Complex Analysis Methods in Hybrid GIS Using Uncertain Data

Michael Glemser, Ulrike Klein, Dieter Fritsch, Stuttgart, Günter Strunz, Oberpfaffenhofen

Zusammenfassung: Komplexe Analysemethoden in hybriden GIS unter Berücksichtigung von Genauigkeitsangaben

Thema des Beitrags ist die Berücksichtigung der Datenunsicherheit in komplexen Analyseprozessen. Am Beispiel einer Überschwemmungsvorhersage wird erläutert, welche Erweiterungen in der Datenmodellierung und in den Analysefunktionen dazu notwendig sind. Ziel ist es, die Unsicherheit in der räumlichen Ausdehnung der prognostizierten Überschwemmungsfläche zu bestimmen. Insbesondere die Auswirkungen auf Siedlungsgebiete sind von besonderem Interesse. Ein Vergleich mit dem traditionellen Ergebnis ohne Berücksichtigung der Unsicherheit verdeutlicht den erzielten Informationsgewinn.

Abstract

Topic of this contribution is the integration of data uncertainty in complex analysis processes. By means of a flooding forecast as an example application, the enhancements which are necessary in data modelling and analysis functions are explained. The major objective of the research is to determine the uncertainty in the spatial extension of the possible flooded region. In particular, the impact on settlement areas is of special interest. A comparison with the traditional result without consideration of uncertainty clarifies the information profit achieved.

1 Introduction

Over the years, GIS have evolved to powerful tools which solve various tasks in different applications. The ongoing integration of new and improved analysis methods enables users to process even more complex application tasks with combinations of geometrical, topological and thematic aspects using hybrid data, i. e. raster data as well as vector data. A further step in development will be achieved by the opening of the architecture to interoperable systems (McKee and Kuhn, 1997, OGC, 1999, IOGIS, 1999). In such an open environment, the problems are distributed over different systems which act like single components. The final result is obtained in this teamwork-like approach by a combination where each component adds its part.

Data uncertainty plays a special role in such an environment. Although all data is uncertain to a particular degree, the necessary integration in GIS has not taken place so far. The consequence is that uncertainty is actually not taken into account during data processing. This kind of disregard can partly be justified only in the case of a closed system, where the user has full control over all steps from input to presentation. As the user is aware of the data quality, he is able to verify the result using his knowledge and experience. This might be seen as a reason why uncertainty has not been a relevant topic in GIS yet. But this control gets completely lost in an interoperable system. This starts already within the data input step. Today, a great number of existing databases offer a variety of data sets covering different thematic aspects like topographic information, cadastral data, statistical data or digital orthophotos and satellite images. Data collection is changing widely from digitizing own data to retrieving and transferring from existing data bases. For such data sets, the user requires an uncertainty description which has to be added by the producer of the data set as a kind of meta description. This information allows the data to be checked in order to avoid the risk of combinations of data without any practical use. Furthermore, the uncertainty description serves as additional input parameter for the analysis process and thus enables the propagation of the uncertainty onto the result. For this purpose an appropriate uncertainty model has to be developed and integrated in GIS. In particular, the analysis methods require to be enhanced in order to propagate the uncertainty automatically as a kind of background process.

The emphasis of the research is on the description and propagation of the uncertainty through analysis in a GIS. In this article, the uncertainty integration is demonstrated by an example application in order to allow the reader a simple access to the developed approach. As an example application, the forecast of the spatial extent of a possible flooding is selected. The special task to be solved is to detect those parts of settlement areas which are flooded with respect to an assumed high water level. It defines a complex problem in the way that the whole process consists of two successive steps. The first step is based on a flooding model that has to be developed. In the second step, the derived outcome is geometrically overlaid with topographic information, here settlement areas. The analysis is performed in a study area (0.8 by 1.2 qkm) that is located near Stuttgart in the Enz Valley. Figure 4 shows an orthophoto of this area overlaid with settlement and water information. As a result of the analysis not only a geometrical representation of the potentially flooded settlement is requested as it would be obtained by a traditional analysis. The major goal is to determine the uncertainty of the result in addition to the traditional output.

The following sections start with a discussion about the used example data sets. They include raster data as well as vector data. Therefore their management and integration require the definition of a hybrid data model.

Afterwards, an analysis model is presented which solves the previously described task in the ordinary way without taking uncertainty into account. The next section defines an approach to an uncertainty model that allows the integration and management of the included uncertainty. The approach covers both spatial aspects, geometry and attributes. With respect to that model, the used analysis methods are modified and the results of the uncertainty propagation are presented and discussed. Some conclusions and a brief outlook close this contribution.

2 Data Sets

Basically, the analysis requires two different input data sets. The topographic surface strongly influences which areas are flooded. It can be described by a digital terrain model (DTM). As source of the flood, the location of the river is needed as well as the location of the settlement areas which might be in particular affected by the flood. The necessary landuse information can be taken from a topographic data set.

The DTM for the study area is shown in figure 1a. The character of the terrain varies considerably in the test area. A perspective view (figure 1b) gives a good impression about it. Steep slopes with almost 100 m in height difference can be found as well as wide flat areas at the bottom of the valley. The height difference along the course of the river is about 3 m. The heights are structured in a regular grid and therefore they form a raster data set with a ground pixel size of 5 m. The DTM was manually measured on a digital photogrammetric workstation using two digital aerial images. The accuracy of the heights is approximately 1 m. The exact values depend on the slope of the surface at each position. Although the operator tried to work very accurate, he faced some problems during measurement. For example, it was almost impossible to measure the exact height of the river bank because of its dense coverage with trees and bushes. Point density was another problem: While the sampling in open regions (agricultural areas) is regular, the distribution of points in settlement areas follows the given road structure. The DTM is used for the rectification of the aerial images in order to produce the orthophoto which is often used in the figures (e. g. figure 4) as background information. It has a resolution of 0.5 m on the ground.



Figure 1: Digital terrain model (DTM) with 5 m resolution (a) and a perspective view overlaid with the orthophoto (b)

The landuse data set (figure 2) is extracted from an existing digital topographic database, called ATKIS, which is the German topographic cartographic spatial database (ATKIS, 1997). The database defines a digital model of the physical landscape. The relevant contents of the landscape are modelled as objects which are successively aggregated to object classes (e. g. roads), object groups (e. g. road transportation) and finally to object domains (e. g. transportation). The scale is 1:25.000. All data is collected in vector format with a positional accuracy of 3 m for important objects (e. g. roads, railways, rivers) and a reduced accuracy of 10 m for all others. For the flooding analysis two object classes are relevant: stream and residential area. The study area contains one stream object (the river Enz) and five residential areas that build together the settlement area of interest.



Figure 2: Landuse data set derived from existing topographic database (ATKIS)

A comparison of the landuse data set with the real world situation can be derived from the overlay with the orthophoto (figure 4). The existing boundaries coincide quite well in the limits of their accuracy, although the orthophoto also reveals some misclassifications. Obviously some objects are missing or residential areas are significantly larger than mapped ones due to new buildings. Despite the differences, the original data sets are taken for granted and remain unchanged in the scope of this research.



Figure 3: Object oriented hybrid data model in UML-syntax (Booch et al., 1999)

In the present application both fundamental data types in GIS, raster and vector data, occur. Since they are used together in one analysis step, the data model should be able to manage both in an integrated manner. The integration is achieved by utilising a hybrid data model (Molenaar and Fritsch, 1991). A schematic graphical representation of the developed model can be seen in figure 3. The model is given in UML-syntax (Booch et al., 1999). The approach is based on objects as exclusive representations of spatial phenomena. As a consequence an object building procedure has to be applied in some cases before the data can be integrated (Klein et al., 1998).

The geometric component allows either a raster or vector representation of the object. Furthermore it is possible to attach multiple attributes of any kind to it. A hybrid model enables the user to process data without ever noticing whether the actual type is raster or vector. A user has the advantage that he can fully concentrate on the analysis part of the application.

3 Complex Analysis

The forecast of the spatial extent of a possible flooding is an example for a complex analysis combining a thematic problem with a geometrical operation. For a town situated near a river the area endangered by an increasing water surface has to be determined. The process is divided in two successive steps. In the first step the spatial extension of the flooded area has to be determined. Input data set is the DTM, available in raster format. The height given in each raster cell describes the thematic content of the raster cell by a continuous attribute value. Thus the determination of the flooded area can be considered as a thematic analysis. In the second step the derived flooded region is geometrically overlaid with the settlement areas. Since the settlement areas are given in vector format, the overlay is realised as hybrid method.

3.1 Flood Model

The determination of the flooded area requires the formulation of a flood model. The data required for this thematic problem are the digital height model, the course of the river and an assumption about the flooding height.



settlement area 🔃 river 🔛 flooded area

Assuming that the river has no or only a very small gradient, a quite simple model leads to the flooded area. In each raster cell it has to be examined whether the height of the raster cell is greater or smaller than the height of the river incremented by the assumed increase of water. The raster cells lying below the water surface are flooded and grouped together to the object flooded area. Since this case occurs extremely rare in reality, a more complex flood model has to be formulated. Therefore, for each raster cell of the river it is examined how the increasing water-level affects the neighbouring raster cells. Thus, an individual flooded area is determined for each river cell and the overlay of these partial areas results in the complete area. It is necessary to start this investigation at the deepest part of the river. The result of the forecast of the flooding according to the defined flood model is presented in figure 4. Since the settlement areas are visualised additionally, the endangered regions are already recognisable. In the next step they have to be determined as independent objects using an overlay operation.

Figure 4: Outcome of the simulated flooding

3.2 Hybrid Polygon Overlay

Polygon overlay is a basic functionality which is offered by most GIS products. It takes two sets of input objects and geometrically intersects them. The result is a set of new objects consisting only of the intersection parts but possessing attributes of all the input objects. It is easier to implement the overlay operation with raster data than with vector data. In the raster domain it is a simple boolean And-operator on the raster cells, while in the vector domain the computation of points of intersection is very time-consuming. To take advantage of this fact in hybrid systems, polygon overlay is realised using raster overlay techniques. Therefore the settlement areas given in vector format have to be transformed to objects in raster format by a vector-to-raster conversion. The realisation of a hybrid GIS requires that the conversions run automatically as a kind of background operation without any user interaction, even without being noticed by the user (Fritsch et al., 1998).



Figure 5: Traditional outcome of the overlay (flooded settlement area)

The outcome of the overlay between the intersecting objects – settlement areas and flooded region - is shown in figure 5. The shapes of the five resulting objects can be divided into areas and lines. For a better identification, the regions with the three line objects are shown in an enlarged inset of the orthophoto. The line objects arise when settlement areas and flooded region touch or have a very small overlay. With respect to the result of the analysis such sliver polygons are considered as irrelevant and removed from the database. However, it is doubtful if doing so is warrantable. Since all data is uncertain to a certain degree, the spatial extension of the objects respectively their boundary is to be considered to be uncertain, too. To obtain reliable results for the analysis the uncertainty of the objects needs to be modelled and integrated in all operations.

4 Uncertainty Model

The main goal of this research is to integrate the uncertainty into the data processing and to show the influences on the results. Therefore a description of the initial uncertainty of the different input data sets is required which is subsequently discussed for the two input data sets (DTM and landuse data set) with respect to their different data types (raster and vector data).

As shown before, the DTM given in raster format, takes part in the flooding analysis. A raster data set is formed in the way that the geometric component is controlled because of the definition of a strict raster structure and the attributes are kept variable (Chrisman, 1991). As the raster allows no variation in the geometry, this component can be seen as constant and error-free. In contrast to that, variable attributes indicate that the values have to be measured. Measurements are one of the well-known sources of uncertainty (Burrough, 1986) because they can only be performed with limited accuracy due to resolution of the equipment and human operator interactions. In the present example of the DTM, each raster cell carries a height value as an attribute. Heights belong to a continuous scale, thus they are able to vary along this scale. A stochastic approach is suitable to model this kind of variation. It treats each height as an independent random variable that varies according to a distribution function. In the case of the heights, a Gaussian distribution is assumed. Mean value and variance are the parameters representing the variation in this approach. These two parameters have to be determined for each raster cell. The attribute values which are already collected can be seen as appropriate estimates of the mean. Only the variances require additional effort. A simple approach would be to assume the same height variance for each cell. A more complex determination is based on the dependency of the variance on the slope. Then the variance σ_h^2 can be calculated in the following way:

$\sigma_h = a + b \cdot \tan \alpha$

with *a* as a constant and *b* as a slope dependent part (slope represented through the tangens of the height angle α). Either a comparison with reference data or the existence of prior knowledge are possible ways to estimate the values of *a* and *b*. Just as the mean values, the variances can also be managed in raster form. A variance raster arises that represents the uncertainty in this approach. Figure 6 illustrates the variances of the given DTM for the slope dependent approach. For the purpose of visualisation the absolute values are transformed into qualitative grey values. The darker the grey value the higher the variance and vice versa. The rough character is caused by the slopes. It reveals the lineage and the internal structure of the DTM which was originally derived by a triangulation.



Figure 6: Uncertainty of the DTM data set given by variances of the heights. Variances are transformed into grey values (white = small variance, dark = large variance).

The outcome of the first analysis step, the flooding, is an object which shows the membership to the possible flooding area. The geometry of the object represents the spatial extent of the flooding. Membership defines a discrete attribute with only two valid values: membership and no membership. Uncertainty is expressed for discrete attributes as the degree of truth instead of a measure of variation. The truth can be described with a probability value, i. e. the probability that the assigned attribute value is true. Variances and probabilities are closely related and based on the same theory, the stochastic theory. This facilitates the definition of an universal uncertainty model. A visualisation of the probabilities of the flooded region is found in figure 9.

Beside the flooding object, the settlement areas build a second input data set for the overlay analysis. This data set is given in vector format. In contrast to raster data, the thematical component is controlled (attribute = landuse, value = settlement) and the geometry is let variable (Chrisman, 1991). The user has to measure the spatial extension of the controlled theme, so the geometry is uncertain. Again the uncertainty is introduced by the used measurement methods. But also the fuzziness of the phenomenon plays an important role (Burrough, 1996). Both influences add up to the complete amount of uncertainty. Uncertainty in geometry refers to the variation in position of geometric primitives (points, lines, areas). One approach to model the uncertainty is based on the well-known epsilon-band where an error-band (possessing a width of 2ε) is laid around each primitive (Chrisman, 1982). It is assumed that the true position of the primitive can be found somewhere inside the band. More compatible to the treatment of the thematic uncertainty is to define the primitives as random

variables and to describe them with the same measures (mean and variance of the position) as already known for continuous attributes (Glemser and Fritsch, 1998). The variance for the present data set is provided by prior knowledge. Settlement areas belong to the group of ordinary ATKIS object classes and therefore they possess a standard deviation of 10 m.

The uncertainty of the flooding object is given as probabilities, while the settlement areas are described by variances. But propagation of uncertainty during the overlay step requires a homogeneous representation. It is proposed to choose probabilities as the common basis because the transformation is much more easier to process from variances to probabilities than in the other direction. Probabilities can be calculated for every position in space indicating the membership to the object. The values depend on the local distribution function of the geometric primitive and the distance measured between the certain position and the object (Kraus and Haussteiner, 1993). The formulas for calculation of the probabilities vary for different primitives (Fritsch et al., 1998). For areas, as needed here, the formula is given by:

$$p_{area}(x, y) = p(d) = \int_{-\infty}^{d} f(t) dt = F(d)$$

with $f(t) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2}$ and $\sigma = 10m$

with *d* as the distance of the position (x, y) to the object (*d* has a positive value inside the object), f(t) as the density function and F(d) as the distribution function. The formula defines a spatially continuous probability function which has to be approximated by a discrete raster. Then probabilities are calculated for each cell. The raster forms a probability matrix that can also be used to produce a good graphical visualisation of the uncertainty (Glemser and Fritsch, 1998). Figure 7 shows the probabilities for the settlement areas.



Figure 7: Graphical representation of the probabilities of the settlement data set.

An uncertainty model is derived based on the discussion about the different uncertainty descriptions. Figure 8 gives a graphical representation of the model.



Figure 8: Uncertainty model in UML-syntax (Booch et al., 1999)

The model allows to add uncertainty measures to the geometry as well as to each attribute. The description can either be done by probabilities or variances. So both aspects are covered: the variation and the truth of values. It includes also the transformation between variances and probabilities. The model builds the basis for the propagation of the uncertainty in the following section.

5 Uncertainty Propagation

Since all data only have a limited accuracy it is important for the user to know how the result of the analysis can be assessed with respect to its accuracy. This problem can only be adequately solved if the uncertainty of the input data sets is integrated in the analysis process and propagated onto the result. Theoretical basis is the derived uncertainty model (section 4) which is based on the stochastic approach. Within the thematic analysis the uncertainty of the flooded region has to be estimated. To this end a probability value is calculated for every raster cell indicating the membership to the object flooded area. Since the overlay is implemented for raster data the geometrical representation of all objects including their uncertainty description are required in raster format. In the example application this concerns the settlement area given in vector format. The variances given for the boundaries have to be transformed into probabilities for every raster cell indicating the membership to the objects settlement area (already discussed in section 4). The objects given in raster format and their probability matrices form the basis for the overlay.

5.1 Flood Model

The task to be solved is to detect those parts of the terrain which are flooded with respect to a assumed high water level. A major difficulty is that the functional context between the digital elevation model and the flooded region can only be modelled in a complex form. Thus, the variance propagation based on a functional context is not suitable to determine the uncertainty description of the flooded area. Simulation techniques are used alternatively to solve the problem.

Simulation can be described as generation of realisations of a mathematical system by variation of the input variables. If stochastic data are used, simulation is called stochastic as well. In our application this concerns the generation of the flooded region due to the stochastic behaviour of the elevation model. The problem can be solved by using Monte Carlo methods. Monte Carlo methods are special procedures for the simulation of stochastic problems generating random samples by random numbers. The stochastic behaviour of the samples results from a predefined probability density function. For a given stochastic input variable the outcomes of the simulation are the realisations of a random variable as well. The digital elevation model is the basis of the simulation considering the heights as random variables. Assuming a Gaussian or Normal distribution, the heights are described sufficiently by height value and variance.

The determination of the uncertainty of the flooded region is achieved as follows: in a first step n realisations of the elevation model are generated using the probability density function for each height. Then, according to the formulated flood model (section 3.1), the spatial extent of a possible flooding is determined for each of the n realisations. The outcome of the simulation are n different and independent realisations of the flooded region. For a representative conclusion from the sample to the population the volume n of the sample has to be large enough (e. g. n = 1000). The quality of the derived statement about the uncertainty of the flooded area depends primarily on the volume of the sample. The more realisations are produced, the more reliable is the result - but computing effort increases at the same time.

For the analysis of the iterations the n flooded regions are overlaid and for every raster cell the frequency of flooding is determined. Frequency can be interpreted as probability which indicates the membership of the cell to the object flooded region. The complete probability matrix can be considered as outcome of the performed simulation. Grouping all cells with probabilities larger than 0.0 results in the maximum extent of the flooded area (figure 9).



Figure 9: Graphical representation of the probabilities of the flood region

Because of the integration of uncertainty, the simulation result covers a larger area in comparison with the traditional solution. Not only a modified extension is derived, but also the probability of a flooding is spatially quantified. The graphical representation of the probabilities shows clearly that the uncertainty of the heights causes the uncertainty of the boundary, namely the spatial extent of the flooded area.

5.2 Polygon Overlay

The overlay of two sets of input objects results in a set of new objects consisting only of the intersection parts but possessing all attributes of the input objects. The combination of the attributes requires the integration and the propagation of their uncertainty. Since the overlay operation can be interpreted as a simple boolean Andoperator, the propagation of the probabilities is defined by

$$p(A_1 \cap A_2) = p(A_1) \cdot p(A_2)$$

 $p(A_1)$ and $p(A_2)$ denote the probabilities of the two raster cells possessing the assigned attribute values A_1 and A_2 .

The comparison of figure 10 with figure 5 shows that the integration of the uncertainty results in a larger endangered area. Therefore the sliver polygons can no longer be considered as irrelevant and removed from the database. Disregarding the uncertainty leads to a misinterpretation of the results. This example application shows clearly that it is important to integrate the uncertainty in data processing as well as to include the uncertainty into the visualisation of the results.



Figure 10: Graphical representation of the probabilities of the flooded settlement area

6 Conclusion and Outlook

The paper discusses the influence of data uncertainty on the outcome of a complex analysis process. All investigations are carried out within the example application of a flooding forecast but the presented approach is much more general in the sense that it can be applied to almost any operation in a hybrid GIS. In order to integrate the uncertainty of the input data sets, an uncertainty model is developed which is able to describe thematic as well as geometrical uncertainty of the data. During analysis processes, either a simulation or a mathematical propagation defined by the used analysis function perform the estimation of the requested uncertainty. In connection with interoperable applications, it is essential that all methods generate not only a result but also an uncertainty information belonging to it which supports the correct and reliable interpretation of the result. A second source of uncertainty can be detected in the analysis model applied. A decisive characteristic of a model is that it defines only a kind of approximation of the reality. For example, the used flooding model does not correspond exactly to the real situation of a flooding because of the complexity of the process itself. This aspect has not been covered so far. The further research work will concentrate on this aspect and will try to find a way to integrate model uncertainty as well.

Acknowledgement

This research work is founded by the German Research Foundation (DFG) within the joint project IOGIS (Interoperable Geowissenschaftliche Informationssysteme).

References

ATKIS (1997): Amtlich Topographisches-Kartographisches Informationssystem (ATKIS). Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV). <u>http://www.atkis.de</u>.

Booch, G., Rumbaugh, J., Jacobson, I. (1999): The Unified Modeling Language User Guide. Addison-Wesley.

- Burrough, P. A. (1986): Principles of Geographical Information Systems for Land Resources Assessment. Oxford Science Publications, Clarendon Press, Oxford.
- Burrough, P. A. (1996): Natural Objects with Indeterminate Boundaries. ESF-GISDATA, Vol. 2., Taylor & Francis.

- Chrisman, N. R. (1982): A Theory of Cartographic Error and its Measurement in Digital Databases. Proceedings of Auto-Carto 5, Crystal City, 159-168.
- Chrisman, N. R. (1991): The Error Component in Spatial Data. In: Maguire, D., Goodchild, M., Rhind. D. (Eds.): Geographical Information Systems – Principles and Applications. Longman Scientific & Technical, 165-174.
- Fritsch, D., Glemser, M., Klein, U., Sester, M., Strunz, G. (1998): Zur Integration von Unsicherheit bei Vektorund Rasterdaten. GIS, Vol. 11, Heft 4, 26-.35
- Glemser, M., Fritsch, D. (1998): Data Uncertainty in a Hybrid GIS. In: IAPRS Vol. 32, Part 4, Proceedings of ISPRS Commission IV Symposium, Stuttgart, Germany, 180-187.
- IOGIS (1999): Interoperable Offene Geowissenschaftliche Informationssysteme (IOGIS). http://www.ifgi.unimuenster.de/3_projekte/4dgis/texte/iogis/IOGIS.html.
- Klein, U., Sester, M., Strunz, G. (1998): Segmentation of Remotely Sensed Images Based on the Uncertainty of Multispectral Classification. In: IAPRS Vol. 32, Part 4, Proceedings of ISPRS Commission IV Symposium, Stuttgart, Germany, 299-305.
- Kraus, K. Haussteiner, K. (1993): Visualisierung der Genauigkeit geometrischer Daten. GIS, Vol. 6, Heft 3, 26-32.
- McKee, L., Kuhn, W. (1997): The OpenGIS Consortium's Purpose and Technical Approach. In: Fritsch D., Hobbie, D. (Eds.): Photogrammetric Week '97, Wichmann, Heidelberg.
- Molenaar, M., Fritsch, D. (1991): Combined Data Structures for Vector and Raster Representations in Geographic Information Systems. GIS, Vol. 4, Heft 3, 26-32.

OGC (1999): Open GIS Consortium (OGC). http://www.opengis.org.

Autors

Dipl.-Ing. MICHAEL GLEMSER, Wissenschaftlicher Mitarbeiter am Institut für Photogrammetrie der Universität Stuttgart, E-Mail: Michael.Glemser@ifp.uni-stuttgart.de.

Dr.-Ing. ULRIKE KLEIN, Wissenschaftliche Mitarbeiterin am Institut für Photogrammetrie der Universität Stuttgart, E-Mail: Ulrike.Klein@ifp.uni-stuttgart.de.

Anschrift: Institut für Photogrammetrie, Geschwister-Scholl-Straße 24, 70174 Stuttgart.