}$. Consider what would have happened if classification accuracy at t2 had been 100%. In that case, the conditioned confusion probabilities $Pr(LC_{act,t2} \mid LC_{data,t2}\ ,\ LC_{data,t1},\ LC_{act,t1})$ would be exactly the same as the unconditioned probabilities $Pr(LC_{act,t2} \mid LC_{data,t2})$ and as a result there would be no difference between change estimates by methods 2 and 3. As classification accuracy at t2 decreases conditioning of confusion probabilities at t2 on t1 can have a greater impact on change estimation. Because in study area 1 classification accuracy at t1 was lower than at t2 (90.5% vs. 95.8%), conditioning on t2 (method 4) had a greater impact on the posterior change estimates than conditioning on t1 (method 3).

Table 9. Study area 1: land cover change estimates by different methods

	PC		RC		SAD	
Method	ha.	%	ha.	%	ha.	%
Actual *	47	1.4	56	1.7	0	0.0
PCC °	249	7.4	264	7.8	587	17.3
Method 1 +	212	6.3	160	4.7	360	10.6
Method 2 +	519	15.3	56	1.7	945	27.9
Method 3 +	228	6.7	163	4.8	362	10.7
Method 4 +	54	1.6	48	1.4	18	0.5
MonGIS +	232	6.8	248	7.2	576	17.0

PC = Point-wise change, RC = Region-wise change, SAD = Sum of Absolute Differences; * = based on exhaustive reference data set, ° = Post Classification Comparison, + = error-reduced estimates

Table 10. Study area 2: land cover change estimates by different methods

	PC		RC		SAD	
Method	#cells	%	#cells	%	#cells	%
Actual *	12.0	48.0	18.0	72.0	0.0	0.0
PCC °	9.0	36.0	14.0	56.0	12.0	48.0
Method 1 +	9.8	39.2	16.4	65.6	6.4	25.6
Method 2 +	13.5	54.1	18.0	72.0	3.8	15.2
Method 3 +	12.7	50.8	18.0	72.1	4.1	16.4
Method 4 +	13.3	53.0	18.0	72.1	3.7	14.8

PC = Point-wise change, RC = Region-wise change, SAD = Sum of Absolute Differences; * = based on exhaustive reference data set, ° = Post Classification Comparison, + = error-reduced estimates

Which method to use?

The previous section showed for two case-study areas that land cover change estimates could be improved with the 4 methods proposed in this study. Improvement was quantified by comparing PC, RC and SAD values of each method with their true values, quantified from 2 exhaustive reference data sets. Method 1 is easiest to apply, but is only applicable under limited conditions. These limited conditions are that classification accuracy at t2 is high and that at t1 the relative areas of land cover classes are systematically overestimated or underestimated (comparison between row and column totals of the error matrix of t1). Method 2 is applicable to a broader range of conditions than method 1, accounting for confusion probabilities at both dates. Method 2 does not account for temporal correlation in confusion probabilities. If this correlation is strong, then if possible one should use either method 3 or 4. How strong this correlation should be before the use of these methods becomes advantageous is still a matter of research. Practically, one should be aware of the fact that data producers rarely report information on temporal correlation in classification accuracy, and they rarely report confusion probabilities for one date conditioned on another date. If such information is available, a choice is to be made between the use of either method 3 or method 4. The choice depends on the classification accuracy at the two dates. If it is lowest at t1, then use method 4, if it is lowest at t2 then use method 3.

4.4 Conclusions

For most users improvement of land cover change estimates by selection of the best change detection method or by improvement of data quality is often not an option. These users buy classified land cover data sets of two dates, often accompanied by quality descriptions such as the error matrix. This study showed that the error matrix can be used for more purposes than only describing the quality of a single date data set: it can also be used to improve land cover change estimates. The methods proposed in this study can be used directly by users. For two case-studies it was shown that these methods provided better land cover change estimates than the standard change detection method (post classification comparison). The methods provided better estimates than an existing model (MonGIS, by de Zeeuw et al., 1999) and are more generally applicable.

Quality descriptions of the fitness of national land cover databases for monitoring are generally lacking (de Zeeuw and Hazeu, 2001; Lunetta 2002). Producers of vector data sets, for example topographical maps, tend to focus on quantifying positional accuracy and rarely report error matrices. Error matrices cannot be derived from standard measures of positional accuracy (e.g. Federal Geographic Data Committee, 1998b), since some but not all misregistrations result in misclassifications. If vector data sets are to be used for monitoring purposes, then to allow for application of the methods proposed in this study, the error matrices associated with these data sets should also be reported. This study showed that estimates of land cover change between two dates could be further improved by accounting for temporal correlation in classification accuracy. In two of the methods proposed the probabilities of confusion between two land cover classes at one date were conditioned on the confusion between land cover classes that occurred at the other date. To my knowledge, producers of land cover data sets do not report such conditional confusion probabilities. They can be quantified when reference data sets have been collected from the same study area at two dates.

In both case-studies presented in this chapter, I used exhaustive reference data sets to quantify error matrices. In practice, error matrices supplied by data producers will be derived from a reference data set that is only a sample of the total area mapped. Further research should increase understanding of the effect of sample design on the reduction of overestimation or underestimation of land cover change that can be achieved by application of the methods proposed in this study.

5. A variance and covariance equation for area estimates[1]

[1] Based on: van Oort, P.A.J., Stein, A., Bregt, A.K., de Bruin, S. and Kuipers, J., 2005. A variance and covariance equation for area estimates with a GIS. Forest Science 51(4): 347-356. Published by the Society of American Foresters, 5400 Grosvenor Lane, Bethesda, MD 20814-2198. © Society of American Foresters.

5. A variance and covariance equation for area estimates

Abstract

Geographical Information Systems (GIS) automatically calculate the area of a polygon from the coordinate values of the vertices that describe its boundary. Uncertainty in these coordinate values results in uncertainty in area estimates. Earlier papers showed how area uncertainty of individual polygons could be calculated. In this chapter we propose a covariance equation to quantify the impact of uncertainty in the position of a vertex on uncertainty in the area of all polygons sharing this vertex. The approach is based on the assumption of absence of spatial correlation in positional errors, but includes variation in the positional accuracy. The equations were implemented and applied to a case study area, described by a set of 97 adjacent polygons with different per unit area utility values (€/ha). We were then able to calculate how uncertainty in the coordinate values propagated into uncertainty in the utility value of the complete set of polygons.

5.1 Introduction

Geographical Information Systems (GIS) are increasingly used in forest science (Heit and Shortreid 1991, Husch et al. 2003). A standard GIS operation is the calculation of polygon area from the coordinates of the vertices that describe the polygons' boundaries. Uncertainty in the coordinate values results in uncertainty in area estimates, which in turn results in uncertainty in the estimated utility value of the polygons. So far, research has been limited to the calculation of uncertainty in the areas of individual polygons (Chrisman and Yandell 1988, Griffith 1989, Bondesson et al. 1998, Magnussen 1996, Næsset 1999). The single exception appears to be the work by Prisley et al. (1989). Their equation required the position of polygon centroids to be known whereas this chapter presents an equation which is independent of the polygon centroid.

In this chapter we propose a method to quantify the implications of uncertainty in area estimates on the uncertainty in the utility value of a set Ω of polygons p_q, $q = 1, \ldots, N$. Each polygon has a, possibly different, utility value per unit area v_q. One may think of a set of forest stands at different growth stages, thus with different monetary value, or of a set of polygons to which subsidies for nature conservation are assigned depending on the type of nature present. The utility value of a polygon, $U(p_q)$, is the product of the value per unit area, v_q, and polygon area, $A(p_q)$ (eq. 1). The utility value of the set of polygons, $U(\Omega)$, is the sum of values $U(p_q)$ of the N individual polygons (eq. 2). For several reasons, knowledge of the uncertainty in $U(\Omega)$ may be of interest:

- When two parties have a different estimate of $U(\Omega)$, then knowledge of the uncertainty in $U(\Omega)$ may help in decide on what to do: accept the difference as being caused by uncertainty in the data, or start to investigate the cause of the difference in estimated $U(\Omega)$ by considering differences in data and methods used by the two parties.

- Knowing in advance how much uncertainty in $U(\Omega)$ is reduced as a result of measuring $A(p_q)$ with greater accuracy may be of interest to the manager who has to decide whether measurement of $A(p_q)$ with greater accuracy is worth the costs.

Implications of uncertainty in v_q on uncertainty in $U(\Omega)$ have been frequently addressed (Buongiorno and Gilless 2003, Husch et al. 2003), implications of uncertainty in $A(p_q)$ scarcely. Næsset (1999), in a study of uncertainty in timber volume estimates, studied uncertainty in both v_q and $A(p_q)$. His study was limited to uncertainty in individual polygons; our study assesses uncertainty in a set of adjacent polygons. In this chapter the focus is entirely on implications of uncertainty in $A(p_q)$ on uncertainty in $U(\Omega)$, $\text{var}(U(\Omega))$. The assumption is made that no uncertainty exists in v_q. The case study in the second part of the chapter provides an example of a case where this assumption is valid. The uncertainty in the area of a single polygon is expressed by $\text{var}(A(p_q))$. The covariance $\text{cov}(A(p_q),A(p_r))$, $r \neq q$ measures the relation between $A(p_q)$ and $A(p_r)$. The uncertainty in $U(\Omega)$, as expressed by $\text{var}(U(\Omega))$, depends on both $\text{var}(A(p_q))$ and $\text{cov}(A(p_q),A(p_r))$ (eq. 4):

$$U(p_q) = v_q \cdot A(p_q) \tag{1}$$

$$U(\Omega) = \sum_{q=1}^{N} U(p_q) \tag{2}$$

$$\text{var}(U(p_q)) = v_q^2 \cdot \text{var}(A(p_q)) \tag{3}$$

$$\text{var}(U(\Omega)) = \sum_{q=1}^{N} v_q^2 \cdot \text{var}(A(p_q)) + 2 \cdot \sum_{q=1}^{N} \sum_{r>q}^{N} v_q \cdot v_r \cdot \text{cov}(A(p_q), A(p_r)) \tag{4}$$

5.1.1 Aim and outline

The aim of this study is to develop a model to calculate $\text{var}(U(\Omega))$. Ingredients for this model are the equations for $\text{var}(A(p_q))$ and $\text{cov}(A(p_q),A(p_r))$. Our derivation of the equation for $\text{cov}(A(p_q),A(p_r))$ is based on the work by Chrisman and Yandell (1988). It differs from the existing equation for $\text{cov}(A(p_q),A(p_r))$ by Prisley et al. (1989) in that it is independent of polygon centroid coordinates. The following section presents an overview of literature on methods to quantify $\text{var}(A(p_q))$ and $\text{cov}(A(p_q),A(p_r))$. The chapter consists of two parts. In part one of the chapter the equations in the model are presented and a simple example serves as an illustration. The equations were derived under a set of simplifying assumptions on error structure which are discussed in this part of the chapter. Part two of the chapter provides a practical illustration of the use of the model in a case study by the National Forest Service of the Netherlands (SBB). In this case study, we consider the case of the manager, who has a limited budget to increase the positional accuracy of the data set of the case study. The model is used to quantify the reduction in $\text{var}(U(\Omega))$, as a result of increasing the positional accuracy of the vertices causing the uncertainty in area estimates.

5.1.2 Uncertainty in area estimates

A polygon is described by a set of vertices connected by arcs. Each vertex has a set of coordinates that describe its position, which in turn allow calculation of the area of the polygon. Uncertainty in polygon area may be caused by uncertainty in the coordinate values of the recorded vertices and by incompleteness of the vertices (i.e. not all vertices are recorded). Incompleteness of vertices may be unavoidable due to fuzziness of the boundary (Magnussen 1996, de Groeve and Lowell 2001). There is no crisp answer to the question on whether a boundary is crisp or fuzzy: it depends on the level of detail at which the boundary is studied. As one zooms in on an apparently crisp boundary more detail will become visible and one may conclude that it is in fact not crisp but fuzzy. Magnussen (1996) included effects of fuzziness explicitly in his area variance equation. Another approach of including fuzziness in the model is to use one error term σ_x that includes both the effect of measurement error $\sigma_{x,m}$ and the effect of fuzziness or generalisation $\sigma_{x,g}$. Assuming independence between measurement and generalisation, the error term σ_x is calculated as $\sigma_x = \sqrt{\sigma_{x,m}^2 + \sigma_{x,g}^2}$. For example see Blachut et al. (1979: 202) and Polman and Salzman (1996: p. 50 and p. 402; in Dutch). In this chapter, we proceed with the assumption that var($A(p_q)$) and cov($A(p_q)$,$A(p_r)$) can be calculated solely from the error terms σ_x and σ_y assigned to the vertices delineating polygons.

Two approaches exist to quantify uncertainty in area estimates: (1) by comparison with another data set (e.g. Gross and Adler 1996, van Neil and McVicar 2001) or (2) by error propagation analysis. The advantage of the first approach is that it does not make any assumptions on the sources of uncertainty. The disadvantage of the approach is that the estimation of the uncertainty in areas of polygons not included in the sample is problematic because the approach does not directly address the sources of uncertainty. If the major source of uncertainty in area estimates is uncertainty in coordinate values, then var($A(p_q)$) and cov($A(p_q)$,$A(p_r)$) can be calculated directly from the uncertainty in these coordinate values using error propagation analysis. The two most commonly applied error propagation analysis methods are Monte Carlo Simulation (MCS) and derivation from the area equation based on probability theory. In MCS the positions of vertices are repetitively permutated, areas recalculated and the variance is calculated from the total number of repetitions. MCS is computationally intensive due to the number of repetitions needed (Næsset, 1999), but also because the permutations may cause topological inconsistencies that have to be eliminated before areas can be recalculated (Hunter et al. 1996). Variance and covariance equations based on derivation from the area equation do not require repetitions and do not cause topological inconsistencies. Prisley et al. (1989) derived variance and covariance equations from an area equation that required the coordinate values of each polygon's centroid. Other studies used the trapezoid rule (Burrough and McDonnel 1998: 62-63, Husch et al. 2003: 61) as the area equation, which is independent from the position of the centroid. Equations of var($A(p_q)$) have been derived from this rule under a range of model assumptions (Chrisman and Yandell 1988, Griffith 1989), but a derivation of the equation of cov($A(p_q)$,$A(p_r)$) was never made. This chapter provides a derivation based on the trapezoid rule, yielding an equation of cov($A(p_q)$,$A(p_r)$) which is independent of the position of the polygon centroids.

5.2 Part 1: equations

5.2.1 Notation

At the highest aggregation level, the vector data structure in a GIS contains a set Ω of N polygons p_q, $\Omega = \{p_q, q = 1, \ldots, N\}$. Building blocks of these polygons are the vertices. Vertices are connected by arcs which together form a closed loop. The boundary of a polygon p_q is described by vertices with labels $s_q(i)$ and coordinate pairs $x_q(i)$,$y_q(i)$, $i = 1, \ldots, n_q$. The boundary of an adjacent polygon $r \neq q$ from the same set Ω is described by vertices with labels $s_r(j)$ and coordinate pairs $x_r(j)$,$y_r(j)$, $j = 1, \ldots, n_r$,. As polygons may share several vertices, we introduce the variable $\delta(s_q(i),s_r(j))$: $\delta(s_q(i),s_r(j)) = 1$ if vertex $s_q(i) = s_r(j)$ occurs in both polygon p_q and p_r, and $\delta(s_q(i),s_r(j)) = 0$ otherwise.

The polygon area is calculated from the coordinates of the vertices that describe the polygon's boundary. The trapezoid rule requires that the vertices are sorted in counter clockwise direction so that vertex i–1 comes just before and vertex i+1 just after vertex i. As a polygon is described by a closed loop of vertices, $s_q(0) \equiv s_q(n_q)$ and $s_q(n_q+1) \equiv s_q(1)$. The polygon area $A(p_q)$ is calculated as:

$$A(p_q) = \frac{1}{2} \cdot \sum_{i=1}^{n_q} X_q(i) \cdot \left(Y_q(i+1) - Y_q(i-1)\right) \tag{5}$$

The true coordinates $x_q(i)$ and $y_q(i)$ are rarely known. Instead we have recorded $X_q(i)$ and $Y_q(i)$ with unknown errors $\varepsilon(i)$ and $\eta(i)$:

$$\mathrm{E}(X_q(i)) = x_q(i) + \mathrm{E}(\varepsilon(i)) \qquad ; \mathrm{E}(Y_q(i)) = y_q(i) + \mathrm{E}(\eta(i))$$

The estimated polygon area $A(p_q)$ can now be decomposed into the unknown true area $a(p_q)$ and two error terms, $B(p_q)$ and $C(p_q)$:

$$A(p_q) = a(p_q) + B(p_q) + C(p_q) \tag{6}$$

with:

$$a(p_q) = \frac{1}{2} \cdot \sum_{i=1}^{n_q} x_q(i) \cdot \left(y_q(i+1) - y_q(i-1)\right)$$

$$B(p_q) = \frac{1}{2} \cdot \sum_{i=1}^{n_q} \varepsilon(i) \cdot (Y_q(i+1) - Y_q(i-1)) + \eta(i) \cdot (X_q(i+1) - X_q(i-1))$$

and

$$C(p_q) = \frac{1}{2} \cdot \sum_{i=1}^{n_q} \varepsilon(i) \cdot (\eta(i+1) - \eta(i-1))$$

5.2.2 Model assumptions

Derivation of the equations for var($A(p_q)$) and cov($A(p_q)$,$A(p_r)$) depends on the assumptions made on the structure of errors ε and η. In this chapter we assume that $\varepsilon(i)$ and $\eta(i)$ are distributed according to a continuous distribution with mean 0 and variances $\sigma^2_x(i)$ and $\sigma^2_y(i)$. We assume that the distribution of errors may be different in different vertices, that errors in the *x*- and *y*-direction of each individual vertex are uncorrelated, and that the errors between vertices are uncorrelated:

$$\begin{array}{lll}
\mathrm{E}(\varepsilon(i)) = 0 & ;\ \mathrm{E}(\eta(i)) = 0 & \\
\mathrm{E}(\varepsilon^2(i)) = \sigma^2{}_x(i) & ;\ \mathrm{E}(\eta^2(i)) = \sigma^2{}_y(i) & ;\ \mathrm{E}(\varepsilon(i)\cdot\eta(i)) = 0 \\
\mathrm{E}(\varepsilon(i)\cdot\varepsilon(k)) = 0\ \forall\ k \neq i & ;\ \mathrm{E}(\eta(i)\cdot\eta(k)) = 0\ \forall\ k \neq i & ;\ \mathrm{E}(\varepsilon(i)\cdot\eta(k)) = 0\ \forall\ k \neq i
\end{array}$$

In the absence of bias in $x_q(i)$ and $y_q(i)$, i.e. $\mathrm{E}(\varepsilon(i)) = 0$ and $\mathrm{E}(\eta(i)) = 0$, one finds that $\mathrm{E}(B(p_q)) = \mathrm{E}(C(p_q)) = 0$. As a result, $A(p_q)$ is an unbiased estimate of the true area $a(p_q)$. Bias may exist, however. If that is the case and the bias is known, then it can be removed by subtracting it from the coordinate values $X_q(i)$ and $Y_q(i)$. Variance and covariance equations can then be applied to the resulting unbiased coordinates. Næsset (1999) found bias in errors in vertices, related to the length of shadows cast perpendicular to the boundaries of patches of forest and found that this led to biased areas estimates.

In the equations proposed in this chapter, we relax the assumption that the position accuracy is the same for all vertices. In equations for $\mathrm{var}(A(p_q))$ based on the epsilon band concept, this relaxation is impossible, as the concept presumes a fixed width of the error-band describing the uncertainty in the position of an arc (Leung and Yan, 1998).

We assume that the errors in *x*- and *y*-direction in each individual vertex are independent. Chrisman and Yandell (1988) examined the Digital Line Graph data set of the U.S. Geological Survey and found a negligibly small correlation, $\rho = 0.026$. Bondesson et al. (1998) showed that such dependency occurs if a vertex's coordinates are measured using trigonometry (i.e. using the angle and distance to other vertices), but does not occur when measurements are made with GPS. Chrisman and Yandell (1988) and Griffith (1989) derived a variance equation that included dependence of errors in *x*- and *y*-direction.

The last assumption is that positional errors between different vertices are independent. Examples of such dependency are (figures 1a and 1b):

- Errors caused by shadow may be positively correlated, because vertices located on the same boundary have shadows casting into the same direction.
- If a road has a fixed width, then errors on either side of the road are positively correlated. As a result, uncertainty in the area of a polygon on one side of the road is correlated with uncertainty in the area of the polygon on the other side of the road (e.g. p_q and p_r in figure 1b). According to the equations proposed in this chapter, the two are independent, because they do not share a single vertex with $\delta(s_q(i),s_r(j)) = 1$

Griffith (1989) derived the variance equation for the case where errors were correlated between adjacent vertices. Keefer et al. (1991) quantified correlation between adjacent vertices using time-series analysis. With the fitted model of correlation between vertices, they then calculated variance in polygon area using Monte Carlo Simulation. Naesset (1999) also measured serial correlation between adjacent nodes, but did not include it explicitly in his computations. Implicitly it was included, because random errors were added to segments, i.e. equal random errors to each of the vertices defining the segment. To our knowledge, correlation between errors of non-adjacent vertices (figure 1b) has not been addressed in literature. The difficulty in modelling this correlation is due to the problem of identifying which vertices have correlated errors, as these cannot simply be identified on the basis of adjacency (as in figure 1a).

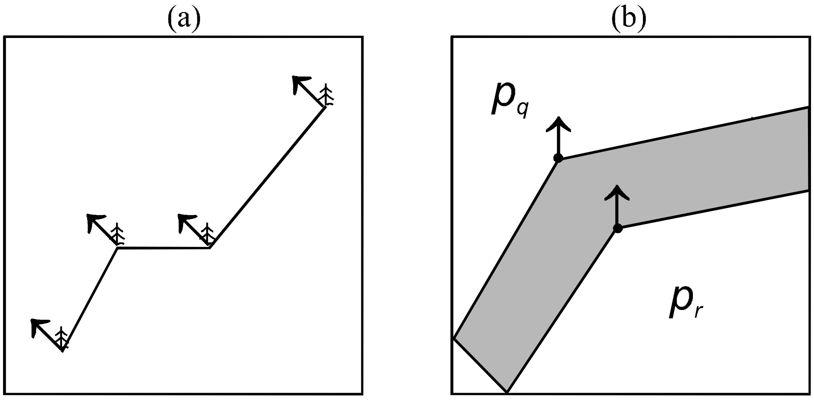


Figure 1. Correlation in errors between vertices; (a) adjacent vertices, (b) non-adjacent vertices. Arrows indicate errors, which are correlated in size and direction.

The review processes of the articles by van Oort et al. (2005) and Bogaert et al. (2005) coincided and for this reason the two could not refer to each other. Reference to the work by Bogaert et al. (2005) is now added in this chapter. Bogaert et al. measured correlation in errors of GPS measurements. They found no correlation in errors in *x*- and *y*-direction for individual vertices. They did however find correlation between different vertices, depending on the time span between GPS measurements. Serial correlation was absent at time spans larger than 30 seconds.

5.2.3 Variance and covariance equations

The variance and covariance in polygon areas are calculated as a function of the variances in the vertex coordinate values. The variance equation (7) is defined as:

$$\mathrm{var}(A(p_q)) = \mathrm{E}(A^2(p_q)) - \{\mathrm{E}(A(p_q))\}^2 \qquad (7)$$

Substitution of $A(p_q)$ with $a(p_q) + B(p_q) + C(p_q)$ and elimination of terms with expectation zero yields:

$$\mathrm{var}(A(p_q)) = \mathrm{E}(B^2(p_q)) + \mathrm{E}(C^2(p_q)) + 2\cdot\mathrm{E}(B(p_q)\cdot C(p_q)) \qquad (8)$$

where:

$$\mathrm{E}(B^2(p_q)) = \frac{1}{4}\cdot\mathrm{E}\left(\sum_{i=1}^{n_q}\left[\varepsilon^2(i)\cdot\left(Y_q(i+1)-Y_q(i-1)\right)^2+\eta^2(i)\cdot\left(X_q(i+1)-X_q(i-1)\right)^2\right]\right)$$

$$\mathrm{E}(C^2(p_q)) = \frac{1}{4}\cdot\mathrm{E}\left(\sum_{i=1}^{n_q}\left[\varepsilon^2(i)\cdot\left(\eta(i+1)-\eta(i-1)\right)^2\right]\right)$$

and

$$\mathrm{E}(B(p_q)\cdot C(p_q))=\frac{1}{4}\cdot\mathrm{E}\begin{pmatrix}\sum_{i=1}^{n_q}\left[\varepsilon^2(i)\cdot(\eta(i+1))-\eta(i-1))\cdot\left(Y_q(i+1)-Y_q(i-1)\right)\right]+\\ \sum_{i=1}^{n_q}\left[\eta^2(i)\cdot(\varepsilon(i+1))-\varepsilon(i-1))\cdot\left(X_q(i+1)-X_q(i-1)\right)\right]\end{pmatrix}=0$$

Note that $a(p_q)$ is a constant and hence does not occur in an expression of the variance. Substitution of $\mathrm{E}(\varepsilon^2(i)) = \sigma^2_x(i)$, $\mathrm{E}(\eta^2(i)) = \sigma^2_y(i)$, $\mathrm{E}(X(i)) = x(i)$ and $\mathrm{E}(Y(i)) = y(i)$, and elimination of terms with expectation zero yields the variance equation:

$$\begin{aligned}\mathrm{var}(A(p_q))&=\frac{1}{4}\cdot\sum_{i=1}^{n_q}\left[\sigma_x^2(i)\cdot\left(y_q(i+1)-y_q(i-1)\right)^2+\sigma_y^2(i)\cdot\left(x_q(i+1)-x_q(i-1)\right)^2\right]\\&+\frac{1}{4}\cdot\sum_{i=1}^{n_q}\left[\sigma_x^2(i)\cdot\left(\sigma_y^2(i+1)+\sigma_y^2(i-1)\right)\right]\end{aligned}\qquad(9)$$

Although our derivation is similar to that in Chrisman and Yandell (1988), the final form of (9) differs in two aspects:

1. their equation allowed for the errors *x*- and *y*-direction to be correlated, whereas we assumed absence of correlation;
2. their equation did not allow for differences in the positional accuracy of vertices, whereas (9) allows for such differences to exist.

The covariance equation is derived in a similar way as the variance equation, starting from equation 10:

$$\mathrm{cov}(A(p_q),A(p_r)) = \mathrm{E}(A(p_q)\cdot A(p_r)) - \mathrm{E}(A(p_q))\cdot\mathrm{E}(A(p_q))\qquad(10)$$

Substitution of $A(p_q) = a(p_q) + B(p_q) + C(p_q)$ and $A(p_r) = a(p_r) + B(p_r) + C(p_r)$ and elimination of terms with zero expectation yields:

$$\mathrm{cov}(A(p_q),A(p_r)) = \mathrm{E}(B(p_q)\cdot B(p_r)) + \mathrm{E}(C(p_q)\cdot C(p_r))\qquad(11)$$

where

$$\mathrm{E}(B(p_q),B(p_r))=\frac{1}{4}\cdot\mathrm{E}\left(\sum\begin{bmatrix}\varepsilon(i)\cdot\varepsilon(j)\cdot\left(Y_q(i+1)-Y_q(i-1)\right)\left(Y_r(j+1)-Y_r(j-1)\right)+\\ \eta(i)\cdot\eta(j)\cdot\left(X_q(i+1)-X_q(i-1)\right)\left(X_r(j+1)-X_r(j-1)\right)\end{bmatrix}\right)$$

$$\mathrm{E}(C(p_q)\cdot C(p_r))=\frac{1}{4}\cdot\mathrm{E}\left(\sum\left[\varepsilon(i)\cdot\varepsilon(j)\cdot(\eta(i+1)-\eta(i-1))\cdot(\eta(j+1)-\eta(j-1))\right]\right)$$

and summation is carried out over the vertices shared by both polygons.

Substituting $\sigma^2_x(i)$, $\sigma^2_y(i)$, $x_q(i)$, $y_q(i)$, $x_r(j)$ and $y_r(j)$ into $\mathrm{E}(B(p_q)\cdot B(p_r))$ and $\mathrm{E}(C(p_q)\cdot C(p_r))$ we obtain the covariance equation (12). The binary $\delta(s_q(i),s_r(j))$ is used to identify where substitution is possible. For example if $\delta(s_q(i),s_r(j)) = 1$ then $\mathrm{E}(\varepsilon(i)\cdot\varepsilon(j)) = \sigma^2_x(i)$, else $\mathrm{E}(\varepsilon(i)\cdot\varepsilon(j)) = 0$. If $\delta(s_q(i),s_r(j)) = 1$ and $\delta(s_q(i+1),s_r(j-1)) = 1$ then $\mathrm{E}(\varepsilon(i)\cdot\varepsilon(j)\cdot\eta(i+1)\cdot\eta(j-1)) = \sigma^2_x(i)\cdot\sigma^2_y(i+1)$ in the term $\mathrm{E}(C(p_q)\cdot C(p_r))$. The covariance equation under these conditions equals:

$$\operatorname{cov}(A(p_q),A(p_r)) =$$

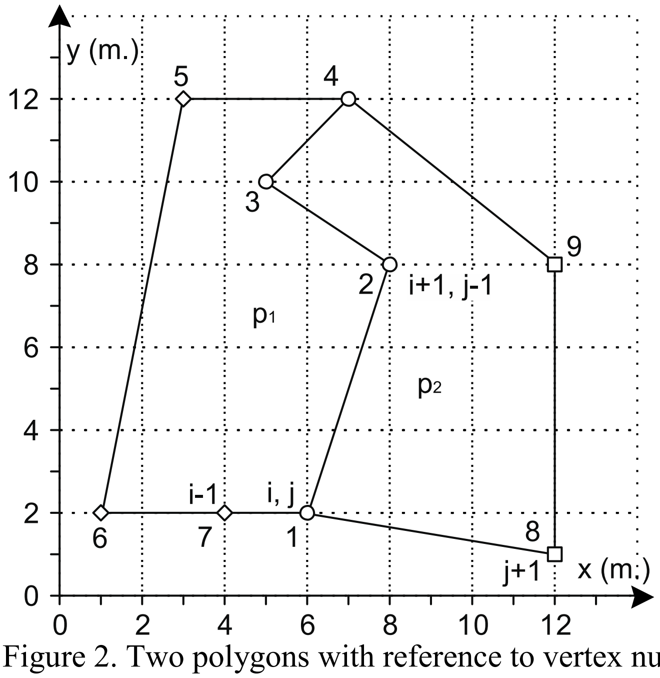
$$\frac{1}{8}\cdot\sum_{i=1}^{n_q}\sum_{j=1}^{n_r}\delta(s_q(i),s_r(j))\cdot\begin{bmatrix}\sigma_x^2(i)\cdot\left(y_q(i+1)-y_q(i-1)\right)\cdot\left(y_r(j+1)-y_r(j-1)\right)+\\ \sigma_y^2(i)\cdot\left(x_q(i+1)-x_q(i-1)\right)\cdot\left(x_r(j+1)-x_r(j-1)\right)\end{bmatrix} \qquad (12)$$

$$-\frac{1}{8}\cdot\sum_{i=1}^{n_q}\sum_{j=1}^{n_r}\delta(s_q(i),s_r(j))\cdot\sigma_x^2(i)\cdot\begin{bmatrix}\delta(s_q(i+1),s_r(j-1))\cdot\sigma_y^2(i+1)+\\ \delta(s_q(i-1),s_r(j+1))\cdot\sigma_y^2(i-1)\end{bmatrix}$$

Note that in this equation summation occurs twice over each vertex $\delta(s_q(i),s_r(j)) = 1$, so the result is divided by 2, resulting in the (1/8) instead of (1/4).

5.2.4 Illustration

The equations presented in the previous section were implemented in a script in the programming language AML of the GIS software package ArcInfo by ESRI. This section serves as a simple illustration of the equations presented. Figure 2 shows two adjacent polygons $p_{q=1}$ and $p_{r=2}$, with reference to the vertex numbers as stored in the GIS. See the figure for notation. The axes show the coordinate values in x- and y-direction.



Vertex label numbers in the figure as stored in the GIS.

Notation in equations:
For p_1: let $s_1(i-1)=7$, $s_1(i)=1$ and $s_1(i+1)=2$
For p_2: let $s_2(j-1)=2$, $s_2(j)=1$ and $s_2(j+1)=8$
Then for the binary δ:
$\delta(s_1(i-1),s_2(j+1))=\delta(7,8)=0$
$\delta(s_1(i),s_2(j))=\delta(1,1)=1$
$\delta(s_1(i+1),s_2(j-1))=\delta(2,2)=1$

Figure 2. Two polygons with reference to vertex numbers.

Table 1 lists the contribution of each vertex to var($A(p_1)$), var($A(p_2)$) and cov($A(p_1)$,$A(p_2)$), with elaborate notation for vertex $s_1(i) = s_2(j) = 1$. Note that for vertex 1 the binary $\delta(s_1(i-1),s_2(j+1))$ equals 0 and that for vertex 4 the binary $\delta(s_1(i+1),s_2(j-1))$ equals 0. The contribution of vertex 2 to var($A(p_1)$) is equal to the contribution of vertex 2 to var($A(p_2)$), both of which are equal to the contribution of vertex 2 to $-1\cdot$cov($A(p_1)$,$A(p_2)$). If the per unit area values of the polygons are v_1 and $v_2 = v_1 + d$, then the contribution of uncertainty in the position of vertex 2 to uncertainty in the utility value $U(\Omega)$ is (eq. 4): $\operatorname{var}(U(\Omega)) = v_1^2 \cdot \operatorname{var}(A(p_1)) + v_2^2 \cdot \operatorname{var}(A(p_2)) + 2\cdot v_1\cdot v_2\cdot\operatorname{cov}(A(p_1),A(p_2)) = d^2\cdot\operatorname{var}(A(p_1))$, where var($A(p_1)$), var($A(p_2)$) and cov($A(p_1)$,$A(p_2)$) refer to the individual contribution of this vertex. We notice in passing that this is the contribution of the uncertainty in the polygons for that single vertex and is therefore different from var($A(p_1)$) as calculated with 9, which expresses contribution from all vertices delineating polygon 1. This also applies to var($A(p_2)$), cov($A(p_1)$,$A(p_2)$) and the notation in table 1. The result d^2·var($A(p_1)$) implies that the larger the difference between v_1 and v_2, the greater is the vertex's contribution to var($U(\Omega)$). If the polygons p_1 and p_2 have the same per unit area value ($d = 0$), then uncertainty in the position of vertices 2 and 3 does not contribute to var($U(\Omega)$).

Table 1. Contribution of uncertainty in the position the vertices in figure 2 to var($A(p_1)$), var($A(p_2)$) and cov($A(p_1)$,$A(p_2)$). The table lists only the non-zero contributions.

vertex $s_1(i) = s_2(j) = 1$	
var($A(p_1)$)	$= 0.25\cdot[\sigma^2_x(i)\cdot(8–2)^2 + \sigma^2_y(i)\cdot(8–4)^2] + 0.25\cdot\sigma^2_x(i)\cdot[\sigma^2_y(i+1) + \sigma^2_y(i–1)]$ $= 9\cdot\sigma^2_x(i) + 4\cdot\sigma^2_y(i) + 0.25\cdot\sigma^2_x(i)\cdot[\sigma^2_y(i+1) + \sigma^2_y(i–1)]$
var($A(p_2)$)	$= 0.25\cdot[\sigma^2_x(j)\cdot(1–8)^2 + \sigma^2_y(j)\cdot(12–8)^2] + 0.25\cdot\sigma^2_x(j)\cdot[\sigma^2_y(j+1) + \sigma^2_y(j–1)]$ $= 12.25\cdot\sigma^2_x(i) + 4\cdot\sigma^2_y(i) + 0.25\cdot\sigma^2_x(i)\cdot[\sigma^2_y(i+1) + \sigma^2_y(i–1)]$
cov($A(p_1)$,$A(p_2)$)	$= 0.25\cdot1\cdot[\sigma^2_x(i)\cdot(8–2)\cdot(1–8) + \sigma^2_y(i)\cdot(8–4)\cdot(12–8)]$ – $0.25\cdot1\cdot\sigma^2_x(i)\cdot[1\cdot\sigma^2_y(i+1) + 0\cdot\sigma^2_y(i–1)]$ $= –10.5\cdot\sigma^2_x(i) + 4\cdot\sigma^2_y(i) – 0.25\cdot\sigma^2_x(i)\cdot\sigma^2_y(i+1)$
vertex $s_1(i) = s_2(j) = 2$	
var($A(p_1)$)	$= 16\cdot\sigma^2_x(i) + 0.25\cdot\sigma^2_y(i) + 0.25\cdot\sigma^2_x(i)\cdot[\sigma^2_y(i+1) + \sigma^2_y(i–1)]$
var($A(p_2)$)	$= 16\cdot\sigma^2_x(i) + 0.25\cdot\sigma^2_y(i) + 0.25\cdot\sigma^2_x(i)\cdot[\sigma^2_y(i+1) + \sigma^2_y(i–1)]$
cov($A(p_1)$,$A(p_2)$)	$= –16\cdot\sigma^2_x(i) – 0.25\cdot\sigma^2_y(i) – 0.25\cdot\sigma^2_x(i)\cdot[\sigma^2_y(i+1) + \sigma^2_y(i–1)]$
vertex $s_1(i) = s_2(j) = 3$	
var($A(p_1)$)	$= 4\cdot\sigma^2_x(i) + 0.25\cdot\sigma^2_y(i) + 0.25\cdot\sigma^2_x(i)\cdot[\sigma^2_y(i+1) + \sigma^2_y(i–1)]$
var($A(p_2)$)	$= 4\cdot\sigma^2_x(i) + 0.25\cdot\sigma^2_y(i) + 0.25\cdot\sigma^2_x(i)\cdot[\sigma^2_y(i+1) + \sigma^2_y(i–1)]$
cov($A(p_1)$,$A(p_2)$)	$= –4\cdot\sigma^2_x(i) – 0.25\cdot\sigma^2_y(i) – 0.25\cdot\sigma^2_x(i)\cdot[\sigma^2_y(i+1) + \sigma^2_y(i–1)]$
vertex $s_1(i) = s_2(j) = 4$	
var($A(p_1)$)	$= 1\cdot\sigma^2_x(i) + 1\cdot\sigma^2_y(i) + 0.25\cdot\sigma^2_x(i)\cdot[\sigma^2_y(i+1) + \sigma^2_y(i–1)]$
var($A(p_2)$)	$= 1\cdot\sigma^2_x(i) + 12.25\cdot\sigma^2_y(i) + 0.25\cdot\sigma^2_x(i)\cdot[\sigma^2_y(i+1) + \sigma^2_y(i–1)]$
cov($A(p_1)$,$A(p_2)$)	$= 1\cdot\sigma^2_x(i) + 3.5\cdot\sigma^2_y(i) – 0.25\cdot\sigma^2_x(i)\cdot\sigma^2_y(i–1)$
vertex $s_1(i) = 5$	
var($A(p_1)$)	$= 25\cdot\sigma^2_x(i) + 9\cdot\sigma^2_y(i) + 0.25\cdot\sigma^2_x(i)\cdot[\sigma^2_y(i+1) + \sigma^2_y(i–1)]$
vertex $s_1(i) = 6$	
var($A(p_1)$)	$= 25\cdot\sigma^2_x(i) + 0.25\cdot\sigma^2_y(i) + 0.25\cdot\sigma^2_x(i)\cdot[\sigma^2_y(i+1) + \sigma^2_y(i–1)]$
vertex $s_1(i) = 7$	
var($A(p_1)$)	$= 0\cdot\sigma^2_x(i) + 6.25\cdot\sigma^2_y(i) + 0.25\cdot\sigma^2_x(i)\cdot[\sigma^2_y(i+1) + \sigma^2_y(i–1)]$
vertex $s_2(j) = 8$	
var($A(p_2)$)	$= 9\cdot\sigma^2_x(j) + 9\cdot\sigma^2_y(j) + 0.25\cdot\sigma^2_x(j)\cdot[\sigma^2_y(j+1) + \sigma^2_y(j–1)]$
vertex $s_2(j) = 9$	
var($A(p_2)$)	$= 30.25\cdot\sigma^2_x(j) + 6.25\cdot\sigma^2_y(j) + 0.25\cdot\sigma^2_x(j)\cdot[\sigma^2_y(j+1) + \sigma^2_y(j–1)]$

As a numerical example consider positional accuracies equal to $\sigma_x(i) = \sigma_y(i) = 0.5$ m for all vertices. Then:

$A(p_1) = 47$ m^2 ; $A(p_2) = 46$ m^2 ;
var($A(p_1)$) = 25.47 m^4 ; var($A(p_2)$) = 26.31 m^4 ;
cov($A(p_1)$,$A(p_2)$) = –5.72 m^4

With utility values per unit area $v_1 = 10$ €/m^2 and $v_2 = 20$ €/m^2, the utility of the set $\Omega = \{p_1, p_2\}$ is: $U(\Omega) =$ € 1390 (eq. 2) and the uncertainty in $U(\Omega)$ is (eq. 4): $\mathrm{var}(U(\Omega)) =$ €^{2} 10784.38. Assuming normality, the 95% confidence interval on $U(\Omega)$ is $[1186.46,1593.54] = (1\pm0.146)\cdot U(\Omega)$ = within ±14.6% of $U(\Omega)$. Note that $\mathrm{cov}(A(p_1),A(p_2))$ is negative, because the expansion of one polygon at the expense of another leads to a negative covariance.

5.3 Part 2: case study

5.3.1 Introduction

Part 2 of the paper provides a practical illustration of the use of the equations derived in part 1. The National Forest Service of the Netherlands (SBB) receives subsidies from the ministry of agriculture, nature conservation and food quality for the management and conservation of nature. The subsidy per unit area depends on the type of nature. Figure 3 shows that 8 nature types are present in the study area, table 2 shows the total subsidy (in €) received for each. The total area shown in figure 3 is 225 ha, 184 ha of which is covered by nature types for which subsidies are received. The data set contains 1436 vertices, 1527 arcs and 97 polygons. The subsidy value of set of polygons in the study area is $U(\Omega) =$ € 27455.

For convenience of notation we will report in this part of the paper standard deviations instead of variances, thus σ instead of $\sigma^2_x(i)$ and $\sigma^2_y(i)$ and $\mathrm{std}(U(\Omega))$ instead of $\mathrm{var}(U(\Omega))$.

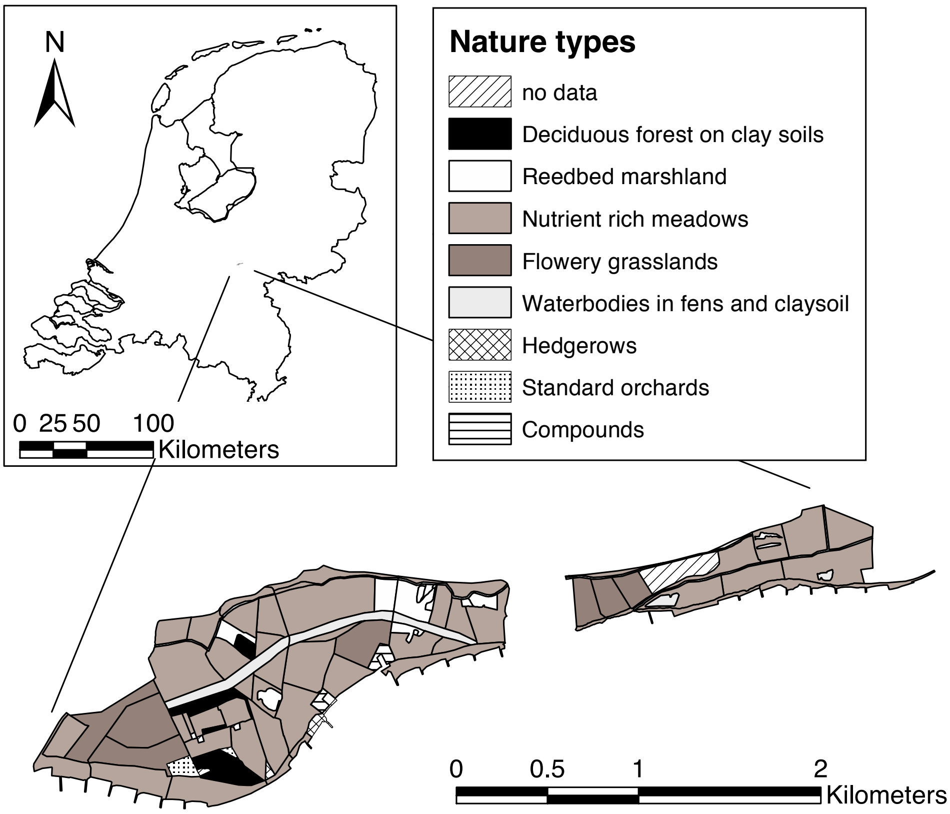


Figure 3. Case study area with types of nature.

Table 2. Nature in figure 3: description and subsidy value.

Description	Area (ha)	Subsidy (€)
Deciduous forest on clay soils	8.1	214
Reedbed marshland	11.1	5157
Nutrient rich meadows	147.5	20414
Flowery grasslands	33.7	0
Waterbodies in fens and claysoil	14.7	1170
Hedgerows	1.1	112
Standard orchards	1.6	387
Compounds	1.7	0

5.3.2 Data acquisition

Currently, data are acquired using the national topographical map, scale 1:10000, and lines drawn upon this map during field visits by SBB employees. The positional accuracy of the data is $\sigma = 2.0$ m for all vertices, according to SBB experts. Differences exist in accuracy between the vertices, but it is currently impossible to identify subsets of vertices with different accuracy. An accuracy of $\sigma = 2.0$ m is not unrealistic in the field of forestry, for example Naesset (1999) reports $\sigma = 2.4$ m, but most studies report larger errors: Magnussen (1996): $\sigma = 10$ m (from 1:15,000 photographs), Gross and Adler (1996): $\sigma = 6.5$ m (from a 1:7,000 photograph).

The data acquisition procedure may be changed, for example:

- Await improvements in the data acquisition procedure of the national topographical service and buy their updates;
- Switch from the national topographical map (scale 1:10000) to the use of municipality maps (scale 1:2000), provided that these are complete for the study area;
- Supply the employees with handheld GPS during their field visits;
- Revisit the study area with a map indicating the vertices which contribute most to the uncertainty in the total subsidy claimed from the area and record the position of these vertices with greater accuracy.

In terms of the terminology introduced in part 1 of the paper these changes may be redefined as:

1. Increase/decrease in the positional accuracy of all vertices;
2. Increase in the positional accuracy of a subset of vertices.

A subset can be created using simple selection criteria based on equations 4, 9 and 12. An alternative is to calculate each vertex's individual contribution to std($U(\Omega)$) and then select the vertices that contribute most to std($U(\Omega)$). The latter option is computationally more intensive, so the question is if similar reductions in std($U(\Omega)$) can be realized with simple selection criteria. In the following sections we will compare 5 subsets.

Aim

SBB would like to base decision making with regard to changes in the data acquisition procedure on the expected implications on the uncertainty in the total subsidy claimed from the ministry, std($U(\Omega)$). Our model is the whole set of equations presented in part 1 of the paper and implemented in a GIS. Our aim is to show how this model can support SBB in this aspect of their decision-making.

5.3.3 Methods

In two scenarios the effects of changing the positional accuracy on std($U(\Omega)$) were quantified. In scenario 1, the effect of increasing or decreasing the accuracy of all vertices within the range σ = 1.0 m to 10.0 m was quantified. In scenario 2, we increased the accuracy of five subsets of 500 vertices (35% of all) from σ = 2.0 m to 0.5 m. The subsets are described in table 3. The first four subsets were created using simple selection criteria, subset 2.5 was identified after calculation of the contribution of individual vertices to std($U(\Omega)$).

Table 3. Scenario 2: subsets of vertices for which the effect of increasing the positional accuracy on std($U(\Omega)$) is quantified.

Subset	Selection criterion	Motivation
2.1	vertices on located on the outer boundary of the study area	The contribution of 'inner' vertices to cov($A(p_q)$,$A(p_r)$) ≤ 0; the contribution of 'outer' vertices is 0 because these vertices have no adjacent polygon.
2.2	vertices delineating the polygons with the highest $v_q{\cdot}A(p_q)$	Generally if $A(p_q)$ is larger var($A(p_q)$) is also higher. $A(p_q)$ is weighted by v_q in this subset because the reduction in var($U(\Omega)$) depends on both var($A(p_q)$) and v_q (eq. 4).
2.3	vertices part of arcs with longest arc-length*	The contribution of uncertainty in the position of a vertex $s_q(i)$ to var($A(p_q)$) depends on the distance between the vertices $s_q(i+1)$ and $s_q(i–1)$ (eq. 9).
2.4	vertices part of arcs with highest product (arc-length)·$\lvert v_q–v_r \rvert$ †	For a vertex completely surrounded by polygons with the same utility value per unit area, i.e. $v_q = v_r$, uncertainty in the position of the vertex does not affect var($U(\Omega)$) at all.
2.5	vertices with highest contribution to var($U(\Omega)$)	Calculate directly the contribution of each vertex to var($U(\Omega)$), then select the 500 vertices with the highest contribution.

* the arc connecting vertices $s_q(i)$ and $s_q(i+1)$ has length $[(X_q(i) – X_q(i+1))^2 + (Y_q(i) – Y_q(i+1))^2]^{0.5}$

† where v_q and v_r are the per unit area utility values of the two polygons p_q and p_r on either side of an arc

5.3.4 Results

Contribution of individual vertices to std($U(\Omega)$)

Figure 4 shows a part of the data set, with per vertex the contribution to std($U(\Omega)$) in €. Inside the polygons are the per unit area values of the nature types in €/ha. The figure shows that the contribution is larger for vertices delineating 'valuable' polygons, vertices located on longer arcs and larger when the difference between the per-unit-area values of polygons on either side of an arc is larger. The contribution to std($U(\Omega)$) is exactly zero for vertices located on arcs which have on either side the same value per unit area. The contribution of the 1436 vertices to std($U(\Omega)$) ranged from € 0.0 (330 vertices) to € 7.91, with a mean of € 0.61 a median of € 0.36. The contribution to std($U(\Omega)$) of the vertices in subset 2.5 (table 3) was at least € 0.57.

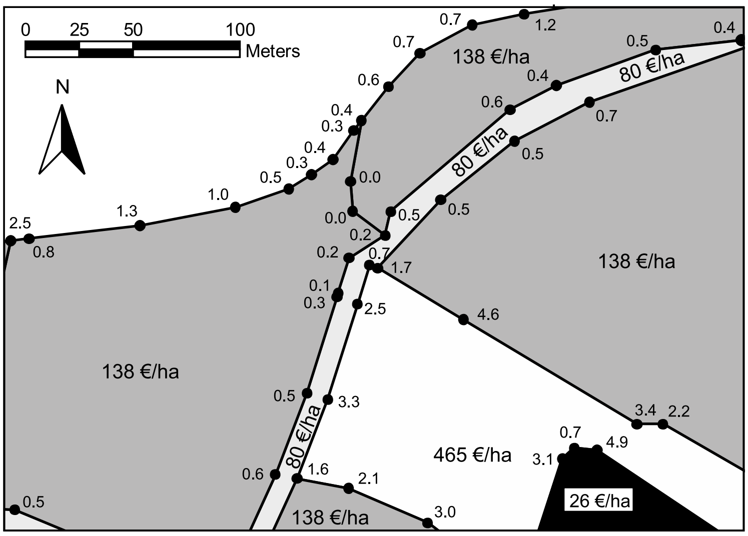


Figure 4. Contribution (in €) of individual vertices to std($U(\Omega)$). Inside the polygons are per unit area values in €/ha.

Increasing the accuracy of all vertices

Recall that the subsidy value $U(\Omega)$ of the study area is € 27455. With an accuracy of σ = 2.0 m for all vertices std($U(\Omega)$) = $(3814 - 2276)^{0.5}$ = € 39.22 and the 95% confidence interval [27377, 27531] = $(1\pm0.0028)\cdot U(\Omega)$ = within ±0.28% of $U(\Omega)$. Increasing or decreasing the positional accuracy of all vertices at the same time (scenario 1), revealed the following relationship between std($U(\Omega)$) and σ:

$$\text{std}(U(\Omega)) = 19.96\cdot\sigma \qquad (13)$$

The 95% confidence limits are then: $U(\Omega) \pm 1.96\cdot\text{std}(U(\Omega)) = U(\Omega) \pm 39.11\cdot\sigma$. Filling in σ = 10 m (Magnussen 1996) in (13) yields a 95% confidence interval of $(1\pm0.014)\cdot U(\Omega)$. To realize a further reduction of the 95% confidence limits on $U(\Omega)$ to $(1\pm0.0010)\cdot U(\Omega)$, the positional accuracy should be increased to $\sigma = 0.0010\cdot U(\Omega)$ / 39.11 = 0.7 m.

Increasing the accuracy of a subset of vertices

Figure 5 shows the upper limit of the two-sided 95% confidence interval on $U(\Omega)$, for each of the subsets defined in scenario 2. These limits were calculated from the variances of each subset as computed with the model. A comparison between scenarios 1 and 2 shows that a 95% confidence limit of $(1\pm0.001)\cdot U(\Omega)$ can be obtained by either increasing the accuracy of all vertices from 2.0 m to 0.7 m (scenario 1) or by increasing the accuracy of 35% of the vertices to 0.5 m (scenario 2, subsets 2.4 and 2.5).

The lowest possible std($U(\Omega)$) given the subset size and the accuracy of the vertices in the subset is calculated with subset 2.5. The question was if an equally low std($U(\Omega)$) could be achieved with subsets identified with simple selection criteria (table 3, subsets 2.1 to 2.4). The effectiveness of these criteria is now discussed. The fact that subset 2.1 yielded the lowest reduction in std($U(\Omega)$) indicates that there exist vertices inside the study area that contribute more to std($U(\Omega)$) than vertices located on the outer boundary of the study area. For example, a larger reduction in std($U(\Omega)$) was obtained by increasing the accuracy of vertices delineating the most valuable polygons (subset 2.2). The selection criterion for subset 2.2 however does not address the source of the contribution of individual vertices to var($A(p_q)$) and cov($A(p_q)$,$A(p_r)$). For example when a 'valuable' polygon is delineated by a relatively large number of vertices $s_q(i)$, then the distance between vertices $s_q(i+1)$ and $s_q(i-1)$ is small. Equation 9 then shows that the contribution of these vertices to var($A(p_q)$) is small. The selection criteria for subsets 2.3 and 2.4 were based on distances between vertices and as a result yielded a larger reduction in std($U(\Omega)$). Yet however large the distance between vertices is, the contribution to std($U(\Omega)$) remains zero if a vertex is shared by polygons with identical per unit area values. This fact is accounted for in subset 2.4. Figure 5 shows that with a simple selection criterion as that applied to obtain subset 2.4, we came very close to the lowest possible std($U(\Omega)$) as defined by subset 2.5.

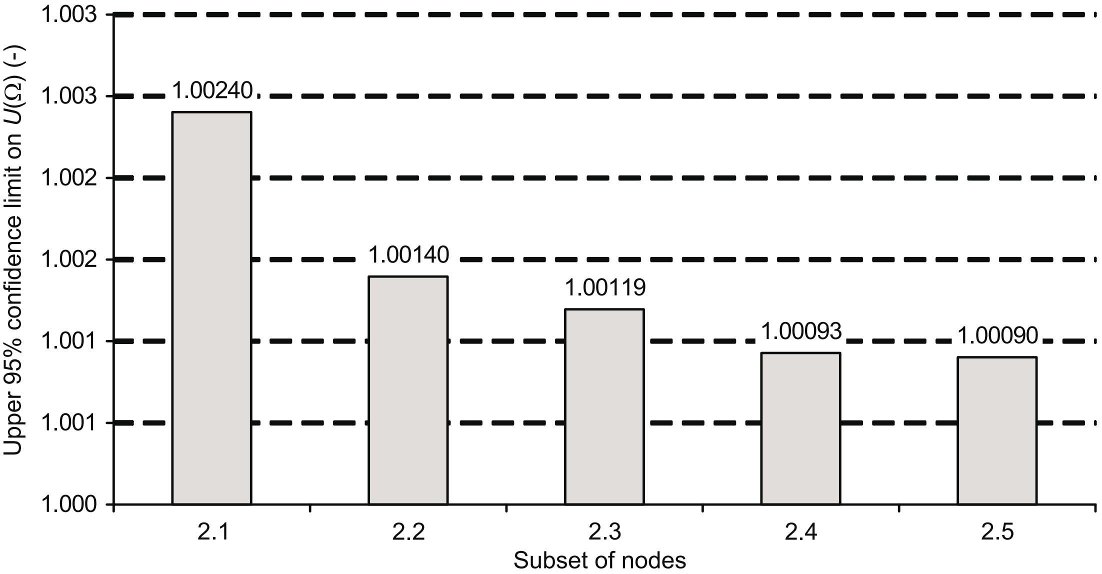


Figure 5. Upper 95% confidence limits on $U(\Omega)$ for 5 subsets of vertices for which the positional accuracy was increased. The vertical axis is scaled to $U(\Omega)$ so that 1.001 corresponds with $(1+0.001)\cdot U(\Omega)$.

5.4 Conclusions and discussion

Our aim was to develop a model to calculate the uncertainty in the utility value of an area described by a set of polygons with different per unit area utility values and uncertainty in the area of these polygons. To achieve this, equations for the variance and covariance were derived from a standard GIS area calculation. The equations were implemented in a GIS and applied to case study by the National Forest Service of the Netherlands (SBB). The following two sections provide a discussion on the relevance of this study for SBB and recommendations for further research.

SBB receives subsidies for the management and conservation of the nature present in a small study area. Knowledge of uncertainty reduction in these subsidies can support SBB managers in decision-making with regard to resources spent on spatial data quality. Two scenarios showed how much the uncertainty in the subsidy value could be reduced as a result of increasing the positional accuracy. A comparison between the scenarios showed that the same uncertainty reduction could be achieved by either increasing the accuracy of all vertices to $\sigma = 0.7$ m or by increasing the accuracy of 35% of the vertices to $\sigma = 0.5$ m. If different costs are associated with these two options, then such a comparison may aid in selecting the cheapest option of the two.

Thus the model provides information that can support SBB in its decision-making. To provide meaningful information for actual decision making on spatial data quality, a cost-benefit analysis of the different options to increase/decrease data quality should be added. Further, the model could be expanded by including measurements instead of expert estimates of positional accuracy. The validity of the model assumptions described in part 1 needs further checking. One assumption was that fuzziness could be captured in the error-terms σ_x and σ_y. Although no evidence is given in this chapter, one may expect that area variance is underestimated if existing fuzziness is not fully captured in these terms (see also Magnussen 1996). Further, a number of assumptions on the correlation of errors were made and require checking. This checking involves measurement of correlation of errors in x- and y- direction and of correlation between errors in different vertices. Keefer et al. (1991) provide an example of how correlation between errors in different vertices can be measured and modelled. Complementary to measurements, a covariance equation could be derived that includes these correlations, just as Chrisman and Yandell (1988), Griffith (1989) and Bogaert et al. (2005) did for the variance equation.

6. Detection and risk in digging activities[1]

[1] van Oort, P.A.J., Bregt, A.K. and de Bruin, S., submitted. Detection and risk for digging activities around underground cables and pipelines: implications of spatial data quality. Submitted to Transactions in GIS.

6. Detection and risk in digging activities

Abstract

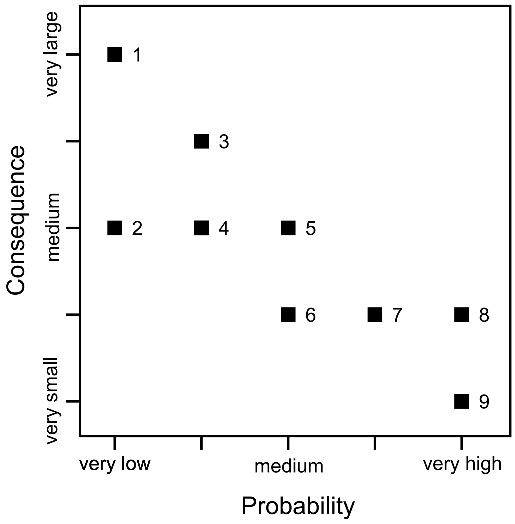
Digging activities are considered the largest cause of damage to underground cables and pipelines. Contractors can reduce the risk through detection, which will cost time and thus money. In the Netherlands, maps are the prime source of information on the location of cables/pipelines and detection time strongly depends on whether maps indicate the presence of cables and pipelines. Poor quality maps can contribute to increased risk or higher risk avoidance costs. The objective of this chapter is to present a model for calculating the trade-off between detection costs and risk and for calculating implications of over- and incompleteness of maps. The model aims to find the optimal detection time at which the sum of detection cost and risk is at its minimum. A case-study showed that it is possible to parameterise the model with data collected from contractors through a questionnaire. The case-study provides a numerical example of calculation of the trade-off between risk and detection costs and provides and example of calculation of costs of incompleteness. We conclude that the model contributes valuable new insight. However, more and location specific data are needed to enable operational use of the model.

6.1 Introduction

6.1.1 Damage to underground cables and pipelines

Underground cable and pipeline networks are used for transportation of utilities such as gas, water, electricity, telecom, sewage and industrial liquids. Studies from the U.S., U.K., Australia, Germany, Japan and the Netherlands, reported in Muhlbauer (2004) and in Pauwels en Wieleman (2004), indicate that damage to these networks is in the order of at least tens of millions of euros per country per year. Figure 1, based on Pauwels and Wieleman (2004), provides qualitative insight into the probability and consequence of damage to the different types of cables and pipelines present in The Netherlands. Five general causes of damage are distinguished by CONCAWE, the oil companies European organisation for environmental and health protection: (1) mechanical failures, (2) operational errors, (3) corrosion, (4) natural hazards and (5) digging activities. Analyses of 30-years damage records on oil and gas pipeline networks in Europe and the U.S. by Guijt (1998) and True (2004) indicate that digging activities represent the largest cause of damage: between 30 to 50% of all damages could be attributed to digging activities. These percentages are European and U.S. averages for pipelines. Higher percentages may be expected where the frequency of digging activities is higher, for example in more densely populated areas such as the Netherlands. Also higher percentages may be expected for cables, which are buried less deep and with higher densities inside urban areas.

VELIN, the Dutch organisation for owners of high-pressure (> 40 bar) pipelines, analysed 262 incidents on 14,500 km of pipeline over a period of 4 years. 97% of the damages was attributed to digging activities (VELIN 2004). Although this percentage is exaggerated by the fact that the analysis also included near misses, the difference with the 30-50% reported by True (1998) and Guijt (2004) is striking. It shows that the impact of digging activities is larger in more densely populated areas. With increasing world population and increasing urbanisation the impact of digging activities may also be expected to increase in other countries of the world.



Legend

1. transmission of gas (40-80 Bar, 11,600 km); and industrial liquids (40-100 Bar, 3,500 km);
2. sewage (82,406 km)
3. gas (1-8 Bar, 34,000 km); electricity (50/110/150 kV, 3,500 km); urban heating (3,600 km)
4. drinking water (109,366 km)
5. electricity (3-25 kV, 103,000 km)
6. gas (0.1 Bar, 88,150 km)
7. electricity (0.4 kV, 150,000 km)
8. Telecom (> 240,000 km)
9. Radio & television cables (> 150,000 km); home-connections (=connection of homes to networks, > 150,000 km)

Probability: very low is <0.03 per 1000 km*year; very high is >5.7 per 1000 km*year
Consequence: very small is <250 euro per incident; very large is >75,000 euro per incident

Figure 1. Probability and consequence of damage to all cables and pipelines in the Netherlands, based on Pauwels and Wieleman (2004).

To reduce damage due to digging activities, understanding of the digging decision-making process is needed. It is generally accepted that contractors can reduce damage by spending efforts in detecting cables and pipelines. Nevertheless, there appears to be hardly any quantitative research into the trade-off between detection costs and damage reduction, nor research into characteristics of digging activities that affect this trade-off. To make inferences on how characteristics of digging activities affect the probability of damage one needs data about both those digging activities that did and those that did not result in damage. Hardly any publication includes data on digging activities that did not result in damage (the study by van Houten en Lourens (1995) is the single exception known to us). Most studies however are based only on records in damage registration systems (for example see Cooke and Jager 1998, True 1998 and Guijt 2004). In the analysis reported in this chapter, data on digging activities with no damage are included.

Contractors, utility companies and researchers may benefit from each others knowledge. Research may provide answers to questions and issues raised by these companies, while addressing these questions can enrich the research agenda. Contractors and utility companies are confronted with questions and issues on spatial data quality (Dijkema en Zweistra 2001, Meeus en Schoof 2003, de Kruif 2004c, van Oort 2004 and van der Stok 2005) and questions and issues on spatial data infrastructures (which are called one-call systems in their terminology, see U.S. DoT 1999, Muhlbauer 2004, de Kruif 2004a, FGDC 2004). Many publications on spatial data infrastructures (SDI) and spatial data quality (SDQ) can be found in geographical information science journals, but none of these report on digging activities around underground cable and pipeline networks. In this chapter we address the issue of SDQ. Despite concerns about SDQ in digging activities around underground cables and pipelines, no conceptual model is available to provide insight into how exactly SDQ can contribute to damage, no mathematical model is available to calculate the implications of SDQ and little quantitative insight exists into the implications.

6.1.2 Aim and outline

Digging activities are a major cause of damage to underground cables and pipelines. Contractors invest in reducing damage through detection. The trade-off between detection and damage has to date never been quantified. Concerns have been raised about the quality of maps, but it is unclear how consequences of errors in maps can be quantified. The aim of this chapter is to develop a model that meets the following criteria:

1. It must formalise (and thereby provide insight into) how contractors make the trade-off between detection costs and risk reduction;
2. It can be used to quantify the trade-off between detection costs and risk reduction;
3. It can be used to quantify economic impacts of errors in cable/pipeline maps;
4. It must be possible to estimate the model parameters from questionnaires distributed among contractors, since they are the only source of information on without-damage digging activities.

Considering the lack of scientific research into how contractors make the trade-off between detection costs and risk reduction, a major output of the research will be the formulation of a model. The model presented in section 6.2 is widely applicable; the case-study in section 6.3 is based on data from the Netherlands.

6.2 Model development

The conceptual model (§2.1) is based on interviews with contractors and utility companies, articles in Dutch professional journals and consultancy reports. The mathematical model (§2.2) is our formalisation of the conceptual model.

6.2.1 Conceptual model

Trade-off between detection and risk

We propose a model as depicted in figure 2. The positively sloping line shows the detection costs which increase linearly with detection time. The lower curve shows the risk which is the product of damage per hit and expected number of hits. The upper curve is the total of detection cost and risk. Clearly, there is an optimal detection time at which the total costs are at their minimum.

Normally, risk is defined as the probability of an event multiplied by the consequence of the event. Here, because more than one cables can be present and can be hit during one digging activity, risk is defined as the expected number of events (=cable/pipeline hits) multiplied by their consequences. The expected number of hits decreases non-linearly with detection time because detection starts with the most efficient detection methods (map interpretation, aboveground detection, see Jeong and Abraham 2004) and then proceeds to less efficient methods (underground detection). The most common form of underground detection is to dig a number of transects perpendicular to the expected trajectory of a cable/pipeline. With an increasing number of transects, the probability of encountering a deviation or extra cable in a new transect decreases.

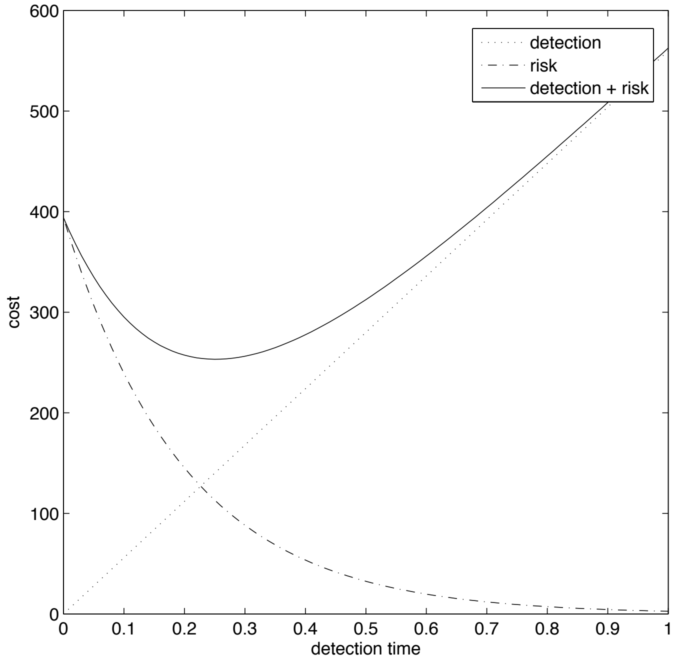


Figure 2. Conceptual model of the trade-off between detection costs and risk.

Spatial data quality

The trade-off between detection costs and risk is affected by three spatial data quality elements: completeness, vertical positional accuracy and horizontal positional accuracy:

1. **Completeness**. If a cable/pipeline is present but not shown, then this is a case of incompleteness resulting in increased risk for the contractor who assumes that detection is unnecessary. In the opposite case of overcompleteness the contractor unnecessarily spends a certain amount of time on detection. Contractors can anticipate on incompleteness by spending some time on detection even if according to maps no cables/pipelines are present. But then they could needlessly be spending time on detection in digging activities if conform the maps no cables/pipelines are present.
2. **Vertical positional accuracy**. If the depth of a digging activity is less than the depth of cables/pipelines shown on the map, then no hits are expected and the optimal detection time is assumed to be zero. If an error exists in the vertical position in the upwards direction then that can result in increased risk.
3. **Horizontal positional accuracy**. As indicated in figure 2, detection reduces risk. The steepness of the slope of the risk curve depends amongst others on the relative horizontal positional accuracy. A higher relative horizontal positional accuracy will result in the slope being less steep. Relative positional accuracy is the accuracy of the position of the cable/pipeline relative to other objects on the same map. Absolute positional accuracy is the accuracy of objects in the map relative to a

coordinate system. Currently it is not an issue for contractors. It would be an issue if digging machinery were fully controlled by digital maps in combination with positioning systems such as GPS, GALILEO or GLONASS.

Although the model presented in the next section could be used to calculate the implications of all three elements, we have chosen to describe the procedure to calculate implications only for the element completeness. Contractors and utility companies pointed to this element as the most relevant element for further analysis, because:

1. According to contractors, maps are used to find the approximate horizontal position of cables and pipelines, not the exact position. Once their location has approximately been found, other aboveground and underground detection methods are used for finding the exact position.
2. According to utility companies, depth can easily change while they remain uninformed about these changes. Depth may change due to people adding or removing a layer of soil, or due to erosion. As a result, utility companies have little control over the reliability of information on this attribute. In the Netherlands, as a consequence of the lack of control, utility companies do not provide information on depth. Good digging practice rules for depths of different cables and pipelines exist but utility companies cannot be held accountable for deviations from these rules.
3. On the other hand, interviews made clear that completeness is an issue and can be improved by the utility companies. For example, they could increase the speed with which their records are updated and could develop protocols to increase the completeness based on notifications from contractors.

6.2.2 Mathematical model

The conceptual model visualised in figure 2 consists of equations 1 to 5, with definitions of variables and parameters in tables 1 and 2 respectively. The expected costs for a contractor are calculated with equation 1. Equation 1 calculates the total costs $Cost_{tot}$ as the sum of detection costs $Cost_{det}$ (eq. 2) and risk $Cost_{risk}$ (eq. 3). Detection costs depend on detection time $Time_{det}$ and on detection costs $Cost_{lab}$. The fact that more than one cables and pipelines can occur on one digging location has implications for the definition of risk (explained in §2.1.1) and for the definition of parameters $\Pr_{liveonly}(map,loc)$ and $\Pr_{live\&dead}(map,loc)$ (see §2.2.2). Note that these two parameters depend on the number of cables/pipelines shown on the map(s) (variable *map*) and on location (variable *loc*). Risk (eq. 3) is calculated as the costs in case of a hit $Cost_{hit}$ (eq. 4) times the expected number of hits $E(hits)$ (eq. 5). In equation 5 an exponential function is used to describe the non-linear decrease of the expected number of hits $E(hits)$ with detection time $Time_{det}$. Note that a decrease of $E(hits)$ with $Time_{det}$ implies negative values for $par_{det,liveonly}$ and $par_{det,live\&dead}$.

$$Cost_{tot} = Cost_{det} + Cost_{risk} \tag{1}$$

$$Cost_{det} = Cost_{lab} \cdot Time_{det} \tag{2}$$

$$Cost_{risk} = Cost_{hit} \cdot E(hits) \tag{3}$$

$$Cost_{hit} = Cost_{claim} \cdot \Pr_{claim} \tag{4}$$

$$
\begin{aligned}
&E(hits) = \\
&\Pr_{liveonly}(map, loc) \cdot \Pr_{vert} \cdot \exp\begin{pmatrix} par_{prsize} \cdot \log(Time_{tot}) + par_{det,liveonly} \cdot (Time_{det} / Time_{tot}) + \\ \sum_{i=1}^{N-1} par_{soil}(i) \cdot soil(i) \end{pmatrix} + \\
&\Pr_{live\&dead}(map, loc) \cdot \Pr_{vert} \cdot \exp\begin{pmatrix} par_{prsize} \cdot \log(Time_{tot}) + par_{det,live\&dead} \cdot (Time_{det} / Time_{tot}) + \\ \sum_{i=1}^{N-1} par_{soil}(i) \cdot soil(i) \end{pmatrix}
\end{aligned}
\tag{5}
$$

Table 1. Variables in the model.

variable	description	unit
Time$_{det}$	detection time	hour
Time$_{tot}$	project time = total time of the digging activity	hour
Time$_{det}$ /*Time*$_{tot}$	fraction detection time	-
map	number of cables / pipelines shown on map(s)	-
loc	location of the digging activity	-
soil(i)	binary indicating whether soil type *i* is present, with *i* = 1,*N* and *N* the total number of soil types	-
Time$_{opt}$	optimum detection time, at which total costs are at their minimum	hour

Expected number of hits

The expected number of hits equals zero if digging occurs at a depth where no cables pipelines are present ($\Pr_{vert}$ = 0 in eq. 5) and on a location where not a single cable or pipeline is present ($\Pr_{liveonly}(map,loc)$ = 0 and $\Pr_{live\&dead}(map,loc)$ = 0 in eq. 5). If one or more cables/pipelines are present within the digging depth, then the expected number of hits depends on the detection time ($Time_{det}$) conditional to the project size ($Time_{tot}$) and the soil type ($soil(i)$).

To allow for a comparison between digging activities of different size, detection time is divided by the total project time to obtain the unit-free fraction detection time ($Time_{det}$ / $Time_{tot}$) in equation 5. Equation 5 takes the natural log of $Time_{tot}$ so that if par_{prsize} = 1 then the expected number of hits increases linearly with project size. If 0 < par_{prsize} < 1 then the odds of a hit are relatively large in smaller projects but still increasing with project size. If par_{prsize} < 0 then more hits are expected in smaller projects. Pauwels and Wieleman (2004) have hypothesised relatively large risks in smaller projects, thus par_{prsize} < 1. One reason to expect a relatively large probability in smaller projects is that requests for maps to the one-call system are more frequently omitted in smaller projects. Another reason occurs when in a larger project cables/pipelines are only present in a small part of the total digging area.

Through $\Pr_{liveonly}$, $\Pr_{live\&dead}$, $par_{det,live}$ and $par_{det,live\&dead}$ the model recognises the possible implications of presence of dead (out-of-order) cables and pipelines. Based on our interviews, we expect that risks are less efficiently reduced if dead cables and pipelines are present, because a contractor may stop his detection activities when first finding a dead cable or pipeline, so that the probability of hitting the live cable or pipeline increases. Thus in terms of equation 5, we expect $par_{det,liveonly} < par_{det,live\&dead} < 0$.

Table 2 Parameters in the model.

parameter	description	unit
$Cost_{tot}$	total costs	euros
$Cost_{lab}$	detection costs (mainly labour, but may also include other costs, e.g. costs of machinery)	euros/hour
$Cost_{claim}$	damage claim paid by contractor in case of registered damage	euros/hit
$\Pr_{claim}$	probability that damage is registered immediately by the utility company. Only in that case is a damage claim sent to the contractor.	-
$\Pr_{vert}$	probability that one or more cables are present within the digging depth.	-
$\Pr_{liveonly}(map,loc)$	probability that one or more live (active) and zero dead (out-of-order) cables or pipelines are present, at location *loc* and with *map* indicating the number of cables pipelines present according to available map(s)	-
$\Pr_{live\&dead}(map,loc)$	probability that one or more live (active) and one or more dead (out-of-order) cables or pipelines are present	-
$\Pr_{live}(map,loc)$	probability that one or more live (active) cables or pipelines are present. Ignoring out-of-order cables and pipelines	-
par_{prsize}	project size parameter	1/hour
$par_{det,liveonly}$	detection parameter for when one or more live (active) and zero dead (out-of-order) cables or pipelines are present	-
$par_{det,live\&dead}$	detection parameter for when one or more live (active) and one or more dead (out-of-order) cables or pipelines are present	-
$par_{det,live}$	detection parameter for when one or more live (active) cables or pipelines are present.	-
$par_{soil}(i)$	soil type parameter	-
N	number of soil types	-

Definition of parameters $\Pr_{live}$, $\Pr_{liveonly}$ and $\Pr_{live\&dead}$

Often, but not always, cables and pipelines lie buried close together at a single digging location. This has implications for the definition of $\Pr_{live}$, $\Pr_{liveonly}$ and $\Pr_{live\&dead}$. Consider the case where maps show 4 cables/pipelines on the digging location, while actually 5 are present. Then two possible definitions of $\Pr_{live}$ are:

1. treat this as 4 cases of a complete map and 1 case of an incomplete map; or
2. treat this as one case of {≥1 shown on the map(s), ≥1 actually present}.

In this chapter, the second definition is used. The first definition implies determining per digging activity an optimum detection time for each of the individual cables/pipelines and then summation of these optimum detection times to obtain the optimum detection time for the digging activity. However, if cables/pipelines are close together, detection times for individual cables/pipelines will coincide so that one cannot simply sum the optimum detection times for individual cables/pipelines. The two motivations to choose for the second definition were:

1. Our interviews indicated that contractors do not distinguish between optimum detection times per individual cable/pipeline but do try to apply optimum detection times per digging activity. Since one of the criteria of the model was that it must be possible to estimate the parameters from questionnaires distributed among contractors, it was of great importance to choose a definition that is in-line with the contractors' perception of their decision-making processes.
2. It simplifies the model while retaining the most determining effect of over- and incompleteness, namely the distinction between zero and more than zero cables/pipelines present.

Ignoring possible implications of dead (out-of-order) pipelines, $\Pr_{liveonly}(map,loc)$ and $\Pr_{live\&dead}(map,loc)$ can be written as $\Pr_{live}(map,loc)$. Based on our interviews and the discussion above, we define the instances of *map* and *loc* as:

- *map* = ["≥1 shown on the map(s)", "0 shown on the map(s)", "no map(s) available"];
- *loc* = ["inside built-up area", "outside built-up area and along infrastructure", "outside built-up area and not along infrastructure"].

We can now define overcompleteness as $\Pr_{live}$("≥1 shown on the map(s)") = 0, that is, in reality not a single cable or pipeline is present while according to map(s) one or more cables or pipelines are present. Similarly, we can define incompleteness as $\Pr_{live}$("0 shown on the map(s)") = 1, that is in reality one or more cables or pipelines are present while none are shown on the maps. We will use subscripts *act* and *asm* to refer to the actual and assumed values of parameters $\Pr_{live}$, $\Pr_{liveonly}$ and $\Pr_{live\&dead}$. Then note that while the actual value of $\Pr_{live,act}$ in an individual digging activity is a binary event, a contractor may assume for $\Pr_{live,asm}$ any value between 0 and 1. If in k out of K digging activities it happens that one or more cables/pipelines are present then a plausible assumption for $\Pr_{live}$ is $\Pr_{live,asm} = k/K$.

Implications of over- and incompleteness

Assuming that contractors select optimum detection times the model can be used to calculate implications of over- and incompleteness and the costs of making the wrong assumption on the over- and incompleteness. With $\Pr_{live}$, the procedure to calculate the extra costs of an assumption is, in three steps:

1. Calculate optimal detection times: $Time_{opt,asm}$ given $\Pr_{live,asm}$ and $Time_{opt,act}$ given $\Pr_{live,act}$;
2. Calculate expected costs: $Cost_{tot,asm}$ given $Time_{opt,asm}$ and $\Pr_{live,act}$ and $Cost_{tot,act}$ given $Time_{opt,act}$ and $\Pr_{live,act}$ (note: different detection times, same $\Pr_{live,act}$);
3. calculate extra costs = $Cost_{tot,act} - Cost_{tot,asm}$

The case-study in section 6.3 presents examples of calculations on the costs of incompleteness.

6.3 Case-study

6.3.1 Methods

Model

The introduction noted a general lack of data on digging activities that did not result in damage. The only way to obtain such data is by attending digging activities or to get the information from contractors through a questionnaire. As will be shown in section 6.4, contractors and utility companies can benefit from a fully parameterised model. However, these benefits do not necessarily occur at the level of individual contractors. For individual contractors filling in questionnaires costs precious time with no direct reimbursement foreseeable. For this reason we anticipated low response rates and a need to limit the length of the questionnaire and thereby the size of the model. Equation 7 is the model parameterised in the case-study. Equation 8 presents the derivation of the optimal detection time for this model. Definitions of parameters and variables of this model are found in tables 1 and 2.

$$\begin{aligned} Cost_{tot} &= Cost_{claim} \cdot \Pr_{claim} \cdot \Pr_{live}(map,loc) \cdot \\ &\exp\left(par_{prsize} \cdot \log(Time_{tot}) + par_{\mathrm{det},live} \cdot (Time_{\mathrm{det}} / Time_{tot})\right) + Cost_{lab} \cdot Time_{\mathrm{det}} \end{aligned} \tag{7}$$

$$Time_{opt} = \frac{Time_{tot}}{par_{\mathrm{det},live}} \cdot \left(\begin{array}{l} \log\left(-\dfrac{Cost_{lab} \cdot Time_{tot}}{Cost_{claim} \cdot \Pr_{claim} \cdot \Pr_{live}(map,loc) \cdot par_{\mathrm{det},live}} \right) - \\ par_{prsize} \cdot \log(Time_{tot}) \end{array} \right) \tag{8}$$

Parameter estimation

This section describes how each parameter was estimated, based on responses to a questionnaire sent out by CUMELA (an organisation representing contractors in the Netherlands). Bruil Ede b.v. and another contractor who chose to remain anonymous gave access to their damage registration systems.

Parameter *Cost_{claim}* was calculated as the mean of damages from the damage registration systems of two large contractors. The means of the two contractors are used separately in two scenarios (table 3).

Parameter $\Pr_{claim}$. There are two ways to estimate parameter $\Pr_{claim}$. The first (eq. 9) is based on data from utility companies and may underestimate $\Pr_{claim}$ because the annual number of failures also includes failures due to other causes than digging activities. The second (eq. 10) is based on data from the questionnaire and may overestimate $\Pr_{claim}$ if not all hits are reported. Both ways to estimate $\Pr_{claim}$ were applied and used separately in two scenarios (table 3)

$$\Pr_{claim} = \frac{\text{annual number of claims}}{\text{annual number of failures}} \tag{9}$$

$$\Pr_{claim} = \frac{\text{number of hits reported to utility company}}{\text{number of hits}} \qquad (10)$$

Parameter Pr$_{live}$(*map*,*loc*). Ideally, parameter Pr$_{live}$(*map*,*loc*) would be estimated from the number of cables/pipelines actually present. The only way to be absolutely sure about this number is to uncover the entire area, which is obviously very costly. As an approximation, the number encountered during digging can be used. In the questionnaire, contractors were asked to report per digging activity (1) the number of cables/pipelines reported in/by the map(s), (2) the number of cables/pipelines encountered during digging and (3) location (“inside built-up area”, “outside built-up area and along infrastructure” or “outside built-up area and not along infrastructure”).

Parameters *par$_{prsize}$* and *par$_{det,live}$*. Contractors were asked to register per digging activity values of *Time$_{det}$*, *Time$_{tot}$*, *map*, *loc* and *hits*. *hits* is a discrete variable, with a high number of 0s (no damage) and incidentally a positive discrete number of hits. For such a data set, the appropriate method to estimate *par$_{prsize}$* and *par$_{det,live}$* is through Poisson regression (Neter et al. 1996). We may expect a contractor to apply a larger detection time *Time$_{det}$* if he anticipates a high value of *Cost$_{claim}$* . Parameters *par$_{prsize}$* and *par$_{det,live}$* are estimated from *Time$_{det}$*, *Time$_{tot}$* and *hits* and may be dependent on contractors' anticipated values of parameters *Cost$_{claim}$*, *Pr$_{claim}$* and Pr$_{live}$(*map*,*loc*). We would like to estimate *par$_{prsize}$* and *par$_{det,live}$* independent of anticipated values of *Cost$_{claim}$*, *Pr$_{claim}$* and Pr$_{live}$(*map*,*loc*). Therefore, *par$_{prsize}$* and *par$_{det,live}$* were estimated only from recordings of digging activities where we were sure that the estimate of Pr$_{live}$(*map*,*loc*) would be very close to one. A test showed no significant correlation between the fraction detection time (*Time$_{det}$* / *Time$_{tot}$*) and the expected damage in case of a hit (*Cost$_{claim}$*). Interference of Pr$_{claim}$ was not investigated.

Parameter *Cost$_{lab}$*. Parameter *Cost$_{lab}$* was calculated as the mean of personnel costs per hour per contractor. Other costs (machinery) were not considered because they are relatively small and because they can be quite difficult to estimate.

6.3.2 Results & discussion

400 contractors were sent a questionnaire, 19 contractors returned a total of 81 questionnaires. In the text below, the number of records from which parameters were estimated is often lower than 81, because often not all fields in the questionnaire were filled in.

Parameters

Table 3 shows parameter estimates. The two scenarios reflect the uncertainty on the values of parameters *Cost$_{claim}$* and Pr$_{claim}$. Scenario 1 is the scenario with the lowest risk, scenario 2 is the scenario with the highest risk. Parameter *Cost$_{claim}$* was estimated from damage registration systems of two different contractors. Investigation of the cause of the large difference was beyond the scope of this chapter. From de Kruif (2004b), based on an interview with experts of a utility company, Pr$_{claim}$ was estimated as in equation 9: 1100/1700 = 0.65. From the questionnaire we obtained that 25 out of 29 damages were reported to the owner, thus Pr$_{claim}$ = 0.862.

For parameters *par$_{prsize}$* and *par$_{det,live}$* it was tested whether they deviated significantly from zero. Note that since detection reduces risk *par$_{det,live}$* must be lower than 0, so that the alternative hypothesis for $par_{det,live}$ was H_1: $par_{det,live} < 0$. Tests yielded:

1. Test H_0: $par_{prsize} = 0$, H_1: $par_{prsize} \neq 0$, result: $par_{prsize} = -0.3109$, s.e. = 0.1349, p-value 0.0212;
2. Test H_0: $par_{det,live} = 0$, H_1: $par_{det,live} < 0$, result: $par_{det,live} = -4.9869$, s.e. = 4.4646, p-value 0.1320.

Both null hypotheses were rejected at $\alpha = 0.15$, suggesting that the expected number of hits $E(hits)$, equation 5, is as we proposed correlated with total project time and with detection time. The standard errors however are so large that accurate prediction of $E(hits)$ is impossible. The negative value for par_{prsize} is consistent with results from interviews by Pauwels and Wieleman (2004) which suggested that $par_{prsize} < 1$. Histogram 6, see the discussion in §3.2.4, suggests a possible underestimation of par_{prsize}.

Table 3. Parameter estimates.

parameter	estimate		unit
	scenario 1*	scenario 2*	
$Cost_{claim}$	407	894	euros
Pr_{claim}	0.65	0.862	-
Pr_{live}	see table 4		-
par_{prsize}	-0.3109	-0.3109	1/hour
$par_{det,live}$	-4.9869	-4.9869	-
$Cost_{lab}$	28	28	euros/hour

* the two scenarios reflect uncertainty in the parameters $Cost_{claim}$ and Pr_{claim}.

Table 4 lists for 48 digging activities whether 0 or ≥1 cables/pipelines were encountered. From table 4, one can calculate the probability that ≥1 cables/pipelines are encountered, given that one is inside built-up area, with 0 cables/pipelines shown on the map. This probability is calculated as Pr_{live}("0 shown on the map(s)", "inside built-up area") = 2/(2+3) = 0.4.

Table 4. Frequency of presence and absence of cables and pipelines.

	Location					
	inside built-up area		outside built-up area		not specified	
number encountered* →	≥1	0	≥1	0	≥1	0
≥1 shown on map(s)	15	1	11		9	
0 shown on map(s)	2	3	1	4	1	1

* encountered during the digging activity

Trade-off between detection and risk

Figure 3 shows how the expected number of hits decreases with the fraction detection time and with increasing project size. Figure 4(a) shows, for a project of size $Time_{tot}$ = 20 hours, the trade-off between detection and risk. The black and white markers reflect the uncertainty in parameters $Cost_{claim}$ and Pr_{claim} as in scenarios 1 and 2 (table 3). In scenario 1, the optimum fraction detection time is around zero. Scenario 2 would give an optimum fraction detection of approximately 0.2. The circles in figure

4(b) show the expected number of hits associated with these optima, respectively 0.39 and 0.15. Although no alternative quantitative data are available to compare with, these optima appear rather high. Further research should indicate whether they can be explained by uncertainty in model parameters, or by other factors.

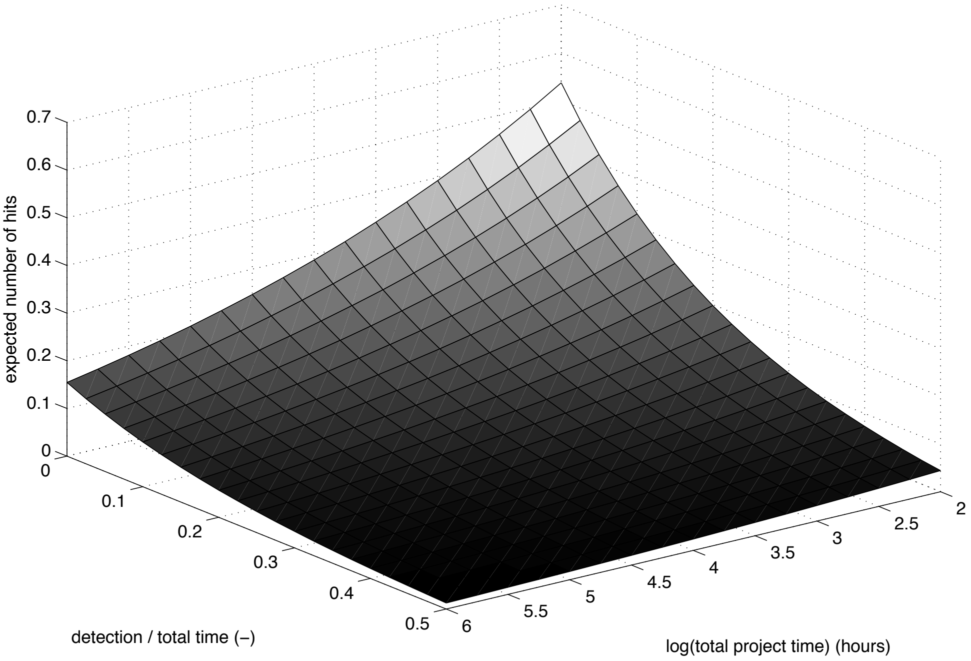


Figure 3. Expected number of hits decreases with fraction detection time and is higher in smaller projects.

With the scenario 2 parameters $Cost_{claim}$ = 849, Pr_{claim} = 0.862, $\mathrm{Pr}_{live}(map,loc)$ = 1, par_{prsize} = -0.3109, $par_{det,live}$ = -4.9869 and $Cost_{lab}$ = 28 and with total project time $Time_{tot}$ = 20 hours the model yields (eq. 8) an optimal detection time of $Time_{opt}$ = 4 hours, corresponding with an optimal fraction detection time of $Time_{opt}$ / $Time_{tot}$ = 0.2. At this optimum, the expected costs (eq. 7) are $Cost_{tot}$ = 224 euros.

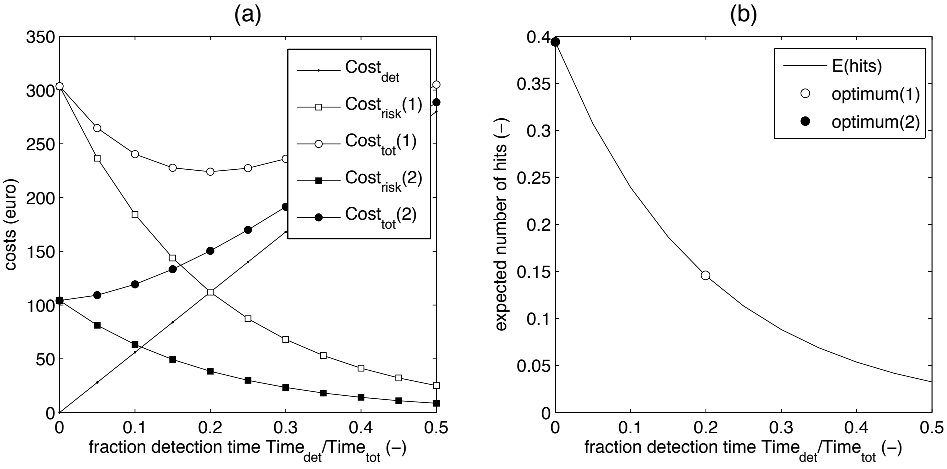


Figure 4. Trade-off between detection and risk. Part a shows the costs, b shows the expected number of hits. Black and white markers, corresponding with scenarios 1 and 2 (table 3), reflect the uncertainty in parameters $Cost_{claim}$ and Pr_{claim}.

Implications of over- and incompleteness

Continuing with the scenario 2 parameters and with total project time $Time_{tot}$ = 20 hours, some calculations on incompleteness ($\mathrm{Pr}_{live,act}$("0 shown on the map(s)",*) = 1). The asterix (*) is used to indicate that the calculation applies for any location. The subscripts *act* and *asm* are used for actual and assumed. For a contractor who assumes error-free maps, following the procedure outlined in section 6.2.2, we obtain:

1. Optimum detection times given $\mathrm{Pr}_{live,asm}$("0 shown on the map(s)",*) = 0 and $\mathrm{Pr}_{live,act}$("0 shown on the map(s)",*) = 1 are: $Time_{opt,asm}$ = 0 hours and $Time_{opt,act}$ = 4 hours;
2. Given $\mathrm{Pr}_{live,act}$("0 shown on the map(s)",*) = 1 the expected costs are: $Cost_{tot,asm}$ = 304 euros and $Cost_{tot,act}$ = 224 euros;
3. The extra costs due to assuming an error-free map are 304 - 224 = 80 euros.

Table 5 shows the expected total costs and extra costs for three different assumptions $\mathrm{Pr}_{live,asm}$("0 shown on the map(s)",*), when in a digging activity maps are complete and incomplete. From table 5, one can calculate the expected costs when incompleteness occurs for example in 4 or in 0 out of 10 digging activities. Results of these calculations are shown in table 6. In case of incompleteness in 4 out of 10 the best thing to do is to assume $\mathrm{Pr}_{live,asm}$("0 shown on the map(s)",*) = 0.4, with costs 1,211 euros. If maps are always error-free (0 out of 10 incomplete) then the best thing to do is to assume $\mathrm{Pr}_{live,asm}$("0 shown on the map(s)",*) = 0.0, with costs 0 euros. Over 10 digging activities of $Time_{tot}$ = 20 hours the costs of incompleteness in 4 out of 10 digging activities are then at least 1,211 euros.

Table 5. Total and extra costs depending on $\Pr_{live,asm}$ and $\Pr_{live,act}$.

		$Cost_{tot}$ (eq.7): total costs (euros) per digging activity[1]		extra costs (euros) per digging activity[1,2]	
$\Pr_{live,asm}$ ("0 shown on map")	$Time_{opt,asm}$	$\Pr_{live,act}$ ("0 shown on map") = 1 (incomplete)	$\Pr_{live,act}$ ("0 shown on map") = 0 (complete)[3]	$\Pr_{live,act}$ ("0 shown on map") = 1 (incomplete)	$\Pr_{live,act}$ ("0 shown on map") = 0 (complete)[3]
0	0	304	0	80	0
0.4	0.3	290	9	66	9
1	4	224	112	0	112

[1] assuming parameters as in scenario 2 (table 3) and total project time $Time_{tot}$ = 20 hours

[2] calculated as outlined in section 6.2.2

[3] $\Pr_{live,act}$ ("0 shown on map") = 0 is complete because conform the map(s) no cables/pipelines are present. The only costs are then the detection costs: $Cost_{det} = Time_{opt,asm} \cdot Cost_{lab}$

Table 6. Example of calculation of costs of incompleteness.

	nr. of digging activities[1] with		sum over all digging activities	
$\Pr_{live,asm}$ ("0 shown on map")	$\Pr_{live,act}$ ("0 shown on map") = 1 (incomplete)	$\Pr_{live,act}$ ("0 shown on map") = 0 (complete)[2]	total costs[3] (euros)	extra costs[3] (euros)
0	4	6	1,215	319
0.4	4	6	1,211	315
1	4	6	1,566	670
0	0	10	0	0
0.4	0	10	88	88
1	0	10	1,117	1,117

[1] assuming parameters as in scenario 2 (table 3) and total project time $Time_{tot}$ = 20 hours in all 10 digging activities

[2] $\Pr_{live,act}$ ("0 shown on map") = 0 is complete because conform the map(s) no cables/pipelines are present

[3] by multiplication of the number of digging activities in this table with costs in table 6

Actual and optimal fraction detection time

In the questionnaire, respondents filled in for individual digging activities the actual detection time $Time_{det}$. With the model, we can compute for each of these individual digging activities the optimal detection time $Time_{opt}$. Then, the difference between the two can be calculated (shown in figure 5) and the extra costs due deviations from the optimum can be calculated (figure 6). In both figures, the top and bottom histogram reflect the uncertainty in parameters $Cost_{claim}$ and $\Pr_{claim}$ as in scenarios 1 and 2 (table 3). In figure 5 the top histogram (mean -0.0648) suggests that contractors took slightly less risk than optimal, the bottom histogram (mean 0.0211) suggests that contractors took slightly more risk than optimal. Visually, the histograms in figure 5 suggest that contractors' detection times are close to optimal. Figure 6 shows that generally, deviations from the optimum bring along extra costs of less than

200 euro per digging activity. For the three digging activities with extra costs higher than 500 euro a partitioning (not shown) into detection costs and expected risk revealed that the extra costs were all due to excessive detection costs and not due to excessive risks. This could mean either that par_{prsize} is underestimated in the model, or risk-adverseness of contractors or their contractors. More data need to be collected in order to determine which of the two is the case.

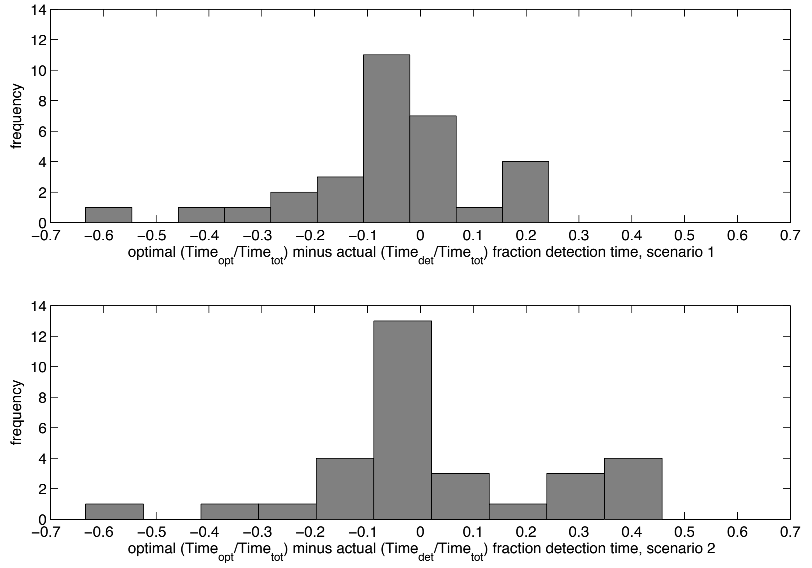


Figure 5. Histograms of difference between optimal and actual fraction detection time for the two scenarios (table 3). The top histogram (mean -0.0648) suggests that contractors took slightly less risk than optimal; the bottom histogram (mean 0.0211) suggests that contractors took slightly more risk than optimal. Expected extra costs due to deviation from the optimum are shown in figure 6.

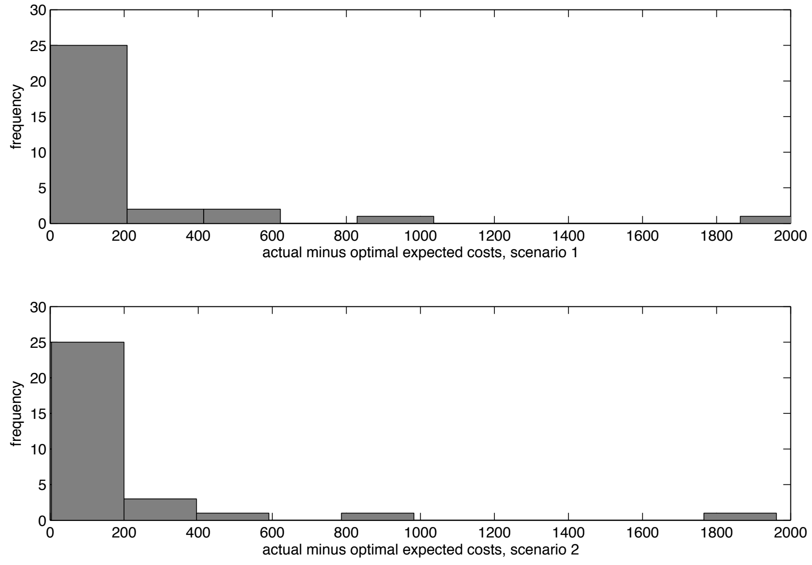


Figure 6. Histograms of expected extra costs caused by deviations from the optimal fraction detection time.

Detection time and expected damage

Clearly, it is advantageous to spend more time on detection if the expected damage in case of a hit is larger. The questionnaire also asked contractors to indicate how much damage they expected in case of a hit. Expected economic damage categories (in euros) were: (1) <500, (2) < 5,000, (3) <50,000 and (4) >50,000. Expected physical danger categories were: (1) “no physical danger”, (2) “physical danger, non lethal” and (3) “lethal”. We tested for correlation between fraction detection time and expected damage and obtained very low correlation coefficients of 0.19 and 0.22, with one-sided p-values of 0.19 and 0.16. There are two possible explanations for this inconsistency: (1) the model does not correctly represent the digging decision-making process and (2) contractors are poorly informed about damage in case of a hit. Results of van Houten and Lourens (1995) support the second explanation.

6.4 Conclusions & further research

Digging activities represent the largest cause of damage to underground cables and pipelines, but very little scientific research is available. Dutch professional journals, consultancy reports and newspaper have raised concerns about spatial data quality. There was a lack of insight in which element of spatial data quality (SDQ) was most relevant and a lack of insight into how implications of SDQ could be quantified. Based on this chapter seven conclusions can be drawn:

1. The formulation of the conceptual and mathematical model is general and therefore widely applicable. The issue addressed occurs all over the world, the case-study presented some quantitative results for the Netherlands;
2. The model provides insight into the trade-off between detection costs and risk;
3. The model can be used to calculate the optimal detection time;
4. Three spatial data quality parameters are relevant: completeness, horizontal positional accuracy and vertical positional accuracy. Completeness is the most relevant element;
5. The model could be used to calculate implications of all three elements. The chapter presents and illustrates a procedure to calculate implications of over- and incompleteness;
6. The case-study proved that it is possible to parameterise the model with data derived from questionnaires among contractors;
7. The case-study results are consistent with our model of the contractor’s behaviour, that is, histograms suggest that their detection times are close to optimal.

6.4.1 Beyond insight: applicability in decision-making

Once parameterised with acceptable accuracy, the following potential applications of the model are foreseen:

1. At the start of a project contractors can use the model to decide on how much to pay for detection. Or conversely, contractors may confront the contractor with the expected damage and expected number of hits for the sum that the contractor offers to pay for detection;

2. The model could be used to substantiate claims by Dutch contractors (Dijkema en Zweistra 2001, de Kruif 2004c, van der Stok 2005) who argue that they suffer the consequences of poor quality maps;
3. For contractors that take more risk than optimal the model could be used to convince them of the benefits of increasing their detection time;
4. When detection efficiencies and mean detection fractions per contractor are known, they could become a selection criterion in bids for contracts. Through economic competition this could lead to an overall increase in safety for public and contractors;
5. Costs for utility companies can be included in the model. Once included in the model, the model could be used to estimate the costs and benefits of increasing/decreasing the spatial data quality of their maps (as already suggested in van Oort 2004).

6.4.2 Further research

The case-study illustrated that the model could, as required, be parameterised with data collected from contractors through a questionnaire. The previous section showed that, once parameterised with acceptable accuracy, applications valuable for society may be expected. The first recommendation is therefore to collect more data so that the model can be put to use. Two modelling challenges lie ahead. Firstly, the challenge of finding a method that accounts of the coincidence of detection times for individual cables/pipelines. Secondly, to make parameters dependent on the type of cable or pipeline, such as the size of the damage claim ($Cost_{claim}$). Once more data have been collected it will be a challenge to parameterise and validate the model.

7. Do users ignore spatial data quality?[1]

[1] Based on: van Oort, P.A.J. and Bregt, A.K., 2005. Do users ignore spatial data quality? – a decision-theoretic perspective. Risk Analysis 25(6) © 2005, the Society for Risk Analysis.

7. Do users ignore spatial data quality?

Abstract

Risk analysis (RA) has been proposed as a means of assessing fitness for use of spatial data but is only rarely adopted. The proposal is that better decisions can be made by accounting for risks due to errors in spatial data. Why is RA so rarely adopted? Most Geographical Information Science (GISc) literature stresses educational and technical constraints. In this chapter we propose, based on decision-theory, a number of hypotheses for why the user would be more or less willing to spend resources on RA. The hypotheses were tested with a questionnaire, which showed that the willingness to spend resources on RA depends on the presence of feedback mechanisms in the decision-making process, on how much is at stake and to a minor extent on how well the decision-making process can be modelled.

7.1 Introduction

Spatial data are data about topography and specific themes of the earth's surface. They are an ingredient in almost all public decision-making (Burrough and McDonnell 1998, Cornélis and Brunet 2002), a popular estimate is that 80% of the data used by managers and decision-makers is spatial. Many researchers have stressed the need to deal with issues of Spatial Data Quality (SDQ), as the risk of misuse of spatial data has greatly increased (see conference proceedings: Lowell and Jaton, 1999; Heuvelink and Lemmens, 2000; Mowrer and Congalton, 2000; Hunter and Lowell, 2002). Important causes of this increased risk of misuse are the growing availability of spatial data, greatly enhanced access to these data, enhanced possibilities of manipulating these data and a growing group of inexperienced users (Aronoff 1989, Morrison 1995; Doucette and Paresi 2000). Furthermore, producers of spatial data sets provide little information regarding the quality of these data sets (Östman 1997, Jakobsson and Vauglin 2001). Geographical Information Systems (GIS) have limited functionality for visualisation of SDQ (van der Wel 2000) and limited functionality for quantifying error-propagation (Forier and Canters 1996, Heuvelink 1998, Duckham, 2002). Apart from these factors, Openshaw (1989) noted that the traditional response of both researchers and users has too often been to simply ignore SDQ. Agumya and Hunter (1999) categorised users into three groups according to how they respond to SDQ in data:

1. Those who establish that the data are suitable prior to using the data;
2. Those who wish to chose the best among several suitable data sets;
3. Those who use data regardless of their suitability, either because they must use it or they choose to ignore its SDQ.

According to Agumya and Hunter (1999), studies about the proportion of users in each class are scant, but there is a general belief among researchers that most users fall into the third class. Fitness for use can be established in various ways: consultation with experts, trial and error, by assessing legal or other constraints to the use of a data set, by assessing for which purposes the data set is already used and for which purposes it is produced, to the most rigid method which is to establish fitness for use through risk analysis (RA). In RA implications of errors or uncertainties on expected outcomes of a decision-making process are quantified and data sets are only used if the risks generated are acceptable (Agumya and Hunter 2002). Several authors have shown that RA can be used as a means of assessing fitness for use, see for example Li et al. (2000), Crosetto and Tarantola (2001), de Bruin et al. (2001) and de Bruin and Hunter (2003). Users are often assumed to 'ignore' SDQ if they did not establish fitness for use through RA prior to using spatial data sets (Openshaw 1989, Agumya and Hunter 1999). Knowledge of the risks caused by SDQ can contribute in several ways to decision-making:

- It helps in finding out which level of SDQ is required for the decision, and this can aid in selecting the data set to be used;
- It helps in deciding how many resources need to be reserved for mitigation;
- It helps in deciding on the deciding the risk-level of an investment;
- It helps in redesigning the decision-making process to reduce sensitivity to low SDQ;
- By being candid towards stakeholders about the risks taken in a decision, it is possible to preserve public trust in public decision-making.

Given these advantages, one may wonder why RA is so rarely adopted as a means of establishing fitness for use. One possible explanation is the lack of time and money. More informative explanations explain why the costs are considered to be too high and why the benefits are considered to be too low. A full list of hypotheses or explanations as to why users so rarely quantify risks due to SDQ prior to using spatial data in decision making is rarely found in geographical information science (GISc) literature (but see Openshaw 1989). Rather, most research articles briefly state that risks due to SDQ are not quantified and that this may be a problem. The greatest part of such a research article then addresses one or more of the following problems:

1. Lack of awareness that SDQ may cause risks;
2. Lack of practical examples illustrating the needs and benefits of RA;
3. Users lack the skills to conduct RA;
4. Earlier research has been limited to error propagation analysis on the geographical analyses conducted using GIS, the outcomes of which are less tangible to users than outcomes reported in terms of risks;
5. Lack of methodologies for calculating risks due to SDQ;
6. Lack of supporting tools in current GIS software (for example error-buttons, visualisation, wizards);
7. Poor documentation of SDQ.

The seven problems mentioned above can be summarised as two classes of hypotheses as to why users ignore SDQ: educational (1 to 4) and technical (5 to 7). The origin of this chapter lies in an effort to address the problems as listed above. In 2002, we conducted open interviews with GIS/spatial data users in public decision-making to find case studies in which these problems could be addressed. Surprisingly, we noted in these interviews that the users concerned were well aware of spatial data quality issues, yet in many cases they considered quantification of implications of SDQ (= Risk Analysis, RA) unnecessary or impossible. They deliberately chose to ignore SDQ. The three problems addressed in this chapter are (1) that existing educational and technical explanations could not fully explain these users' motivations, (2) that there is hardly any empirical research into why users ignore spatial data quality and (3) that there appears to be an unknown discrepancy between researchers and spatial data users in public decision-making.

The aims of this chapter are therefore: (1) to propose new hypotheses as to why users ignore SDQ, (2) to test these hypotheses in a questionnaire and (3) to test for differences between respondents of the questionnaire.

7.2 Methods

7.2.1 Definitions and outline

Before proposing hypotheses, relevant definitions are given in Table 1. Section 7.2.2 then presents hypotheses based on decision-theory.

Table 1. Definitions

Decision (-making): a process consisting of five phases (from Cornélis and Brunet 2002)

1. DO: documentation and information = identification of data needs, problems, objectives and knowledge of how these can be achieved;
2. DA: decision analysis = definition of alternative actions to achieve these objectives;
3. DT: decision taking = selection of one of the alternatives;
4. DI: decision implementation = implementation of the selected action and;
5. DE: decision evaluation = evaluation of the outcomes of the action in terms of objective achievement.

SDQ: Spatial Data Quality. A general term, covering all aspects that affect the quality of a spatial data set. Prominent elements of SDQ are positional accuracy, attribute accuracy, completeness, logical consistency and lineage. See chapter 2 of this thesis.

RA: Risk-analysis. Quantification of risks by multiplying potential outcomes of a decision with the probabilities of these outcomes. Probabilities of outcomes are calculated, using error propagation analysis, from probabilities of errors. Probabilities of errors are derived from SDQ reports of the input data set(s). If necessary, additional information on these errors (and their correlation!) is acquired.

In the decision-making process we distinguish two types of actors:

(Co-) decision-maker: has the power to select one of the alternatives in phase DT of the decision

Stakeholder: does not have the power to select one of the alternatives in phase DT of the decision, but does experience the consequences in DE.

7.2.2 Hypotheses

Hypotheses were derived from open interviews with users of spatial data in various government organisations in the Netherlands and from a selection of decision-theoretic literature references. A list of hypotheses is presented at the end of this section; below, the sources of the hypotheses are listed.

Cornélis and Brunet (2002) described the spatial decision-making process as a sequence of five phases (see Table 1). They described the various sources of uncertainty that occur in the decision-making process and provided a policy-maker point of view. In each phase, we distinguish in this chapter a feedback mechanism that, if present, may reduce risks generated by SDQ. Hypothesis 1 is that if a decision-maker expects that a mechanism sufficiently reduces risks, then this may reduce his or her willingness to spend resources on RA. Concrete descriptions of the feedback-mechanisms are given in Tables 3 and 4 and in the discussion (§7.4.2).

In order to calculate risks as a result of SDQ in a decision-making process, the decision-making process needs to be described in a model. The two most important modelling steps that have to be taken to allow for RA are (Morgan et al. 1990, Jaeger et al. 2001): (1) it must be possible to rank potential outcomes of the decision-making process according to their desirability and (2) the relationship between input data, the alternative actions and outcome of each {input data, action, outcome}-combination must be described by a set of equations. These modelling steps are further referred to as RA modelling criteria. Once meeting these criteria becomes more problematic, users may start to become more concerned about how realistic the risks calculated with such a model are. They may become more sceptical about the potential benefits of RA. See van der Smagt and Lucardie (1991) for a general comment on applicability of models under not well-defined conditions. Hypothesis 2 is that people will be less willing to spend resources on RA if RA modelling criteria are more difficult to be met.

Campbell (1991) distinguished between four kinds of decision-making processes, which in this chapter are reduced to two kinds: data-driven and discussion-driven decisions. Criteria for RA are less likely to be met in discussion-driven decisions: it may be more difficult to model these processes and more difficult to rank expected outcomes according to their desirability. Often, but not necessarily, a greater public interest is at stake in discussion-driven decisions. The opposite is the case in data-driven decisions: RA criteria are more likely to be met, but less may be at stake. Thus in both the data-driven and the discussion-driven decision there is one argument in favour and one against RA. It will be interesting to see if one argument outweighs the other and if so, which. Hypothesis 2 and 3 test whether modelling criteria and the stakes affect the willingness to spend resources on RA, hypotheses 4 and 5 test whether one outweighs the other in data or discussion-driven decisions.

On the basis of the decision-theoretic literature cited above, we propose the following hypotheses: People will be less willing to spend resources on RA if:

1. A feedback mechanism reduces risks due to SDQ;
2. RA modelling criteria are more difficult to be met:
   - Criterion 1: it is difficult to rank possible outcomes of the decision-making process according to their desirability;
   - Criterion 2: it is difficult to cast the decision-making process into a set of model equations;
3. Not much is at stake;
4. The decision is data-driven;
5. The decision is discussion-driven.

7.2.3 Questionnaire

Trials with the hypotheses inserted directly as a questions in a questionnaire proved not to be an option, because they were considered to be too abstract by respondents. Therefore, they were translated into a set of features and associated arguments. The formulation of the hypotheses and their translation into features and arguments is the result of a learning process of the authors, fed by open interviews with GIS/spatial data users in public decision-making, experience in spatial data quality research, literature research and trial questionnaires.

Questions in the questionnaire

The decision-theoretic hypotheses were formalised into nine features and seventeen associated arguments, listed in Tables 3 and 4. Table 2 shows the relationship between the five hypotheses mentioned in §7.3.2 and the nine features in Table 3. Arguments are associated with features such that arguments 1a and 1b (Table 4) are associated with feature 1 (Table 3). Their relationship with the above hypotheses can be found through their association with features. The relationship between the five hypotheses and the features and arguments will also become clear in the discussion in section 7.4.2.1. The questionnaire contained the same questions for each feature and for each argument (see figure 1):

- **Feature question:** "If a decision-making process has this feature, then are you less / neutral / more willing to spend resources on risk-analysis?";
- **Argument question:** "Do you agree with this argument?" (Permissible answers: disagree / neutral / agree).

There were three reasons for adding arguments to the features:

- Arguments clarify the relationship between feature and RA;
- The feature questions require more commitment from respondents than argument questions. As a result respondents may be inclined to answer 'neutral' more frequently in the feature questions than in the argument questions. Arguments may uncover more information;
- If respondents agree with both the argument in favour and against RA, they may fill in 'neutral' in the feature question. In that case it would be wrong to discard the feature as being irrelevant.

The third aim of this chapter was to test for differences between respondents. To achieve this the questionnaire contained a **background question**: for their current job respondents could fill in (1) policy-maker, (2) cartographer, (3) GIS-expert, (4) researcher or (5) other, namely …. The definition of policy-maker given in the questionnaire was: "policy-maker: part of my job is to make decisions on how many resources will be spent on different projects". For the other job-categories no definitions were given, because a trial questionnaire indicated that their meaning was clear to respondents.

Table 2. Relationship between hypotheses and features

hypothesis		feature (see Table 3)								
(see §7.3.2)		1	2	3	4	5	6	7	8	9
1. feedback mechanism	– phase –									
	DO: documentation and information	x								
	DA: decision analysis		x							
	DT: decision taking					x				
	DI: decision implementation							x		
	DE: decision evaluation				x					
2. RA modelling criteria	– criterion –									
	ranking of outcomes			x						
	model as a set of equations					x	x		x	x
3. not much at stake							x			x
4. data-driven									x	
5. discussion-driven						x				x

Table 3. Features

ftr.	feature title
1	Stakeholders often know the area of interest very well. They are invited to inspect and criticise the spatial data to be used, prior to your decision taking.
2	Prior to implementation a field trip is organised. Data acquired during this trip may be reason to change your plans, even before implementation starts.
3	The aim of the decision is to achieve an objective. The extent to which this objective is achieved can be expressed numerically.
4	The decision will have implications for stakeholders. Stakeholders may hold you accountable and claim compensation for damage caused by errors in your data.
5	The decision is taken in negotiation between different parties.
6	After an action to be implemented has been selected it is implemented over the course of 10 years.
7	After an action to be implemented has been selected it is implemented. During implementation there is room for adjustment of the selected action
8	The decision-making process is clearly described.
9	The decision to be taken is a strategic one, in which policy-makers do their best to anticipate expected long-term developments.

7.2.4 Trials and final questionnaire

Before the final questionnaire, two trials were conducted. From the first trial we learned two lessons: (1) it proved better not to provide information about the specific context of decision-making processes and (2) it was better to apply a qualitative willingness-to-pay concept than a quantitative willingness-to-pay concept. In trials where we added context information we observed that respondents would make inferences about how much was at stake and then assign more resources to where they thought more was at stake, without paying further attention to presence or absence of other features/arguments in the context descriptions. Also, we found that respondents

Table 4. Arguments for and against RA, associated with features in Table 3

arg.	Argument title [†]
1a	Before presenting these data to them, I want to be sure that they do not contain errors. Otherwise they could use these errors against me or loose trust in me as a decision-maker. Therefore, in my opinion there is a greater need for RA if this feature is present.
1b	Stakeholders will notice errors and inform me about them. That way, risks are reduced. Therefore, in my opinion there is less need for RA if this feature is present.
2a	Field trips are expensive. If RA can contribute to reducing the number of field trips, then in my opinion there is a greater need for RA.
2b	During the field trip errors in the data are detected and the right action selected. That way, risks are avoided. If risks are avoided anyhow, then in my opinion there is less need to quantify them.
3a	If the objective can be expressed numerically, then it is possible to conduct calculations. Therefore, in my opinion there is a greater need for RA if this feature is present.
4a	I want to know if the risk is acceptable to me, for example I want to know how many claims for compensation may be expected. Therefore, in my opinion there is a greater need for RA if this feature is present.
5a	It is more difficult to model propagation of errors, as the decision-making process also depends on the negotiation behaviour of the actors involved. Therefore, in my opinion there is less need for RA if this feature is present.
5b	As soon as more than one party is involved, it is of importance to know how reliable each party's information is. Therefore, in my opinion there is a greater need for RA if this feature is present.
6a	Precisely because implementation takes place over such a long period of time, conscientious consideration is important. Therefore, in my opinion there is a greater need for RA if this feature is present.
6b	So much changes over the course of 10 years that little may be expected of RA at the start of the period. Therefore, in my opinion there is less need for RA if this feature is present.
7a	Adjustment during implementation will require repetition of the planning phase, which will cost both time and money. If RA can contribute to reducing these costs, then in my opinion there is a greater need for RA if this feature is present.
7b	No risks occur, because there is room for adjustment. Therefore, there is little use in calculating these risks. In my opinion, if this feature is present there is less need for RA.
8a	In the case of a clear procedure I expect that the designers of the procedure have made sure that it doesn't generate unacceptable risks. Therefore, in my opinion there is less need for RA if this feature is present.
8b	If the procedures are clear, then it is easily possible to calculate which risks are caused by SDQ. Therefore, in my opinion there is a greater need for RA if this feature is present.
9a	Great interests are at stake. Therefore, in my opinion there is a greater need for RA if this feature is present.
9b	The uncertainty in projections of long-term developments is often greater than the uncertainty in spatial data. Therefore, in my opinion there is less need for a calculation which shows the risks generated by uncertainty in spatial data.
9c	Characteristic of strategic decisions is that at the onset, it is not precisely clear which alternative actions will be considered. As a result, it is difficult to say what the risks will be. In my opinion, therefore, there is less need for RA.

[†] respondents were asked if they agreed with these arguments (see figure 1)

tended to answer “neutral” or “no response” when they were less familiar with the context of the decision-making process. In trials with more quantitative willingness-to-pay concepts, we found that respondents were only willing to answer quantitatively if much more context information was provided. Also, we found that part of the respondents does not regularly make decisions on resource spending. These respondents considered it difficult to respond quantitatively but did have a clear qualitative opinion.

In the second trial a questionnaire was distributed at a conference organised by a GIS software company attended by approximately 300 visitors. It contained an introductory letter stating the objective of the questionnaire, a simple example of RA on spatial data, a background question on the current job of respondents, five feature questions and one open question inviting to add other relevant features affecting the willingness to spend resources on RA. Two lessons were learned: (1) arguments had to be included for clarification and (2) respondents at the conference only partially represented the target group of our questionnaire. 73% Of the respondents indicated they were GIS-experts, less than 10% indicated they were policy-maker. We observed that a large number of GIS-experts had experienced difficulties in responding, because the questions were too far beyond the scope of their daily work.

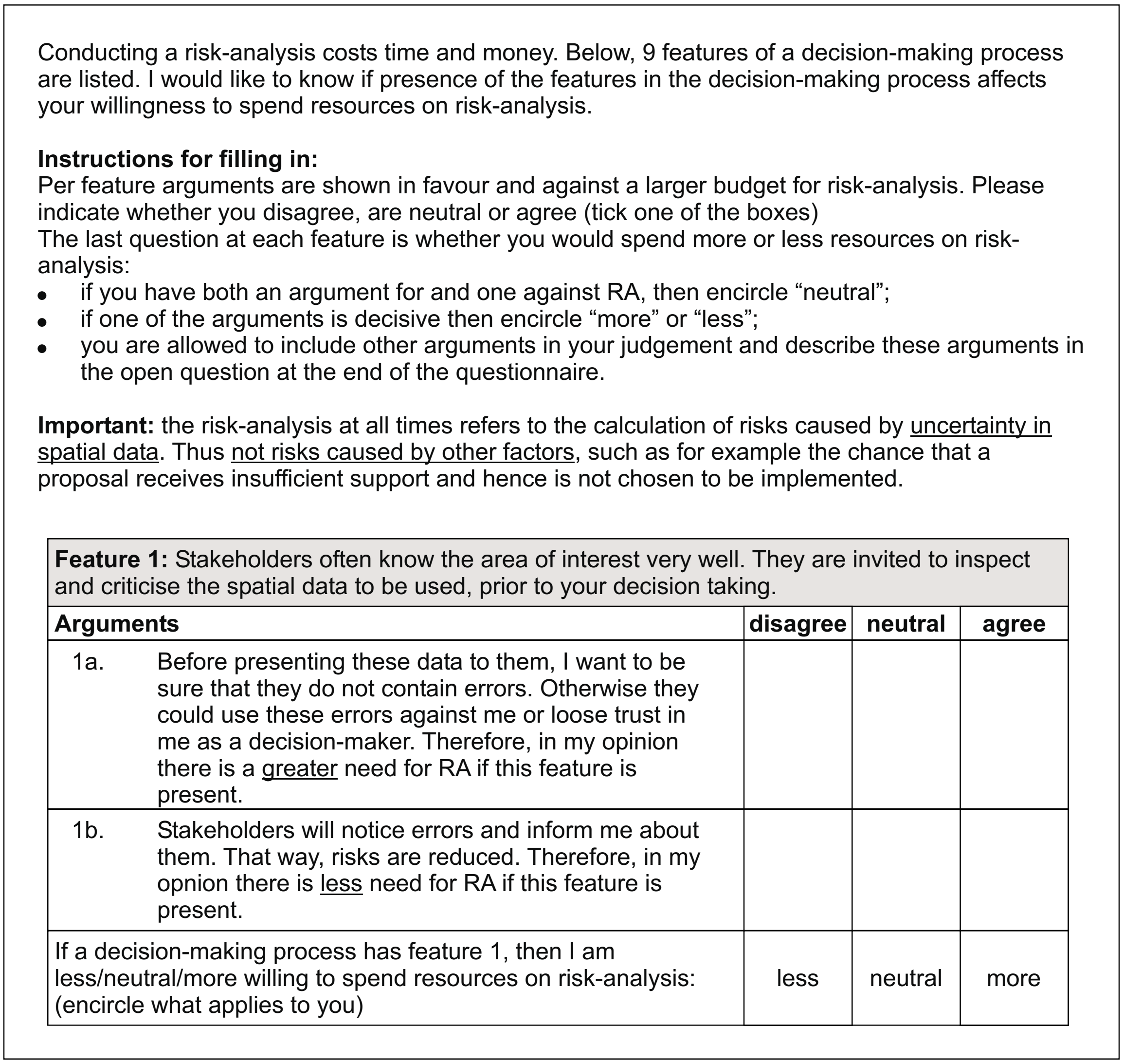
The final questionnaire contained an introductory letter stating the objective of the questionnaire, a simple example of RA on spatial data, a background question on the current job of respondents, all nine features and seventeen arguments of Tables 3 and 4 and an open question inviting to add other relevant features. Special efforts were made to include in the list of respondents policy-makers from various government organisations. Figure 1 is taken from the final questionnaire, it shows the instructions for filling in, one feature q uestion and two associated argument questions.

Statistical testing

The population was defined as Dutch policy-makers, managers, advisors, GIS-experts and researchers. Active use of a GIS was not a criterion. Respondents were personally targeted and contacted by e-mail and telephone and asked only to return the questionnaire if they felt willing and capable of filling it in.

Minimum sample size is calculated as a function of desired precision of the outcomes and prior knowledge on the distribution of the variable of interest. Precision refers to the p-values. In this study, p-values less than 0.15 were considered significant. No prior knowledge of the distribution of our variables was available. Based on the distribution observed from the first 20 questionnaires returned it was estimated that a sample of 50 should be the minimum.

The statistical test used was the Fisher Exact test for ordinal data (Agresti 1990: 60-65), which tests for differences in ranking and is an appropriate test for small sample sizes. The null hypothesis was an equal number of responses in each category (disagree/neutral/agree or less/neutral/more). It is possible that the total number of responses N, divided by the number of response categories (= 3) does not result in an integer. We applied H_0 = 17-18-17 for N = 52, H_0 = 18-17-18 for N = 53 and H_0 = 18-18-18 for N = 54. For example in Table 5 the response for feature 1 was 22-24-8 (N=53), so H_0 = 18-18-18 and we obtained a (one-sided) p-value of 0.052. Next the Fisher Exact test for ordinal data was applied to test for differences in ranking of responses according to the respondent's current job.



Conducting a risk-analysis costs time and money. Below, 9 features of a decision-making process are listed. I would like to know if presence of the features in the decision-making process affects your willingness to spend resources on risk-analysis.

Instructions for filling in:
Per feature arguments are shown in favour and against a larger budget for risk-analysis. Please indicate whether you disagree, are neutral or agree (tick one of the boxes)
The last question at each feature is whether you would spend more or less resources on risk-analysis:

- if you have both an argument for and one against RA, then encircle “neutral”;
- if one of the arguments is decisive then encircle “more” or “less”;
- you are allowed to include other arguments in your judgement and describe these arguments in the open question at the end of the questionnaire.

Important: the risk-analysis at all times refers to the calculation of risks caused by uncertainty in spatial data. Thus not risks caused by other factors, such as for example the chance that a proposal receives insufficient support and hence is not chosen to be implemented.

Feature 1: Stakeholders often know the area of interest very well. They are invited to inspect and criticise the spatial data to be used, prior to your decision taking.				
Arguments		**disagree**	**neutral**	**agree**
1a.	Before presenting these data to them, I want to be sure that they do not contain errors. Otherwise they could use these errors against me or loose trust in me as a decision-maker. Therefore, in my opinion there is a greater need for RA if this feature is present.			
1b.	Stakeholders will notice errors and inform me about them. That way, risks are reduced. Therefore, in my opnion there is less need for RA if this feature is present.			
If a decision-making process has feature 1, then I am less/neutral/more willing to spend resources on risk-analysis: (encircle what applies to you)		less	neutral	more

Figure 1. Instructions for filling in the final questionnaire.

7.3 Results and discussion

7.3.1 Results

From the 250 respondents targeted 55 filled in and returned the questionnaire. Table 5 shows the results for all respondents, with reference to features in Table 3 and arguments in Table 4. The Tables should be read as follows: Table 3, feature 1 states: “Stakeholders often know the area of interest very well. They are invited to inspect and criticise the spatial data to be used, prior to your decision taking”. Table 5 shows that if this feature is present in a decision-making process, then a significant ($p<0.15$) majority of the respondents is less willing to spend resources on RA. Arguments 1a (+, in favour of RA) and 1b (–, against RA) are associated with feature 1. Argument 1b (Table 4) states: “Stakeholders will notice errors and inform me about them. That way, risks are reduced. Therefore, in my opinion there is less need for RA if this feature is present”. Table 5 shows that a significant majority agreed with this argument. Phase 1 (Table 1, Cornélis and Brunet 2002) is: “1. DO: documentation and information = identification of data needs, problems, objectives and knowledge of how these can be achieved”. Recalling the decision-theoretic hypotheses listed in section 7.3.2, we conclude that in the opinion of the respondents there is less need for RA if feedback is received from stakeholders in phase 1 of the decision-making process.

Of the respondents 23 indicated they were policy-makers, 0 cartographer, 7 GIS-experts, 13 researchers and 12 other. The twelve indicating “other” all specified their jobs as either manager, advisor or coordinator. Significant differences were found between all groups. The most and largest differences were found between researchers (N=13) and policy-makers (N=23). Table 6 presents the response of these groups, with the last column showing one-sided p-values from the test for difference in ranking between the two groups.

Table 5. Feature and Argument questions, all respondents

ftr.[†]		less	neutral	more	no response	p-value	sign.[o]
1		22	24	8	1	0.052	*
2		10	24	19	2	0.159	
3		7	27	20	1	0.063	*
4		1	10	42	2	0.000	*
5		10	24	20	1	0.131	*
6		11	27	16	1	0.307	
7		16	22	16	1	0.548	
8		13	30	11	1	0.449	
9		12	19	23	1	0.108	*
arg.[†]	fa[‡]	disagree	neutral	agree	no response	p-value	sign. [o]
1a	+	24	17	14	0	0.125	*
1b	-	14	15	26	0	0.080	*
2a	+	6	12	36	1	0.000	*
2b	-	17	16	22	0	0.278	
3a	+	13	15	27	0	0.051	*
4a	+	3	4	46	2	0.000	*
5a	-	19	16	20	0	0.447	
5b	+	13	16	25	1	0.084	*
6a	+	11	19	24	1	0.070	*
6b	-	22	19	14	0	0.186	
7a	+	11	8	35	1	0.002	*
7b	-	23	11	19	2	0.302	
8a	-	24	15	16	0	0.173	
8b	+	19	21	14	1	0.304	
9a	+	7	12	36	0	0.000	*
9b	-	18	9	28	0	0.114	*
9c	-	17	21	16	1	0.485	

[†] Features: respondents were asked if presence of the feature affected their willingness to spend resources on risk-analysis (RA); Arguments: respondents were asked if they agreed with the argument. See Tables 3 and 4 for definition of features and arguments

[‡] argument is in favour (+) or against (–) RA

[o] p-values smaller than 0.15 were considered significant

Table 6. Feature and Argument questions, differences between respondents

	researchers (N=13)				policy-makers (N=23)					
ftr.[†]	less	neutral	more	no resp.	less	neutral	more	no resp.	p-value	sign.[º]
1	3	7	3	0	10	10	2	1	0.089	*
2	3	5	5	0	4	11	6	2	0.407	
3	2	6	5	0	4	12	6	1	0.311	
4	0	3	9	1	1	4	17	1	0.598	
5	5	5	3	0	3	9	10	1	0.059	*
6	2	5	6	0	6	14	2	1	0.035	*
7	2	5	6	0	7	13	2	1	0.026	*
8	2	7	4	0	6	13	3	1	0.149	*
9	3	2	8	0	5	10	7	1	0.117	*
arg.[†]	disagree	neutral	agree	no resp.	disagree	neutral	agree	no resp.	p-value	sign.[º]
1a	4	3	6	0	9	9	5	0	0.173	
1b	6	2	5	0	3	8	12	0	0.079	*
2a	2	2	9	0	1	7	15	0	0.513	
2b	6	1	6	0	6	9	8	0	0.464	
3a	3	4	6	0	7	7	9	0	0.360	
4a	0	2	11	0	2	2	19	0	0.418	
5a	7	1	5	0	6	9	8	0	0.274	
5b	6	2	4	1	2	9	12	0	0.031	*
6a	4	2	7	0	4	13	5	1	0.181	
6b	7	5	1	0	7	7	9	0	0.040	*
7a	3	1	9	0	4	3	16	0	0.496	
7b	7	2	3	1	8	7	7	1	0.179	
8a	7	3	3	0	11	6	6	0	0.425	
8b	4	4	4	1	9	11	3	0	0.226	
9a	3	0	10	0	3	7	13	0	0.208	
9b	5	3	5	0	7	3	13	0	0.218	
9c	7	4	2	0	5	9	8	1	0.038	*

[†] Features: respondents were asked if presence of the feature affected their willingness to spend resources on risk-analysis (RA); Arguments: respondents were asked if they agreed with the argument. See Tables 3 and 4 for definition of features and arguments

[º] p-values smaller than 0.15 were considered significant

7.3.2 Discussion

The discussion is about the five decision theoretic hypotheses (§7.3.2), formalised as features (Table 3) and arguments (Table 4). Features and arguments, in brackets in the text below, are significant (p-value < 0.15) unless explicitly stated otherwise.

Decision-theory

Hypothesis 1 is about feedback mechanisms in phases 1 to 5 (Table 1, Cornélis and Brunet 2002). Significant outcomes were found for the feedback mechanisms in phases 1, 3 and 5 of the decision-making process and not in phases 2 and 4. Phase 1 is "DO:

documentation and information = identification of data needs, problems, objectives and knowledge of how these can be achieved”. If feedback was received in this phase, then respondents had the opinion that this effectively reduced risks so that they were less willing to spend resources on RA (Table 5, ftr. 1, arg. 1b). Phase 3 is “DT: decision taking = selection of one of the alternatives”. In phase 3 the majority was more willing to spend resources on RA (Table 5, ftr. 5). Note the difference between the role of stakeholder (phase 1) and co-decision-maker (phase 3). Both may provide feedback on SDQ, but the co-decision-maker has the power to influence the decision taken and for this reason respondents regard his feedback with more distrust (Table 6, arg. 5b). Phase 5 is “DE: decision evaluation = evaluation of the outcomes of the action in terms of objective achievement”. Feedback in this phase implies that the decision-maker can be held accountable for damage done due to SDQ. Respondents were more willing to spend resources on RA if this feedback mechanism is present (Table 5, ftr. 4, arg. 4a). Epstein et al. (1998) present a discussion on the legal aspects of implementing this feedback mechanism. According to them, the decision-maker has a strong defence against damage claims if he can prove that he has implemented policies to maintain product quality. Amongst other things, quality depends on the risks generated by SDQ. The first step then in maintaining product quality is to quantify it. This can be done using RA. Other proofs of quality control are, for example, the ISO 9000 accreditation.

Phase 2 is “DA: decision analysis = definition of alternative actions to achieve these objectives”, phase 4 is “DI: decision implementation = implementation of the selected action”. The fact that no significant majorities were found for feedback in phases 2 and 4 (Table 5, ftr. 2 and 7 are not significant) is also interesting. We had anticipated that the presence of these mechanisms would reduce willingness to spend resources on RA, as was the case for feature 1. Significant majorities of the respondents took into account that these mechanisms are expensive, and hoped that applying RA might contribute to reducing these expenses (Table 5, arg. 2a and 7a). Respondents differed in their opinion as to whether or not feedback mechanisms in phases 2 and 4 reduced risks (Table 5, arg. 2b and 7b are not significant). Our experience is that these mechanisms certainly exist in many decision-making processes and thus reduce risks, and that respondents speculate about the necessity of these mechanisms but do not translate this into actively replacing them by RA. One reason for not doing so may be resistance from within the organisation, from the large workforce applying these mechanisms. As one of the respondents commented: “we always do fieldwork prior to decision-taking” and “there is no tradition of conducting RA within my organisation”. If feedback mechanisms in phases 2 and 4 reduce risk, be it not as cheaply as possible, a sense of urgency to replace them for RA may be lacking.

Hypothesis 2 is about RA modelling criteria. Criterion 1 is that it should be possible to rank expected outcomes of the decision according to their desirability; criterion 2 is that the decision-making process can be described in a model using a set of equations. Outcomes were significant for criterion 1 (Table 5, ftr. 3, arg. 3a), but not significant for criterion 2 (Table 5, arg. 5a, 6b, 8b). The only exception to this was that the majority agreed with argument 9b, against RA if criterion 2 is violated.

Hypothesis 3 was confirmed (Table 5, arg. 6a and 9a): a greater need for RA is felt if more is at stake. Hypothesis 4 is that respondents are less willing to spend resources on RA in data-driven decisions. The hypothesis was tested with feature 8 and neither rejected nor accepted. Hypothesis 5 was tested with features 5 and 9. Results from the questionnaire indicate users are more willing to spend resources on RA in discussion-driven decisions (Table 5, ftr. 5 and 9). Results indicate that in discussion-driven decisions, RA is desired because decision-makers want to know how reliable co-decision-makers data are (Table 5, arg. 5b) and because more is at stake in discussion-driven decisions (Table 5, arg. 9a). The argument that more is at stake can override the fact that a decision may be difficult to model (RA modelling criterion 2) and result in a higher willingness to spend resources on RA (Table 5, ftr. 9, arg. 9a, 9b).

Differences between respondents

Table 6 shows the differences in response between researchers and policy-makers. These differences can be summarised as:

- Policy-makers have more trust in feedback mechanisms in phases 1 and 4 of the decision-making process (Table 6, ftr. 1, 7, arg. 1b);
- Policy-makers are more concerned about the reliability of data from co-decision-makers (Table 6, ftr. 5, arg. 5b);
- Policy-makers are less willing to spend resources on RA if the decision is implemented over the course of 10 years (=long term). They agree more with the argument that "so much changes over the course of 10 years that little may be expected of RA at the start of the period. Therefore, in my opinion there is less need for RA" (Table 6, ftr. 6, arg. 6b);
- Policy-makers are less willing to spend resources on RA in strategic decisions. They agree more with the argument that "Characteristic of strategic decisions is that at the onset, it is not precisely clear which alternative actions will be considered. As a result, it is difficult to say what the risks will be. In my opinion, therefore, there is less need for RA" (Table 6, ftr. 9, arg. 9b).

7.4 Conclusion and further research

The title of this chapter is "Do users ignore spatial data quality? – A decision theoretic perspective". The word "ignore" is used in Openshaw (1989) and Agumya and Hunter (1999) where it seems to refer to the phenomenon that users use spatial data without quantifying risks due to SDQ prior to using these data. Strictly according to this definition, we observe that SDQ is frequently ignored. This chapter showed the inappropriateness of the use of the word "ignore". The chapter shows that although risks are rarely quantified, this certainly does not mean that SDQ is in all these cases ignored.

While lack of resources, awareness, skills, tools and poor documentation of SDQ certainly explain some of the reasons why risks due to SDQ are rarely quantified they do not provide the full explanation. Users of spatial data make decisions on the level of resources spent on quantifying risks due to SDQ before using a spatial data set in their decision-making process. These decisions are based on analysing features of the decision-making process. We found that respondents are less willing to spend resources on RA:

- If feedback on SDQ is received from stakeholders at the start of a decision-making process. Respondents expect that stakeholders will detect and inform the decision-makers about errors in the spatial data and so reduce risks to an acceptable level.

We found that respondents are more willing to spend resources on RA:

- if the objective of the decision can be expressed numerically;
- if feedback on SDQ is received from co-decision-makers in the decision-taking phase, because there is more concern about the reliability of co-decision-makers' information;
- if feedback on SDQ is received at the end of the decision-making process, that is if the decision-maker can be held accountable for risks due to SDQ;
- if more is at stake.

In cases where much is as at stake or where decision-makers need to rely on each others data, the desire for RA can override the fact that such decisions are at times difficult to model. This raises the question of how to incorporate factors that are difficult to model such as outcomes of negotiation processes, anticipation of long-term developments or the fact that at the start of the modelling process it is not yet clear for which decision the outcomes of the spatial analysis will be used. The answer to this question may be scenario-analysis, in which these factors are represented as a number of scenarios and propagation of spatial errors is modelled for each scenario.

In the questionnaire respondents were asked whether certain feedback mechanisms reduced risks sufficiently. The results therefore refer to stakeholders' perception of the effectiveness of these mechanisms. Further research should asses the actual effectiveness of these mechanisms and make a comparison with respondent's perceptions. The first mechanism to be addressed is feedback from stakeholders in the data acquisition stage. The majority of the respondents were of the opinion that this mechanism sufficiently reduced risks. Confirmation of rejection of this perception should be of importance to policy-makers who rely on this feedback mechanism as an alternative to RA.

We began with an image of users using spatial data in decisions making without prior consideration of risks due to spatial data quality. Users failed to quantify risks due to constraints such as lack of tools, theory and poor documentation of SDQ. What emerges from this chapter is an image of a user who assesses different aspects of his decision-making process and then decides whether or not quantification of risks due to SDQ is necessary. In view of the authors, this chapter presents only a start to learning more about why users fail to quantify risks due to SDQ prior to using spatial data. More empirical research into how users handle SDQ is needed. At least one of the topics that deserves further research is the confrontation between willingness to invest (outcomes of the questionnaire) and actual resource spending on RA.

8. Conclusions and recommendations

8. Conclusions and recommendations

Since the advance of GIS and automated data acquisition methods (such as satellites), concerns about possible implications of spatial data quality have risen. Based on Aronoff (1989), Morrison (1995) and Longley et al. (1999) the five main reasons for current concern about spatial data quality issues were identified as:

1. There is an increasing availability, exchange and use of spatial data;
2. There is a growing group of users less aware of spatial data quality;
3. GIS enable the use of spatial data in all sorts of applications, regardless of the appropriateness with regard to data quality;
4. Current GIS offer hardly any tools for handling spatial quality;
5. There is an increasing distance between those who use the spatial data (the end users) and those who are best informed about the quality of the spatial data (the producers).

To deal with these concerns, it is necessary to (1) formalise and standardise descriptions of spatial data quality and (2) to apply these descriptions in assessing the suitability (fitness for use) of spatial data, before using the data. The aim of this thesis was twofold: (1) to enhance the description of spatial data quality and (2) to improve our understanding of the implications of spatial data quality. To achieve these aims, six research questions were formulated. Conclusions on these questions are presented in section 8.1; section 8.2 lists recommendations for users, producers and researchers of spatial data.

8.1 Research questions

8.1.1 What is spatial data quality?

A definition of quality was given in chapter 2: "the totality of characteristics of a product that bear on its ability to satisfy stated and implied needs". Five sources of definitions of characteristics were compared, characteristics which are usually referred to as elements of spatial data quality. A comparison between these five sources showed that (1) for some elements the location within the meta-data report differs - not all elements are in the data quality section and (2) the explicitness with which elements are named as individual elements differs - elements are present but not explicitly named as individual elements. Apart from these differences, there appears to be little disagreement on which elements together define spatial data quality. Eleven elements were identified. Three of these elements ("variation in quality", "meta-quality" and "resolution") are often encountered not as individual elements but as sub-elements of other individual elements. One element ("semantic accuracy") is likely to dissolve in other elements in which case it is no longer recognised as an individual element.

8.1.2 Does land cover classification accuracy depend on landscape heterogeneity?

This question was answered positively in chapter 3. We quantified landscape heterogeneity using four landscape metrics and found classification accuracy to be positively correlated with these metrics. The study confirmed results of Smith et al. (2002, 2003) indicating that classification accuracy also depends on land cover class, focal heterogeneity and patch size.

8.1.3 How can knowledge of classification errors be used to improve land cover change estimates?

As an answer to this question, chapter 4 presented five methods, one based on expert knowledge and four based on the Bayes theorem and on theory on how classification errors propagate in the AND overlay (Veregin 1989). It was shown that error matrices can be used to improve land cover change estimates. It was also shown that the method based on error matrices performs poorly when classification errors at different dates are correlated. Correlation has a positive effect: it reduces the error in change estimates. For example if correlation is absent, the result of an AND overlay between data sets of with accuracies 0.8 and 0.7 will be $0.8 \cdot 0.7 = 0.56$. In case of 100% correlation the resulting accuracy will be $\min\{0.8,0.7\} = 0.7$. Methods that ignore this correlation (e.g. Fuller et al. 2003) may come to overly pessimistic conclusions on the fitness of spatial data sets for monitoring purposes.

The methods presented in chapter 4 all use input data derived from the change detection error matrix. In chapter 4, exact estimates of the true change could be obtained from the exhaustive change detection error matrix. The question then arises why one should choose to use the methods presented. The point is, often no change detection error matrix is available: most producers only provide error matrices for single dates. On other occasions, there is a change detection error matrix based on a small sample and an error matrix for one of the dates based on a large sample. In those cases, application of the methods presented in chapter 4 may be beneficial because they allow for the incorporation of information from error matrices. The situation of a small change detection error matrix and a large error matrix for one of the two dates is likely to occur in the analysis of historical land cover databases. In that case collection of reference data for change detection error matrices is difficult or even impossible, while error matrices will be readily available for the most recent dates of the time series.

8.1.4 How do positional errors in vertex coordinates propagate into errors in area estimates for polygons sharing a boundary?

Chapter 5 presented the derivation of both a variance and a covariance equation. The derivation of the covariance equation was new. In the chapter, the assumptions underlying the derivation are extensively discussed. It is shown how these equations can be put to use in the calculation of the uncertainty in the total subsidy value (in euros), for a spatial data set consisting of polygons sharing boundaries and with different subsidy values (euros per hectare) attached to each polygon. Further, it is shown that the model can be used to answer some interesting what-if questions. How much would the variance in the total subsidy value change if the positional accuracy of all vertices were increased or decreased by amount X? How much would the variance in the total subsidy value change if the positional accuracy of a particular subset of vertices were increased or decreased by amount X? Which vertex contributes most to the uncertainty in the total subsidy value?

8.1.5 How can implications of spatial data quality be calculated in the case of digging activities around underground cables and pipelines?

To reduce the chance of hitting an underground cable or pipeline, contractors spend time on detection. They continuously make the trade-off between risk and detection costs. Chapter 6 presents a model with which risk and detection costs can be calculated as a function of detection time. With this model, an optimal detection time can be calculated. The amount of time spent on detection depends strongly on spatial data on the location of underground cables and pipelines. Poor quality data can result in the contractor either spending more or less time than optimal on detection, resulting in sub-optimal costs for the contractor. The extra costs are then the difference between the optimal and sub-optimal costs. In the chapter it is shown how these extra costs can be quantified. In a case-study, the chapter illustrates how the model can be parameterised and presents some calculations of costs of incompleteness.

8.1.6 Which decision-theoretic factors determine the willingness of researchers and policy makers to spend resources on the quantification of implications of spatial data quality?

This question was answered in chapter 7. From open interviews with GIS/spatial data users in public decision-making we noted that the users concerned were well aware of spatial data quality issues. In many cases, they considered quantification of implications of spatial data quality either unnecessary or impossible. They deliberately chose to ignore spatial data quality and often used decision-theoretical arguments to motivate their choice. Using a questionnaire we further analysed these arguments. We found that respondents are less willing to spend resources on the quantification of implications of spatial data quality:

- If feedback on spatial data quality is received from stakeholders at the start of a decision-making process. Respondents expect that stakeholders will detect and inform them about errors in the spatial data and so reduce risks to an acceptable level;

We found that respondents are more willing to spend resources on the quantification of implications of spatial data quality:

- If the objective of the decision can be expressed numerically;
- If feedback on spatial data quality is received from co-decision-makers in the decision-taking phase, because there is more concern about the reliability of co-decision-makers' information;
- If feedback on spatial data quality is received at the end of the decision-making process, that is if the decision-maker can be held accountable for risks due to spatial data quality;

- If more is at stake.

8.2 Recommendations

8.2.1 Spatial data quality description

With regard to spatial data quality description, I would recommend:

1. To include in the element lineage also a detailed description of how reference data were collected;
2. Most spatial analyses and most spatial decisions are based on multiple data sets. To model the propagation of errors in a spatial analysis on multiple data sets, one needs to know how errors are correlated (see also chapter 4). Producers however only document the quality of their own data set and not the correlation with other data sets. With the huge number of data sets nowadays available, measurement of correlation between all data sets may appear a formidable task. Taking a pragmatic approach it is recommended that users and producers together assess which data sets are often co-analysed and then measure correlation only between those data sets. To measure the correlation, producers must coordinate their reference data collection campaigns so that reference data are collected at more or less the same location and time. Producers of land cover databases for monitoring purposes are recommended to collect reference data at the same locations for each time slice and report the results in change detection error matrices;
3. Fisher (1999, 2000) identified three forms of uncertainty: error, vagueness and ambiguity. Often, two or even three forms occur in the same data set, but most publications on spatial data quality are limited to one form. There is a need for models in which all three are quantitatively combined;
4. Producers are recommended to implement a spatial data quality standard and a metadata standard. The promised virtues of these standards can only be realised if they are actually implemented. Producers are recommended to engage researchers to learn from the experiences in adoption. Once adopted, it will be easier to realise tools that support the user in handling spatial data quality.

8.2.2 Spatial data quality application

With regard to the application of spatial data quality descriptions in fitness-for-use assessment, I would recommend:

1. On use of error matrices for monitoring purposes (chapter 4), two recommendations can be given: (1) test the effectiveness of the methods when they are calibrated from a sample reference data set rather than from an exhaustive reference data set and (2) proceed with research on error propagation analysis on categorical variables while taking into account both spatial and temporal correlation;
2. On uncertainty in area estimates (chapter 5), three recommendations can be given: (1) include in the model effects of fuzziness and generalisation, (2) include in the covariance equation various forms of correlation (3) check the validity of model assumptions with field measurements;
3. On cable and pipeline damage (chapter 6), the first recommendation is to collect more data so that the model can be used to support decision-making. The second

recommendation is to find a way of calculating optimum detection times for individual cables and pipelines, while taking into account that detection times will coincide if cables and pipelines are buried close together;

4. On the willingness to spend resources on the quantification of implications of spatial data quality (chapter 7), note that we researched perceptions on factors that increased or decreased the need to quantify implications. The next step in research should be to check the validity of these perceptions. Secondly, we identified a lack of empirical research into how users cope with spatial data quality. More research is needed in that area;
5. Although not systematically investigated, my experience over the past four years was that awareness of spatial data quality issues can especially be raised by quantification of implications as risks. I recommend continuing this line of research.

References

Aalders, H.G.J.L., 2002. The Registration of Quality in a GIS. In: Shi, W., Fisher P.F. and Goodchild M.F. (eds.). Spatial Data Quality. London: Taylor and Francis, pp. 186-299.

Achard, F., Eva, H. D., Stibig, H. J., Mayaux, P., Gallego, J., Richards, T, and Malingreau, J. P., 2002. Determination of Deforestation Rates of the World 's Humid Tropical Forests. Science 297(5583): 999-1002.

Agresti, A., 1990. Categorical data analysis. New York: Wiley, 558 p.

Agumya, A. and Hunter G.J., 2002. Responding to the consequences of uncertainty in geographical data. International Journal of Geographical Information Science 16 (5): 405-417.

Agumya, A. and Hunter, G.J., 1999a. A Risk-Based Approach to Assessing the 'Fitness for Use' of Spatial Data. URISA Journal, 11(1): 33-44.

Agumya, A. and Hunter, G.J., 1999b. Translating Uncertainty in Geographical Data into Risk in Decisions. In: Shi, W., Goodchild, M.F. and Fisher, P.F. (eds.): Proceedings of The International Symposium on Spatial Data Quality, 18-20 July 1999, Hong Kong, pp. 574-584.

Aronoff, S., 1989. Geographic information systems: a management perspective. Ottawa: WDL, 294 p.

Baker, W.L. and Cai, Y., 1992. The r.le programs for multiscale analysis of landscape structure using the GRASS geographical information system. Landscape Ecology, 7(4): 291-302.

Bierkens, M.F.P. and Burrough, P.A., 1993. The indicator approach to categorical soil data. Journal of Soil Science 44(2): 361-368.

Bio, A.F.M., 2000. Does vegetation suit our models? : data and model assumptions and the assessment of species distribution in space. Utrecht, Nederlandse Geografische Studies, The Royal Dutch Geographical Society, 195p.

Blachut, T.J., Chrzanowski, A. and Saastamoinen, J.H. , 1979. Urban surveying and mapping. New York: Springer, 372 p.

Bogaert, P., 2002. Spatial prediction of categorical variables: the Bayesian maximum entropy approach. Stochastic Environmental Research and Risk Assessment 16(6): 425-448.

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

Bondesson, L., G. Ståhl, and S. Holm. 1998. Standard errors of area estimates obtained by traversing and GPS. Forest Science, 44(3): 405-413.

Boschetti, L., Flasse S.P. and Brivio, P.A., 2004. Analysis of the conflict between omission and commission in low spatial resolution dichotomic thematic products: The Pareto Boundary. Remote Sensing of Environment, 91(3-4): 280-292.

Brassel, K., Bucher, F., Stephan, E.M and Vckovski, A., 1995. Completeness. In: Guptill, S.C. and Morrison, J.L. (eds.), Elements of spatial data quality. International Cartographic Association, Tokyo: Elsevier Science, pp. 81-108.

Brimicombe, A.J., 2003. GIS Environmental Modelling and Engineering. London: Taylor & Francis, 312 p.

Broos, M.J., Aarts, L., van Tooren, C.F. and Stein, A., 1999. Quantification of the effects of spatially varying environmental contaminants into a cost model for soil remediation. Journal of Environmental Management 56(2): 133-145

Buongiorno, J. and J.K. Gilless. 2003. Decision methods for forest resource managers. Academic Press, Amsterdam, The Netherlands. 439 p.

Burrough, P.A. and Frank, A.U. (eds.), 1996. Geographic objects with indeterminate boundaries. London: Taylor & Francis, 345 p.

Burrough, P.A. and McDonnell, R.A., 1998. Principles of Geographical Information Systems. Oxford UK: Oxford University Press, 333 p.

Burrough, P.A., 1992. Development of intelligent geographical information systems. International Journal of Geographical Information Science, 6(1): 1-11.

Burrough, P.A., 2001. GIS and geostatistics: Essential partners for spatial analysis. Environmental and Ecological Statistics 8(4): 361-377.

Burrough, P.A., van Gaans, P.F.M. and Hootsmans, R., 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77(2/4): 115-135.

Campbell, H., 1991. Organizational issues in managing geographical information. In: Masser, I. and Blakemore, M. (eds.), Handling geographical information. Longman/Wiley, 1991, pp. 259-282.

Carmel, Y., Dean, D. J. and, Flather, C. H., 2001. Combining location and classification error sources for estimating multi-temporal database accuracy. Photogrammetric Engineering and Remote Sensing, 67(7): 865-872.

CEN/TC 287, 1998. ENV 12656:1998 Geographic Information - Data description – Quality. 46 p.

Chrisman, N.R. and B.S. Yandell. 1988. Effects of point error on area calculations: A statistical model. Surveying and Mapping, 48(4): 241-246.

Chrisman, N.R., 1984. The role of quality information in the long-term functioning of a geographic information system. Cartographica, 21(2-3): 79-87.

Chrisman, N.R., 1987. A draft proposed standard for digital cartographic data quality, in H. Moellering (ed.), Report 8, NCDCDS, Columbus OH. [appeared unchanged, but unattributed in The American Cartographer (1988), pp. 129-135.

Chrisman, N.R., 1991. The error component in spatial data. In: Maguire, D.J., Goodchild, M.F and Rhind, D.W. (eds.), Geographical Information Systems: Principles and Applications. Volume 1, p. 165-174.

Collet, D., 1991. Modelling binary data. London: Chapman & Hall, 369 p.

Comber, A., Fisher, P. and Wadsworth, R.., 2004. Integrating land-cover data with different ontologies: identifying change from inconsistency. International Journal of Geographical Information Science, 18(7): 691-708

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing and Environment, 37(1): 35-46.

Cooke, R. and Jager, E., 1998. A probabilistic model for the failure frequency of underground gas pipelines. Risk Analysis 18(4): 511-527.

Cornélis, B. and Brunet S., 2002. A Policy-maker Point of View on Uncertainties in Spatial Decisions. In: Shi, W., Fisher P.F. and Goodchild M.F. (eds.), Spatial Data Quality. London: Taylor and Francis, pp. 168-185.

Cressie, N.A.C., 1991. Statistics for spatial data. New York: Wiley, 900 p.

Crompvoets J., Bregt A., Rajabifard A. and Williamson, I., 2004. Assessing the worldwide developments of national spatial data clearinghouses. International Journal of Geographical Information Science 18(7): 665-689

Crosetto, M. and Tarantola, S., 2001. Uncertainty and sensitivity analysis: Tools for GIS-based model implementation. International Journal of Geographical Information Science, 15(5): 415-437.

Dai X. L., and Khorram S., 1998. The effects of image misregistration on the accuracy of remotely sensed change detection, IEEE Transactions on Geoscience and Remote Sensing, 36(5): 1566-1577.

de Bruin S., Bregt A. and van de Ven, M., 2001. Assessing fitness for use: The expected value of spatial data sets. International Journal of Geographical Information Science, 15(5): 457-471.

de Bruin, S. and Hunter, G.J., 2003. Making the trade-off between decision quality and information cost. Photogrammetric Engineering & Remote Sensing 69(1): 91-98.

de Bruin, S., de Wit, A.J.W., van Oort, P.A.J. and Gorte, B.G.H., 2004. Using quadtree segmentation to support error modelling in categorical raster data. International Journal of Geographical Information Science. 18(2). 151-168

de Groeve, T. and K. Lowell. 2001. Boundary uncertainty assessment from a single forest-type map. Photogrammetric Engineering & Remote Sensing, 67(6): 717-726.

de Gruijter, J., 1999. Spatial sampling schemes for remote sensing. In: Stein, A., van der Meer, F. and Gorte, B. (eds.), Spatial Statistics for Remote Sensing. Dordrecht: Kluwer, pp. 211–284.

de Kruif, J., 2004a. De geo-info van kabels en leidingen Kabels en Leidingen Informatie Centrum (KLIC). Geo-info Nederland 1(2): 60-67. (in Dutch, with English abstract)

de Kruif, J., 2004b. Datakwaliteit is speerpunt bij Essent netwerk: de kabel- en leidinginformatie bij Essent is klaar voor de toekomst. Geo-info Nederland 1(4): 144-149. (in Dutch, with English abstract)

de Kruif, J., 2004c. Alles heeft zijn historie : grondroerder wordt soms verdrietig van de geleverde tekeningen. Geo-info Nederland 1(6): 280-288. (in Dutch, with English abstract)

de Wit, A.J.W., 2002 Homepage of the Dutch land use database http://cgi.girs.wageningen-ur.nl/cgi/projects/lgn/index_uk.htm

de Wit, A.J.W., van der Heijden, T.G.C. and Thunnissen, H.A.M., 1999. Vervaardiging en nauwkeurigheid van het LGN3-grondgebruiksbestand. Wageningen: DLO-Staring Centre, 84 p.

de Zeeuw, C. J., and Hazeu G. W. (eds.), 2001. Monitoring land use changes using geo-information. Possibilities, methods and adapted techniques. Alterra report 214, Alterra, Green World Research, Wageningen, 84 p.

de Zeeuw, C.J., Bregt, A.K., Sonneveld, M.P.W. and van den Brink, J.A.T., 1999. Geo-information for monitoring land use; From map overlay to object-structured noise reduction. Proceedings of the tenth international workshop on database and expert systems applications, 1-3 September 1999, Florence, Italy. IEEE, Los Alamitos, California, pp. 419-425.

Department of Commerce, 1992. Spatial Data Transfer Standard (SDTS) (Federal Information Processing Standard 173): Washington, Department of Commerce, National Institute of Standards and Technology.

Deutsch, C.V. and Journel, A.G., 1998. GSLIB: Geostatistical Software Library and User's Guide. Oxford University Press, New York.

Devillers, R., Bedard, Y. and Jeansoulin, R., 2005. Multidimensional management of geospatial data quality information for its dynamic use within GIS. Photogrammetric Engineering & Remote Sensing, 71(2): 205-215.

Dijkema, H. en Zweistra, R., 2001. Rechtspraak bij kabel- en leidingschades: onderaannemer staat juridisch zwak. Het loonbedrijf 54(12): 18-21. (in Dutch)

Doucette, M. and Paresi, C., 2000. Quality management in GDI. In: Groot, R. and MacLaughlin, J. (eds.), Geospatial data infrastructure : concepts, cases, and good practice. Oxford, Oxford University Press, pp. 85-96.

Duckham, M., 2002. A user-oriented perspective of error-sensitive GIS development. Transactions in GIS 6(2): 179-193

Epstein, E.F., Hunter, G.J. and Agumya, A., 1998. Liability insurance and the use of geographical information. International Journal of Geographical Information Science, 12 (3): 203-214.

FGDC (Federal Geographic Data Committee), 1998a. FGDC-STD-001-1998. Content standard for digital geospatial metadata (revised June 1998). Federal Geographic Data Committee. Washington, D.C.

FGDC (Federal Geographic Data Committee), 1998b. FGDC-STD-002-1998. Spatial Data Transfer Standard. Computer Products Office, National Technical Information Service, Springfield, VA.

FGDC (Federal Geographic Data Committee), 2004. ROADIC_Final_Report. www.fgdc.gov/fgdc/coorwg/2004/ROADIC_FGDC_presentation/ ROADIC_Final_Report.pdf, last visited 14 October 2005.

Fischhoff, B., Lichtenstein, S. and Slovic, P., 1981. Acceptable risk. Cambridge: Cambridge University Press, 185 p.

Fisher, P., 2000. Sorites paradox and vague geographies. Fuzzy Sets and Systems, 113(1): 7-18.

Fisher, P.F., 1997. Elements of Spatial Data Quality. Book review. International Journal of Geographical Information Science, 11(4): 407-408.

Fisher, P.F., 1999. Models of uncertainty in spatial data. In: Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (eds.), Geographical Information Systems: Principles and technical issues. New York: John Wiley & Sons, 2nd ed., vol. 1, pp. 191-205

Foody, G. M., 2002. Status of land cover classification accuracy assessment. Remote Sensing of Environment, 80(1): 185-201.

Foody, G.M. and Atkinson, P.M. (eds.), 2002. Uncertainty in Remote Sensing and GIS. Chicester: Wiley, 307 p.

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

Foody, G.M., Campbell, N.A., Trodd, N.M. and Wood, T.F., 1992. Derivation and applications of probabilistic measures of class membership from the maximum-likelihood classification. Photogrammetric Engineering and Remote Sensing 58(9): 1335-1341.

Forier, F. and F. Canters, 1996. A user-friendly tool for error modelling and error propagation in a GIS environment, Spatial Accuracy Assessment in Natural Resources and Environmental Sciences: Second International Symposium, May 21-23, Fort Collins, Colorado, pp. 225-234.

Forman, R.T.T. and Godron, M., 1986. Landscape Ecology. New York: John Wiley & Sons, 619 p.

Frank, A.U., 2001. Tiers of ontology and consistency constraints in geographical information systems. International Journal of Geographical Information Science 15(7): 667-678

Fuller, R.M, Smith, G.M. and Devereux, B.J., 2003. The characterisation and measurement of land cover change through remote sensing: Problems in operational applications? International Journal of Applied Earth Observation and Geoinformation 4(3): 243-253

Goodchild, M. F., 1995. Attribute accuracy. In: Guptill, S.C. and Morrison, J.L. (eds.), Elements of Spatial Data Quality. Amsterdam: Elsevier, pp. 59–79.

Goodchild, M.F. and Gopal, S. (eds.), 1989. Accuracy of Spatial Databases. London: Taylor & Francis, pp. 81-90

Goodchild, M.F. and Jeansoulin, R. (eds.), 1998. Data Quality in Geographic Information, from Error to Uncertainty. Paris: Hermes,

Goodchild, M.F., 2002. Introduction to part I: Theoretical Models for Uncertain GIS. In: Shi, W., Fisher P.F. and Goodchild M.F. (eds.), Spatial Data Quality. London: Taylor and Francis pp. 1-4

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, 483 p.

Griffith, D.A. 1989. Distance calculations and errors in geographic databases. P. 81-90 in The accuracy of spatial databases, Goodchild, M.F. and S. Gopal (eds.), 1989. Taylor & Francis, New York.

Gross, C.P. and P. Adler, 1996. Reliability of area mapping by delineation in aerial photographs. P. 267-271 in Spatial accuracy assessment in natural resources and environmental sciences: second international symposium, Mowrer et al. (eds.). Fort Collins, Colorado.

Guijt, W., 2004. Analyses of 30 years of incident data show US and European pipelines becoming safer. Oil & Gas Journal 102 (4): 68-73.

Guptill, S.C. and Morrison, J.L. (eds.), 1995. Elements of spatial data quality. International Cartographic Association, Tokyo: Elsevier Science, 202 p.

Haasbroek, N.D.,1968. Gemma Frisius, Tycho Brahe and Snellius and their triangulations. Delft: Rijkscommissie voor Geodesie, 119 p

Heit, M. and A. Shortreid (eds.). 1991. GIS applications in natural resources. GIS World, Fort Collins, Colorado. 381 p.

Heuvelink, G.B.M. and Burrough, P.A. (eds.), 2002. Developments in Statistical Approaches to Spatial Uncertainty and its Propagation. Themed issue International Journal of Geographical Information Science 16(2): 111-113

Heuvelink, G.B.M. and Lemmens, M.J.P.M (eds.), 2000. Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences. Delft: University Press, 772 p.

Heuvelink, G.B.M. and Lemmens, M.J.P.M. (eds.), 2000. Proceedings of the 4th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences. Delft: Delft University Press, 772 p.

Heuvelink, G.B.M., 1993. Error propagation in quantitative spatial modelling: applications in geographical information systems. Netherlands Geographical Studies No. 163, Utrecht: Koninklijk Nederlands Aardrijkskundig Genootschap, 151 p.

Heuvelink, G.B.M., 1998. Error propagation in environmental modelling with GIS. London: Taylor & Francis, 127 p.

Houghton, R.A., Hackler, J.L. and Lawrence K.T., 1999. The US carbon budget: Contributions from land-use change. Science 285(5427): 574-578.

Hunsaker, C.T., Goodchild M.F., Friedl M.A. and Case T.J. (eds.), 2001. Spatial Uncertainty in Ecology: Implications for Remote Sensing and GIS Applications. New York, Springer-Verlag, 402 p.

Hunter, G. and Lowell, K. (eds.), 2002. Accuracy 2002. Proceedings of the 5th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Melbourne, Australia, 547 p.

Hunter, G.J., B. Hock, M. Robey and M.F. Goodchild. 1996. Experimental development of a model of vector data uncertainty. In: Mowrer et al. (eds.), Spatial accuracy assessment in natural resources and environmental sciences: second international symposium. Fort Collins, Colorado, pp. 217-224.

Husch, B., T.W. Beers and J.A. Kershaw. 2003. Forest mensuration. John Wiley and Sons, Hoboken, New Jersey. 443 p.

Isaaks, E.H. and Srivastava, R.M., 1990. Applied Geostatistics. Oxford University Press, 561 p.

ISO (International Standardisation Organisation), 2002. ISO 19113:2002 Geographic information — Quality principles, 29 p.

ISO (International Standardisation Organisation), 2003a. ISO 19114:2003 Geographic information — Quality evaluation procedures, 63 p.

ISO (International Standardisation Organisation), 2003b. ISO 19115:2003 Geographic information — Metadata, 140 p.

Jaeger, C.C., Renn, O., Rosa, E.A. and Webler, T., 2001. Risk, uncertainty, and rational action. London: Earthscan, 320 p.

Jakobsson, A. and Vauglin, F., 2001. Status of Data Quality in European National Mapping Agencies. Proceeding of the 20th International Cartographic Conference, Volume 4, pp. 2875-2883.

Janssen, L.L.F. and van der Wel, F.J.M., 1994. Accuracy Assessment of Satellite Derived Land-Cover Data: A review. Photogrammetric Engineering and Remote Sensing, 60(4): 419-426.

Jeong, H.S. and Abraham, D.M., 2004. A decision tool for the selection of imaging technologies to detect underground infrastructure. Tunnelling and Underground Space Technology 19(2): 175-191.

Keefer, B.J., J.L. Smith and T.G. Gregoire. 1991. Modeling and evaluating the effects of stream mode digitizing errors on map variables. Photogrammetric Engineering & Remote Sensing, 57(7): 957-963.

Kiiveri, H. T., Caccetta, P., and Evans, F., 2001. Use of conditional probability networks for environmental monitoring. International Journal of Remote Sensing, 22(7): 1173-1190.

Kramer, H. and Knol, W.C., 2004. Historische geodata : een nieuwe dimensie! Het grondgebruik vanaf 1850 ontsloten als geodata. Geo-info Nederland, 1(6): 244-251 (in Dutch, with English summary)

Kresse, W. and Fadaie, K., 2004. ISO standards for geographic information. Berlin: Springer, 322 p.

Laurini, R. and Thompson, D., 1992. Fundamentals of spatial information systems. London: Academic Press, 680 p.

Leung, Y. and J. Yan. 1998. A locational error model for spatial features. International Journal of Geographical Information Science, 12(6): 607-620.

Leyk, S., 2005. Computing the past – Utilizing Historical Data Sources for Map-Based Retrospective landscape Research. University of Zurich, PhD thesis, 172 p.

Li, H. and Reynolds, J.F., 1993. A new contagion index to quantify spatial patterns of landscapes. Landscape Ecology, 8(3): 155-162.

Li, Y., Brimicombe, A.J. and Ralphs, M.P., 2000. Spatial data quality and sensitivity analysis in GIS and environmental modelling: The case of coastal oil spills. Computers Environment and Urban Systems, 24(2): 95-108.

Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W., 1999. Introduction to Data Quality. In: Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (eds.), Geographical Information Systems: Principles and technical issues. New York: John Wiley & Sons, 2nd ed., vol. 1, pp. 175-176.

Lowell, K. and Jaton, A. (eds.), 1999. Spatial Accuracy Assessment. Land Information Uncertainty in Natural Resources. Chelsea, Michigan: Ann Arbor Press, 455 p.

Lunetta, R. S., Ediriwickrema, J., Johnson, D. M., Lyon J. G., and McKerrow, A., 2002. Impacts of vegetation dynamics on the identification of land-cover change in a biologically complex community in North Carolina, USA. Remote Sensing of Environment, 82(2-3): 258-270.

Macleod, R.D., and Congalton, R.G., 1998. A quantitative comparison of change detection algorithms for monitoring eelgrass from remotely sensed data. Photogrammetric Engineering and Remote Sensing, 64(3): 207–216.

Magnussen, S., 1996. A coordinate-free area variance estimator for forest stands with a fuzzy outline. Forest Science, 42(1): 76-85.

Maling, D.H., 1989. Measurements from maps. New York: Pergamon Press, 577 p.

Mas, J.F., 1999. Monitoring land-cover changes: a comparison of change detection techniques. International Journal of Remote Sensing, 20(1): 139-152.

Matheron, G.,1971. The Theory of Regionalised Variables and its Applications, Cahier No. 5, Centre de Morphologie Mathématique de Fontainebleau, 211 p.

McGwire, K.C. and Fisher, P., 2001. Spatially Variable Thematic Accuracy: Beyond the Confusion Matrix. In: Hunsaker, C.T., Goodchild M.F., Friedl M.A. and Case T.J. (eds.), Spatial Uncertainty in Ecology: Implications for Remote Sensing and GIS Applications. New York, Springer-Verlag, pp. 308-329.

Meeus TJ en Schoof R., 2003. Wachten op de grote klap: Niemand is verantwoordelijk voor het ondergrondse labyrint in Nederland. NRC Handelsblad 19 april 2003. (in Dutch)

Moellering, H. (ed.), 1988. The Proposed Standard for Digital Cartographic Data. The American Cartographer 15(1)

Molenaar, M., 1998. An introduction to the theory of spatial object modelling for GIS. London: Taylor & Francis, 246 p.

Morgan, M.G., Henrion, M., and Small, M., 1990. Uncertainty : a guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge: Cambridge University Press, 332 p.

Morrison, J.L., 1995. Spatial data quality. In: Guptill, S.C. and Morrison, J.L. (eds.), Elements of spatial data quality. International Cartographic Association, Tokyo: Elsevier Science, pp. 1-12.

Mowrer, T.H. and Congalton, R.G., 2000. Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing. London: Taylor & Francis, 350 p.

Muhlbauer, W.K., 2004. Pipeline Risk Management Manual: Ideas, Techniques, and Resources - 3rd Edition. Amsterdam, The Netherlands: Elsevier, 395 p.

Næsset, E. 1999. Effects of delineation errors in forest stand boundaries on estimated area and timber volumes. Scandinavian Journal of Forest Research, 14(6): 558-566.

Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W., 1996. Applied linear statistical models : regression, analysis of variance, and experimental designs – 4th ed. Boston: U.S.A.: McGraw-Hill, 1408 p.

O'Hagan, A., 1994. Kendall's advanced theory of statistics, Vol. 2B: Bayesian inference New York: Halsted, 330 p.

O'Neill, R.V., Krummel, J.R., Gardner, R.H., Sugihara, G., Jackson, B., Deangelis, D.L., Milne, B.T., Turner, M.G., Zygmunt, B., Christensen, S.W., Dale, V.H. and Graham, R.L., 1988. Indices of landscape pattern. Landscape Ecology, 1(3): 153-162.

Openshaw, S., 1989. Learning to live with error in spatial databases. In: Goodchild, M.F. and Gopal, S. (eds.), Accuracy of Spatial Databases. London: Taylor & Francis, pp. 263-276.

Östman, A., 1997. The specification and evaluation of spatial data quality. In: Proceedings of the 18th ICA International Cartographic Conference, Sweden, (Gävle: Swedish Cartographic Society), pp. 836-847.

Pauwels H.J.M.B. en Wieleman R.W., 2004. Verplichte Informatie-uitwisseling Ondergrondse Kabels en Leidingen: Graven naar informatie. Delft: The Netherlands: Nederlands Normalisatie Instituut, 143 p. (in Dutch)

Pebesma, E.J., Duin, R.N.M. and Burrough P.A., 2005. Mapping Sea Bird Densities over the North Sea: Spatially Aggregated Estimates and Temporal Changes. Environmentrics 16(6): 573-587.

Petit, C.C., and Lambin, E.F., 2002. Impact of data integration technique on historical land-use/land-cover change: Comparing historical maps with remote sensing data in the Belgian Ardennes. Landscape Ecology, 17(2): 117-132.

Polman, J. and Salzmann, M.A., 1996. Handleiding voor de technische werkzaamheden van het Kadaster. Apeldoorn: Kadaster, 685 p. (in Dutch)

Power C., Simms A., and White R., 2001. Hierarchical fuzzy pattern matching for the regional comparison of land use maps. International Journal of Geographical Information Science, 15(1): 77-100.

Prisley, S.P., T.G. Gregoire and J.L. Smith. 1989. The mean and variance of area estimates computed in an arc-node geographic information system. Photogrammetric Engineering & Remote Sensing, 55(11): 1601-1612.

Raper, J.F., 1999. Spatial representation: the scientist's perspective. In: Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (eds.), Geographical Information Systems: Principles and technical issues. New York: John Wiley & Sons, 2nd ed., vol. 1, pp. 61-70

Riiters, K.H., O'Neill, R.V., Hunsaker, C.T., Wickham, J.D., Yankee, D.H., Timmins, S.P., Jones, K.B. and Jackson, B.L., 1995. A factor analysis of landscape pattern and structure metrics. Landscape Ecology, 10(1): 23-39.

Rosenfield, G.H., 1982. Sample Design for Estimating Change in Land Use and Land Cover. Photogrammetric Engineering and Remote Sensing, 84(5): 793-801.

Shi, W., Fisher, P.F. and Goodchild, M, F. (eds.), 2002. Spatial Data Quality. London: Taylor & Francis, 313 p.

Singh, A., 1989. Digital change detection techniques using remotely-sensed data, International Journal of Remote Sensing, 10(6): 989-1003.

Smith, J.H., Wickham, J.D., Stehman, S.V. and Yang, L., 2002. Impacts of patch size and land-cover heterogeneity on thematic image classification accuracy. Photogrammetric Engineering and Remote Sensing, 68(1): 65-70.

Smith, J.H., Wickham, J.D., Stehman, S.V. and Yang, L., 2003. Effects of landscape characteristics on land-cover class accuracy. Remote Sensing of Environment, 84(3): 342-349.

Smith, J.H., Wickham, J.D., Stehman, S.V. and Yang, L., 2003. Effects of landscape characteristics on land-cover class accuracy. Remote Sensing of Environment, 84(3): 342-349.

Sonneveld, M.P.W., van den Brink, J.A.T., de Zeeuw, C.J., and Bregt, A.K., 2000. Noise reduction in GIS-based land use monitoring. Proceedings of the ninth international symposium on spatial data handling, 10-12 August 2000, Beijing, China. Section 5a, pp. 44-56.

Steele, B.M., Winne J.C. and Redmond, R.L., 1998. Estimation and mapping of misclassification probabilities for thematic land cover maps. Remote Sensing of Environment 66(2): 192-202.

Stein, A. and van Oort, P., in press. Concepts of handling spatio-temporal data quality for cost functions. In: Devillers, R. and Jeansoulin, R. (eds.), Spatial Data Quality. An Introduction. Hermes Science Publishing Ltd, London, 320 p. (expected publication date: march 2006)

Stein, A., Staritsky, I., Bouma, J. and van Groeninen, J.W., 1995. Interactive GIS for environmental risk assessment. International journal of geographical information systems 9(5): 509-525

Stow, D.A., and Chen, D.M., 2002. Sensitivity of multitemporal NOAA AVHRR data of an urbanizing region to land-use/land-cover changes and misregistration. Remote Sensing of Environment, 80(2): 297– 307.

Thunnissen, H.A.M. and de Wit, A.J.W., 2000. The national land cover database of the Netherlands. In Geoinformation for all; XIX congress of the International Society for Photogrammetry and Remote Sensing (Lemmer: GITC): 223-230.

Thunnissen, H.A.M. and Noordman, E., 1996. National land cover database of The Netherlands : classification methodology and operational implementation. Delft: BCRS, 95 p.

Townshend, J. R. G., Justice, C. O., Gurney, C., and McManus, J., 1992. The impact of misregistration on change detection, IEEE Transactions on Geoscience and Remote Sensing. 30(5): 1054-1060

True, W.R., 1998. European pipeline performance improving, spill study shows. Oil & Gas Journal 96 (49): 53-57

U.S. DoT (U.S. Department of Transportation) - Office of Pipeline Safety, 1999. Common Ground: Study of One-Call Systems and Damage Prevention Best Practices, 262 p.

Uitermark, H.T.J.A., 2001. Ontology-based geographic data set integration. PhD thesis University of Twente, The Netherlands, 139 p.

van Asperen, P., 1997. Controlling the process: digital map revision experiences from Holland. Geomatics Info Magazine, 11(6): 76-79.

van der Smagt, A. G. M. and Lucardie, G. L., 1991. Decision making under not well-defined conditions: from data processing to logical modelling. Journal of Economic and Social Geography, 82(4): 295-304.

van der Stok, T., 2005. Overheid neemt regie in handen: spelregels moeten kabel- en leidingschades verminderen. Het loonbedrijf 57(1): 11-13. (in Dutch)

van der Wel, F.J.M., 2000. Assessment and visualisation of uncertainty in remote sensing land cover classifications. Phd. Thesis, Utrecht university, 200 p.

van Houten, Y.A. en Lourens, P.F., 1995. Preventie van gasleidingbeschadigingen onderzocht vanuit een gedragswetenschappelijk kader. Verkeerskundig Studie Centrum –Rijksuniversiteit Groningen, 112 p. (in Dutch)

van Niel T.G. and T.R. McVicar. 2001. Assessing positional accuracy and its effects on rice crop area measurement: an application at Coleambally Irrigation Area. Australian Journal of Experimental Agriculture, 41(4): 557-566.

van Oort, P.A.J. and Bregt, A.K., in press. Do users ignore spatial data quality? – a decision-theoretic perspective. Accepted for publication in Risk Analysis.

van Oort, P.A.J., 2004. Kwaliteit van geo-informatie in kabels en leidingen : risico-analyse kan zorgen voor een optimale kwaliteit van geo-informatie. Geo-info Nederland 1(12): 528-533 (in Dutch, with English abstract)

van Oort, P.A.J., 2005. Improving land cover change estimates by accounting for classification errors. International Journal of Remote Sensing 26(14): 3009-3024

van Oort, P.A.J., Bregt, A.K. and de Bruin, S., submitted. Detection and risk for digging activities around underground cables and pipelines: implications of spatial data quality. Submitted for publication to Transactions in GIS on 19 September 2005

van Oort, P.A.J., Bregt, A.K., de Bruin, S., de Wit, A.J.W. and Stein, A., 2004. Spatial variability in classification accuracy of agricultural crops in the Dutch national land-cover database. International Journal of Geographical Information Science, 18(6), 611-626.

van Oort, P.A.J., Stein, A., Bregt, A.K., de Bruin, S. and Kuipers, J., 2005. A variance and covariance equation for area estimates with a GIS. Forest Science 51(4): 347-356.

van Soest, F., Stein, A., Dekkers, A.L.M. and van Duijvenbooden, W., 2001. A quantitative evaluation of monitoring networks for region-specific nitrate reduction policies. Journal of Environmental Management, 61(3): 215-225.

Vauglin, F., 2002. A Practical Study on Precision and Resolution in Vector Geographical Databases. In: Shi, W., Fisher P.F. and Goodchild M.F. (eds.). Spatial Data Quality. London: Taylor and Francis pp. 127-139

VELIN, 2004. Registratie en analyse van pijpleiding incidenten 1999 tot en met 2003 - eerste voortgangsverslag projectgroep incidenten reductie oktober 2004. VELIN, Tilburg, the Netherlands, 15 p.

Verbyla, D. L., and Boles, S.H., 2000. Bias in land cover change estimates due to misregistration. International Journal of Remote sensing 21(18): 3553–3560.

Veregin, H., 1989. Error modelling for the map overlay operation. In: Goodchild, M.F. and Gopal, S. (eds.), Accuracy of Spatial Databases. London: Taylor & Francis), pp. 3-18.

Veregin, H., 1999. Data Quality Parameters. In: Longley, P.A., Goodchild, M.F., Rhind D. and Maguire, D. (eds.), Geographical Information Systems, Principles and Applications – Volume 1, Principles and Technical Issues, 2nd edition. New York: John Wiley & Sons, pp. 177-189.

Vitousek, P. M., 1994. Beyond global warming: ecology and global change. Ecology 75(7): 1861-1876.

Worboys, M., 1998. Imprecision in finite resolution spatial data. GeoInformatica, 2(3): 257-279.

Abstract

The growing availability of spatial data along with growing ease to use the spatial data (thanks to wide-scale adoption of GIS) have made it possible to use spatial data in applications inappropriate considering the quality of the data. As a result, concerns about spatial data quality have increased. To deal with these concerns, it is necessary to (1) formalise and standardise descriptions of spatial data quality and (2) to apply these descriptions in assessing the suitability (fitness for use) of spatial data, before using the data. The aim of this thesis was twofold: (1) to enhance the description of spatial data quality and (2) to improve our understanding of the implications of spatial data quality.

Chapter 1 sets the scene with a discussion on uncertainty and an explanation of why concerns about spatial data quality exist. Knowledge gaps are identified and the chapter concludes with six research questions.

Chapter 2 presents an overview of definitions of spatial data quality. Overall, I found a strong agreement on which elements together define spatial data quality. Definitions appear to differ in two aspects: (1) the location within the meta-data report: some elements occur not in the spatial data quality section but in another section of the meta-data report; and (2) the explicitness with which elements are recognised as individual elements. For example, the European pre-standard explicitly recognises the element ‘homogeneity’. Other standards recognise the importance of documenting the variation in quality, without naming it explicitly as an individual element.

In chapter 3 we quantified the spatial variability in classification accuracy for the agricultural crops in the Dutch national land cover database (LGN). Classification accuracy was significantly correlated with: (1) the crop present according to LGN, (2) the homogeneity of the 8-cell neighbourhood around each cell, (3) the size of the patch in which a cell is located, and (4) the heterogeneity of the landscape in which a cell is located.

In chapter 4 I present methods that use error matrices and change detection error matrices as input to make more accurate land cover change estimates. It was shown that temporal correlation in classification errors has a significant impact and must be taken into account. Producers of time series land cover data are recommended not only to report error matrices, but also change detection error matrices.

Chapter 5 focuses on positional accuracy and area estimates. From the positional accuracy of vertices delineating polygons, the variance and covariance in area can be derived. Earlier studies derived equations for the variance, this chapter presents a covariance equation. The variance and covariance equation were implemented in a model and applied in a case-study. The case-study consisted of 97 polygons with a small subsidy value (in euros per hectare) assigned to each polygon. With the model we could calculate the uncertainty in the total subsidy value (in euros) of the complete set of polygons as a consequence of uncertainty in the position of vertices.

Chapter 6 explores the relationship between completeness of spatial data and risk in digging activities around underground cables and pipelines. A model is presented for calculating the economic implications of over- and incompleteness. An important element of this model is the relationship between detection time and costs. The model can be used to calculate the optimal detection time, i.e. the time at which expected costs are at their minimum.

Chapter 7 addresses the question why risk analysis (RA) is so rarely applied to assess the suitability of spatial data prior to using the data. In theory, the use of RA is beneficial because it allows the user to judge if the use of certain spatial data does not produce unacceptable risks. Frequently proposed hypotheses explaining the scarce adoption of RA are all technical and educational. In chapter 7 we propose a new group of hypotheses, based on decision theory. We found that the willingness to spend resources on RA depends (1) on the presence of feedback mechanisms in the decision-making process, (2) on how much is at stake and (3) to a minor extent on how well the decision-making process can be modelled.

Chapter 8 presents conclusions on the six research questions (chapters 2-7) and lists recommendations for users, producers and researchers of spatial data. With regard to the description, four recommendations are given. Firstly, spend more effort on documenting the lineage of reference data. Secondly, quantify and report correlation of quality between related data sets. Thirdly, investigate the integration of different forms of uncertainty (error, vagueness, ambiguity). Fourthly, study the implementation and use of spatial data quality standards. With regard to the application of spatial data quality descriptions, I have two main recommendations. Firstly, to continue the line of research followed in this thesis: quantification of implications of spatial data quality, through development of theory along with tangible illustrations in case-studies. Secondly, there is a need for more empirical research into how users cope with spatial data quality.

Samenvatting

Sinds de jaren 1960 is de hoeveelheid beschikbare geo-informatie sterk toegenomen; met Geografische Informatie Systemen (GIS) kan deze geo-informatie gebruikt worden in een veelheid aan toepassingen, ook in toepassingen waarvoor de kwaliteit van de informatie eigenlijk onvoldoende is. Om tegemoet te komen aan zorgen over kwaliteit van geo- informatie zijn nodig: (1) de beschrijving van kwaliteit van geo- informatie door middel van duidelijke en gestandaardiseerde definities en (2) de geschiktheid van de geo-informatie beoordelen voordat de geo-informatie wordt gebruikt in toepassingen. De doelen van dit proefschrift waren: (1) onderzoek naar de beschrijving van kwaliteit van geo-informatie (2) inzicht verkrijgen in de implicaties van kwaliteit van geo-informatie.

Hoofdstuk 1 beschrijft verschillende vormen van onzekerheid en verklaart de toegenomen interesse in de kwaliteit van geo-informatie. Het schetst de leemtes in de huidige kennis en eindigt met zes onderzoeksvragen.

Hoofdstuk 2 biedt een overzicht van diverse definities van het begrip kwaliteit van geo-informatie. Ik vond grote overeenstemming tussen de vergeleken definities. De definities bleken op twee onderdelen te verschillen. Ten eerste de locatie binnen de meta-data: niet alle elementen van kwaliteit worden noodzakelijkerwijs beschreven in de kwaliteit sectie van de meta-data. Ten tweede de mate waarin elementen met naam genoemd worden. Bijvoorbeeld het element “homogeniteit”, ofwel variatie in de kwaliteit, wordt in de Europese voornorm expliciet genoemd. In andere standaarden wordt het belang van dit element wel onderkend, maar het wordt niet expliciet als een element genoemd.

Voor hoofdstuk 3 kwantificeerden wij voor de landbouwgewassen in het Landelijk Grondgebruiksbestand Nederland (LGN) de ruimtelijke variatie in classificatie nauwkeurigheid. De classificatie nauwkeurigheid bleek gecorreleerd met: (1) het volgens LGN aanwezige landbouwgewas, (2) de homogeniteit van de 8 grid cellen rondom iedere grid cel, (3) de grootte van het perceel waarin een grid cel ligt en (4) de heterogeniteit van het landschap waarin een grid cel ligt.

In hoofdstuk 4 worden methoden gepresenteerd om, gebruik makend van fouten-matrices, nauwkeurigere schattingen van veranderingen in landgebruik te maken. Het komt relatief vaak voor dat op twee tijdstippen dezelfde misclassificatie optreedt. Het is belangrijk om daar rekening mee te houden en hoofdstuk 4 presenteert ook de daarvoor geschikte methoden.

Hoofdstuk 5 gaat over positionele nauwkeurigheid en oppervlakte schattingen. Vertices beschrijven de omtrek van polygonen. Met behulp van fouten-voortplanting kunnen vergelijkingen voor de variantie en covariantie van oppervlakte schattingen worden afgeleid, als functie van onzekerheid in de coördinaten van de vertices. Eerdere studies presenteerden variantie vergelijkingen, dit hoofdstuk presenteert voor het eerst ook een covariantie vergelijking. Beide vergelijkingen zijn ingevoerd in een model en toegepast in een case-studie. De case-studie bestond uit 97 polygonen met ieder een kleine subsidie (in euros per hectare). Met behulp van het model konden wij de onzekerheid in het totale subsidie bedrag (in euros) berekenen.

Hoofdstuk 6 legt het verband tussen de volledigheid van geo-informatie en de graafschade risico’s bij graafwerkzaamheden nabij ondergrondse kabels en leidingen. In het hoofdstuk presenteren wij een model waarmee de financiële implicaties van over- en incompleetheid berekend kunnen worden. Belangrijk onderdeel van dit model

is de relatie tussen detectie tijd en kosten. Met het model kan de optimale detectie tijd (de tijd waarbij de verwachte kosten het laagst zijn) berekend worden.

Hoofdstuk 7 gaat in op de vraag waarom zo zelden fouten-voortplanting en risico-analyse (RA) worden gebruikt vooraf aan het gebruik van geo-informatie. Theorie suggereert dat RA verstandig zou zijn omdat het de gebruiker in staat stelt om te beoordelen of gebruik van bepaalde geo-informatie leidt tot onaanvaardbare risicos. De meeste hypotheses voor het schaarse gebruik van RA zijn van technische aard of gerelateerd aan kennis van de gebruiker. In hoofdstuk 7 zijn hypotheses op het bestuurlijke vlak geformuleerd en getoetst. Daaruit bleek dat de bereidheid om geld en tijd te spenderen aan RA afhangt van (1) terugkoppelings mechanismen in het besluitvormingsproces; (2) hoeveel er op het spel staat en (3) in mindere mate van hoe goed de beslissing in een model te beschrijven is.

Hoofdstuk 8 geeft antwoorden op de zes onderzoeksvragen (hoofdstukken 2-7) en doet aanbevelingen voor gebruikers, producenten en onderzoekers van geo-informatie. Ten aanzien van de beschrijving van kwaliteit worden vier aanbevelingen gedaan: (1) meer aandacht voor de beschrijving van hoe referentie data werden verkregen (2) onderzoek naar de correlatie tussen fouten in verschillende data sets, (3) onderzoek naar methoden om meerdere vormen van onzekerheid (error, vagueness, ambiguity) in een model op te nemen en (4) onderzoek naar de implementatie van standaarden. Ten aanzien van het gebruik van kwaliteits- beschrijvingen voor het bepalen van geschiktheid voor gebruik heb ik twee aanbevelingen. Ten eerste, doorgaan op de in dit proefschrift ingeslagen weg: het kwantificeren van implicaties van kwaliteit van geo-informatie, door ontwikkeling van theorie en parallel daaraan aansprekende toepassing van de theorie in case-studies. Mijn tweede aanbeveling is om meer empirisch onderzoek te doen naar hoe gebruikers omgaan met kwaliteit van geo-informatie.

Curriculum vitae

My name is Pepijn Adrianus Johannes van Oort. I was born on the 9th of February 1977, in Oirschot, The Netherlands. From 1995 to 2001 I studied Tropical Land Use at Wageningen Agricultural University. During this study, my interest shifted from development work to mathematics, statistics and geographical information science. It accumulated in an internship at the SysNet project at the International Rice Research Institute (IRRI) in the Philippines, a thesis at the Centre for Geo-Information (CGI) where I built a spatially explicit model for simulating land cover change and a thesis with the Plant Production Systems (PPS) group where I successfully developed a model of nutrient flows in grass-clover swards under various grassland management options. After receiving my MSc, I worked for two months in the Stability of Coalitions (STACO) project, in which coalition-formation in Kyoto negotiations was simulated using an approach based on game-theory.

From 2001 to 2005 I worked on my PhD thesis at the CGI. The thesis, which you have hopefully just read after arriving at this section, addresses a broad range of topics on spatial data quality. Results were published in a large number of peer reviewed journals, in a Dutch professional journal and occasionally I was co-author in publications of colleagues. Supported by the C.T. de Wit graduate school for Production Ecology and Resource Conservation (PE&RC) I have presented at national and international conferences and taken courses in statistics and modelling. Within the same graduate school I was have acted for three years as the coordinator of a PhD discussion group on statistics, mathematics and modelling. My interests are broad, but a certain wish is to build further on the work laid down in this thesis. If you are interested, feel free to contact me.

List of publications

Peer reviewed journals

van Oort, P.A.J., Bregt, A.K., de Bruin, S., de Wit, A.J.W. and Stein, A., 2004. Spatial variability in classification accuracy of agricultural crops in the Dutch national land-cover database. International Journal of Geographical Information Science, 18(6), 611-626.

van Oort, P.A.J., 2005. Improving land cover change estimates by accounting for classification errors. International Journal of Remote Sensing 26(14): 3009-3024

van Oort, P.A.J., Stein, A., Bregt, A.K., de Bruin, S. and Kuipers, J., 2005. A variance and covariance equation for area estimates with a GIS. Forest Science 51(4): 347-356.

van Oort, P.A.J. and Bregt, A.K., 2005. Do users ignore spatial data quality? - a decision-theoretic perspective. Risk Analysis 25(6).

de Bruin, S., de Wit, A.J.W., van Oort, P.A.J. and Gorte, B.G.H., 2004. Using quadtree segmentation to support error modelling in categorical raster data. International Journal of Geographical Information Science 18(2): 151-168.

Peer reviewed, submitted publications

Crompvoets, J., de Bree, F., van Oort, P.A.J., Bregt, A., Wachowicz, M., Rajabifard, B. and Williamson, I., submitted. Worldwide impact assessment of spatial data clearinghouses. Submitted to Computers, Environment and Urban Systems

van Oort, P.A.J., Bregt, A.K. and de Bruin, S., submitted. Detection and risk for digging activities around underground cables and pipelines: implications of spatial data quality. Submitted to Transactions in GIS

Other publications

Roetter R, Laborte A.G., van Oort, P, Hoanh C.T., Cabrera J.M.C.A., Lucas, M., Francisco, S. and van Keulen, H, 1999. Resource-use analysis at regional scale: Explorations for rice systems in SE Asia. CD-ROM Proceedings SAAD-3, CIP, Lima Peru, Nov. 8-10, 1999.

van Oort, P., 2004. W. Kresse, K. Fadaie, et al. (Eds.), ISO Standards for Geographic Information. Book review. International Journal of Applied Earth Observation and Geoinformation 6(2): 159-160.

van Oort, P.A.J., 2004. Kwaliteit van geo-informatie in kabels en leidingen: risico-analyse kan zorgen voor een optimale kwaliteit van geo-informatie. Geo-info Nederland 1(12): 528-533, (in Dutch, English summary.

Stein, A. and van Oort, P.A.J., in press. Concepts of handling spatio-temporal data quality for cost functions. In: Devillers, R. and Jeansoulin, R. (eds.), Spatial Data Quality. An Introduction. Hermes Science Publishing Ltd, London, 320 p. (expected publication date: March 2006).