Data Quality: A Model for Resolvable Objects

Thomas K. Windholz & Kate M. Beard
Department of Spatial Information Science and Engineering,
University of Maine, Orono
E-mail:twindhol@spatial.maine.edu,
beard@spatial.maine.edu

Michael F. Goodchild
Department of Geography,
University of California, Santa Barbara
E-mail: good@ncgia.ucsb.edu

Abstract

The paper elaborates on the possibilities of determining resolution within spatial datasets. In the discussion we use the term resolution as an indicator for the ability to identify a certain object within a given GIS representation. This interpretation of the term resolution combines some properties of the photographic heritage related to the degree of discernable detail, and some of the properties inferred by the scale of a paper map–the users expectation to identify specific features at a certain scale. Resolution is a source of uncertainty as it constrains both what we can observe and represent. Without a model and measures of resolution we can not formulate a measure of what may be missing from a spatial representation. A brief clarification of terms associated with data quality lead the way to existing problems and a subsequent discussion of our suggested approach. The model considers the combined resolution of spatial and thematic dimensions–we consider the case of resolving "objects" in "fields". When generating a three-dimensional representation of the residuals (stored representation vs. higher accuracy) we obtain a relief map showing the minimal determinable variations–which can be used to detect the minimal size of a resolvable object. Thus, the resolvability of a spatial object can be determined by a function of the spatial extension of an object, its attribute value, and the three-dimensional relief of the inherent accuracy of the thematic representation. "Objects" could be patches of higher concentration, density, etc. A large patch may not be resolvable if its attribute value is weak compared to the accuracy of the "field" representation. The paper includes a case study of a sea surface temperature dataset collected off the coast of Maine. The approach for the case study is focused on–but not limited to–the investigation of the properties of resolution when generating kriged maps (i.e., a continuous raster-based representation) from a given sample dataset. We investigate the resulting differences in the ability to detect a certain object by reducing the number of sample points. Based on the residuals from a comparison of a kriged map versus a representation that is accepted as being ground truth (which could also be generated by applying simulation algorithms). The result shows a way for visualizing resolution–as a presentation of one of the inherent uncertainties to the GIS user. Furthermore, the model provides the user with the possibility to analyze a stored representation for the ability to detect an object of a certain spatial extension (i.e., x,y-coordinates) and a given attribute value.

1. Introduction

The computer is a finite system. We cannot duplicate the infinite real world. Any representation stored in a Geographic Information System (GIS) is imperfect. The quality of the data within a GIS depends on several different components (e.g., inaccuracy, inconsistency and resolution). Consequently, any query result—generated by this GIS—introduces a level of uncertainty about the state of the quality of the outcome. It is important to provide the GIS user with the necessary awareness that these problems exist. Although there is a growing interest in improving data quality standards (CEN 1995, ISO 1998, SDTS 1994) commercial GIS packages put little or no effort in calculating and communicating the inherent imperfections to the user. In literature (Beard 1996, Chrisman 1983, Goodchild 1989, 1993, Guptill 1995, Heuvelink 1993, 1998, Hunter 1991, Parsons 1996), however, we can find several different approaches handling either a single imperfection (e.g., inaccuracy) or a conglomerate of imperfections (e.g., imprecision and inconsistency).

1.1. Problem Statement

There are several methods of generating a GIS map. One of them is to generate a raster representation of a continuous variable (e.g., sea surface temperature). For example, we could sample the variable and then generate a kriged map. Then we could ask the question if the resulting map is "good" enough for the purpose of finding an object of a certain spatial extension that is of the same variable within that field representation (e.g., an area of warmer water with an extension of one square mile). Thus, this paper investigates a model that provides the GIS user with the necessary tools to judge the quality of a stored map with respect to its ability to identify a certain object in a continuos field representation.

1.2. Terminology—From Scale and Resolution to Detect-ability

In general one can say that storing a certain representation within a GIS requires a model of the real world at a meaningful (for a specific purpose) scale and at a meaningful—in respect to a chosen scale—resolution. This representation cannot be identical with the real world and thus introduces imperfections (e.g., inaccuracies). However, in order to avoid any confusions with the used terminology in this paper we want to clarify the term resolution and detect-ability.

The stated problem of deciding whether one is able to detect a specific object within a given field representation is dependent on the combined imperfections within the represented area—one of these components could be interpreted as the resolution of the stored field representation (see section 3. The Model—How to Determine Detect-ability). Here the term resolution is a combination of some properties of the photographic heritage—the definition: the degree of discernable detail (Gonzales and Woods, 1993)—and some of the decisive properties inferred by the scale of a paper map—the users expectation to identify specific features at a certain scale. On the other hand we could also use the term "level of geographic detail", which is discussed by Goodchild and Proctor (1997) as a possible augmentation of the term "scale" in the digital geographic world.

Since some of the terminology is used differently in different disciplines we do not want to re-use terms like resolution or scale for the model introduced in this paper. Thus, we introduce another term, detect-ability, which combines properties of the field (that can be seen as aspects of resolution) and properties of the object. Their distinct dependencies (e.g., sample size or object size) are explained in more detail in the following section (2. Dependencies of Detect-ability).

2. Dependencies of Detect-ability

This section is a discussion on the parameters that influence the outcome of the question where (within the field representation) one can identify objects—we refer to this as the dependencies of detect-ability. An intuitive approach to this question suggests that there are two main components influencing the results. On the one hand there is the field representation and on the other hand there is the object. However, here we are interested in a more detailed list (**Figure 1.**).

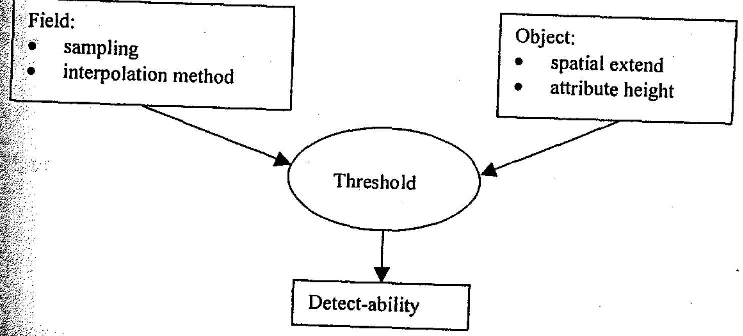

Figure 1. Dependencies of Detect-ability

First let us take a look at the decisive parameters when generating the field (Peuquet, 1999) representation, where we assume that we start with sampling the underlying variable. Sampling itself introduces two aspects: a) the number of sample points and b) their distribution. The next step when aiming at generating a continuous coverage is deciding which interpolation process (e.g., universal kriging) to choose. The final field representation will differ if any of these three components vary. Some of the results are more accurate than others (e.g., more sample points) and some of them will be smoother compared to others—depending on the interpolation method. Moreover, the accuracy or smoothness of the representation distinguishes the influence on the decision we want to make on whether we can detect an object or not.

Second, we would like to focus on some properties of the object itself. There are two components that are of interest when formulating its detect-ability within a field representation: a) the spatial extent of the object and b) the attribute height (or strength) of the object. Assuming that an object within a field would show a compact outline, its spatial extensions can be given by a single value, namely by its area in square units. The attribute height of the object is in the same units as the field representation and is a relative comparison to its neighborhood. The influence on its detect-ability results in the following facts. For an object showing a small spatial extend we can say that it will be detected easier with increasing attribute height. On the other hand an object having a small attribute height its detect-ability will increase when enlarging its spatial extend.

The third dependency is given by the threshold. The threshold determines the percentage of the object that has to be visible for its detection and it could be varied—up to a certain degree of freedom—by a GIS user. The determination of the visibility and thus, detect-ability is discussed in the following section.

3. The Model—How to Determine Detect-ability

The applied method for generating a representation (e.g., sampling followed by kriging) introduces some constraints on the level of detail that one is able to provide within the GIS. In this section we discuss a model that results in a binary map that identifies areas where a certain object can be determined and where not.

3.1. Approach

The model is based on the residuals calculated by subtracting the generated field representation from the ground truth. For an implementation we can substitute ground truth with any layer that we accept as being true. This could either be a comparable representation of higher accuracy (if available) or multiple (e.g., n = 100) generations of simulated realizations using conditional simulation (e.g., Sequential Gaussian Simulation). The residuals can then be seen as a result of a) the sample method and b) the model effects inherent in the interpolation method used to generate the field representation (e.g., kriging). The residuals represent an indicator of how well the representation matches reality—where one could say that this is the accuracy of the map. This is one way of interpreting these residuals, however, here we are looking beyond the numeric information, where we consider the spatial distribution of the values of the residuals. These residuals can be used to determine the detect-ability (or resolution) of a given representation.

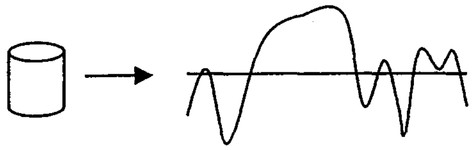

Figure 2. Object representation and relief map of the residuals

Within the field representation—at any given location—one cannot determine a feature occurring in the real world if the variations are smaller than the residuals. Let us take this idea a step further: if we generate a three dimensional representation of the residuals we obtain a relief map (similar to a DEM) of the minimal determinable variations. Looking at this relief map we can now specify a representative object and compare it to the outlines of the relief map (**Figure 2.**). If the object is hidden by the relief map then we say it cannot be detected from the kriged map representation. On the other hand if the object is fully visible on the outside of the relief map than we would be able to determine the object within the kriged map.

Next we would like to discuss the generation of a representative object for the comparison mentioned above. We suggest a representative object in the form of a cylinder. This is a result of the fact that a circle is the most compact form and that the height is a parallel movement of the objects outline. The radius of the circle is determined by the spatial extension of the object (e.g., we want to identify an object that has an area of π square units than the radius of the cylinder would equal 1 unit). The height of the cylinder represents the attribute value.

Finally we could combine the relief map with the cylinder. In order to determine the areas of possible detect-ability the cylinder is moved over the relief map. At each location of the cylinder (i.e., representation of the object) we now have to determine whether the top of the cylinder extends beyond the relief map (i.e., inherent inaccuracies/noise) or not. If the top of the cylinder is visible we can infer that an object located at this position would not be covered by the inaccuracies and thus, be detectable. However, we can say that if the spatial extend of the cylinder is represented by, for example, 100 pixels it is still sufficient enough to see 99 pixels in order to detect the cylinder. Thus, the introduction of a threshold for the detect-ability allows a percentage (e.g., 5%) of the cylinder to be obscured by the relief map.

Figure 3. shows a schematic representation of calculating the detect-ability from the relief map and a representative object. The result is a binary map, where areas of positive detect-ability (i.e., the object can be detected) are marked white and areas of negative detect-ability (i.e., the object cannot be detected) are marked black. The areas refer to the center of the object. Thus, if parts of a given object are within a black area, but it is centered within a white area, we would still be able to detect the object. Regarding the visualization of the resulting binary map it might be better to represent areas of positive detect-ability green and areas of negative detect-ability red. These color settings might improve the communication of the inherent imperfections to the GIS user.

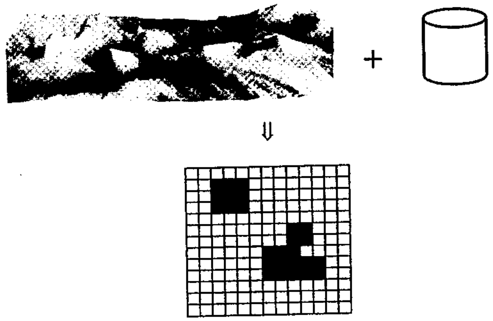

Figure 3. Schematic representation of the moving object and the resulting binary map

Table 1. Implied inferences

	Certain about statement	Uncertain about statement
Object is in observed area	white pixel □	black pixel ■
Object is *not* in observed area	white pixel □	black pixel ■

The resulting binary map needs some more discussion in order to clarify the inferences we can make about the areas of positive (white) and negative (black) detect-ability. For the white areas we can say that whether an object is present or not, the field representation is "good" enough to state that we are certain about the represented facts (i.e., there is an object or not). Whereas, for the black areas we have to state that the field representation does not allow us to make any inferences about the existence or non-existence of the defined object. Thus, all inferences made about objects within a black area introduce uncertainty in any derivations made from these field representations. This relationship is shown in **Table 1.**

3.2. Applications

In this section we take a closer look at some interesting applications of the discussed approach. In general one can divide the applications into two major categories. On the one hand there are those applications where the whole area of interest is already sampled or where—in addition to sampling—the kriged map is already generated. Here the model would be able to tell the user if the quality of the representation is high enough to derive conclusions with a desired certainty. The model could also be used to determine the appropriate sample size for a specific purpose (i.e., detecting objects of a certain size).

First, we would like to discuss issues of examples where the whole study area has been sampled. Applications could be the identification of, for example, warm core rings (i.e., warmer water pools), which would lead to a different ecological system within a cold water area. This phenomenon occurs in the Gulf of Maine when warm core rings get separated from the Gulf-stream. The size of these separations have to fulfill minimum requirements regarding their spatial extent in order to have an impact on the ecological system. The issue is to prove that the change in an ecological system was initialized by one of these pools. Thus, it is of interest to have the ability to say—with certainty—that there was no such object (i.e., pool) within a given field representation (i.e., map of sea surface temperature generated from sample points—see section 4. Case Study). Another application could be the detection of patches of high concentration of soil pollution in a rural area. This case introduces another interesting aspect, where operators of a chemical plant might have an interest to prove—with certainty—that there are no high concentrations of soil pollution in a specific sub-area. Thus, here we deal with a legal issue to prove that a map is fit for the specific purpose.

Second, a slight modification of the discussed model could be used to determine whether a proposed sample size is efficient for detecting a certain object prior to sampling the whole area of interest. Here the problem is more focused on the determination of whether the combination of the applied methods (i.e., sampling and interpolation method) will yield a sufficiently accurate field representation. The first step would require collecting sample points within a predefined sub-area, where objects do not necessarily have to be located. Then, at arbitrary locations within the sub-area, perturbations of the size of the given object would be introduced. Finally an application of the suggested model to determine the detect-ability would clarify if the applied methods (i.e., sampling and interpolation method) are sufficiently accurate. If there are any black areas in the resulting binary map, changes are necessary (e.g., increasing the sample size). This method would require the implementation of conditional simulations—as discussed earlier.

4. Case Study

In this case study we want to determine whether we can detect pools (with a radius of about 10km) of different water temperature (e.g., ±2°C and ±5°C) using sample points and kriging.

4.1. The used Data

On the one hand we use a satellite image showing the sea surface temperature (**Figure 4.**) in the Gulf of Maine and on the other hand we use a set of 231 sample points within the area shown in **Figure 4.** The sample points follow a regular distribution with a spacing of about 20km between them.

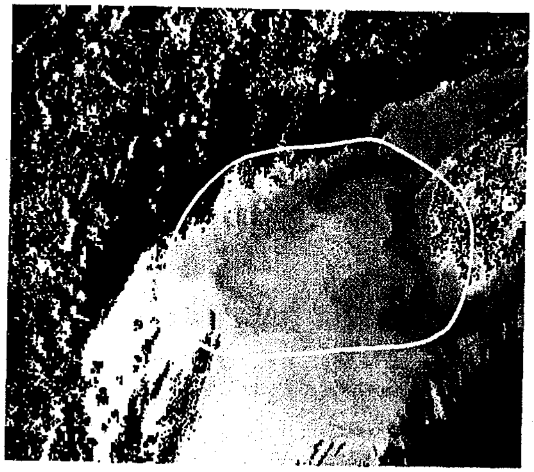

Figure 4. satellite image, showing sea surface temperature

4.2. Processing the Data

Process:

We start with removing the linear trend surface inherent in the sample data. This step is necessary since GS+ only supports ordinary kriging.

Next we determine the semi-variogram (used software: GS+).

Now we can start punctual kriging which results in an interpolated continuos representation (used software: GS+).

Followed by adding the trend surface (using a short c++ program) which results in an interpolated continuos representation of the sea surface temperature in the surveyed area.

The next step is the generation of the relief map of the residuals. This is accomplished by simply subtracting the interpolated surface from the satellite image (i.e., ground truth) (used software: ARC/INFO).

Next we use an AML in ARC/INFO to calculate the resulting binary maps—with the definition: `cylinder.aml <input grid> <output grid> <attribute height> <threshold> <radius>`. A defined cylinder is centered over each pixel within the relief map. At each location we can now calculate the number of pixels where the relief map exceeds the cylinder. A threshold decides whether the center pixel results in a white (i.e., detect-able) or in a black (i.e., not detect-able) output pixel.

4.3. Results

When applying the discussed model we investigate the detect-ability for two different objects. One of them with an attribute height of 2°C and the other one with an attribute height of 5°C, where the remaining dependencies (e.g., radius = 10km, threshold = 85%) of detect-ability are kept constant. The results can be seen in **Figure 5a.**—for the 2°C object—and in **Figure 5b.**—for the 5°C object.

The comparison of the two results (shown in **Figure 5.**) confirm the assumption that the areas where inferences about an object of an attribute height of 2°C are clearly smaller than the areas where inferences about an object of an attribute height of 5°C can be made with certainty. These results lead to the following conclusions:

- If objects of 2°C attribute height should be detected the used method (e.g., sample sp within the area of interest is *not* sufficient.
- If objects of 5°C attribute height should be detected the used method is sufficient.

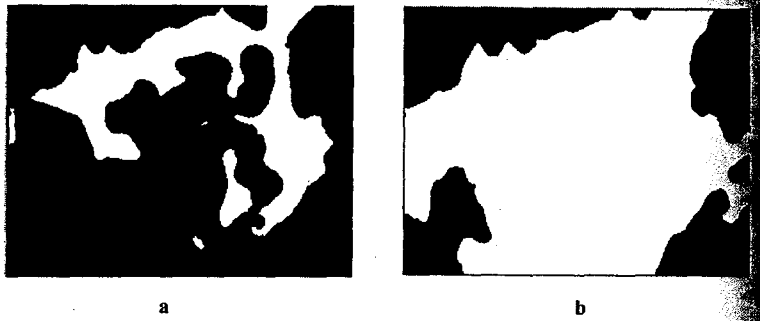

Figure 5. resulting binary maps a) for the 2°C object and b) for the 5°C object

5. Conclusions—Future Work

The discussed model enables the GIS user to determine whether the quality of a given fi representation is sufficient to detect a representative object. The result is presented via a binary ra representation where we can identify areas of positive and negative detect-ability. The user has provide the spatial extent and the attribute height of the object. Furthermore, if required, the us should have the ability to vary—up to a certain degree (e.g., 0% to 20%)—the threshold for th determination of detect-ability.

In our case study we used a satellite image as a reference (i.e., ground truth) to calculate th necessary residuals for the relief map. Future work will focus on including the model of condition simulations (as mentioned in section 3. The Model—How to Determine Detect-ability). For each o the simulated realizations we will be able to create a binary map. The final product will be a summar of all, for example, 100 cases.

Another promising research area using the discussed model is the investigation of the influenc on the binary result map when varying the dependencies of detect-ability. For example, we could reduce or increase the number of sample points and then analyze the relation between the number o sample points and the area of positive detect-ability. It would also be of interest to investigate the implementation of representative object shapes. Here we would like to look into the possibility o exchanging the cylinder by a line.

This paper investigates a simple approach to communicate aspects of inaccuracy and resolution of a field representation to the GIS user. Future work will show aspects of a) an implementation using conditional simulations and b) an exploration of the effects of the dependencies of detect-ability on the results.

Acknowledgements

The authors wish to thank Andrew Thomas and Elizabeth Cough for providing the data on the sea surface temperature in the Gulf of Maine. Most valuable inputs from Gerard Heuvelink helped to focus on the essential parts of the problem. Furthermore, he provided some interesting guidelines for future work on this model. We gratefully acknowledge NIMA for support of this research under grant NMA202-97-1-1021. We would also like to thank Craig Miller for his help on processing the data.

nces

M.K., 1996, A Structure for Organizing Metadata Collection. In *Proceedings 3rd International Conference on Integrating GIS and Environmental Modeling.* Sante Fe, NM. Jan 21-26: Santa Barbara, CA: NCGIA. URL: http://www.ncgia.ucsb.edu/conf/sante-fe_cd_rom/main.html.

C287, WG 2, 1995, Geographic Information - Data description-Quality. draft for discussion URL: http://forum.afnor.fr/afnor/WORK/AFNOR/GPN2/Z13C/PUBLIC/DOC/.

MAN, N., 1983, The Role of Quality Information in the Long Term Functioning of a Geographic Information System. In *Proceedings Auto Carto 6*, **1**, pp. 303-312.

LES, R., and WOODS, R., 1993. *Digital image processing* (Addison-Wesley Publishing Company).

HILD, M. F., and GOPAL, S., 1989, *Accuracy of Spatial Databases* (London: Taylor and Francis).

HILD, M. F., 1993, Data Models and Data Quality: Problems and Prospects. In *Environmental modeling and GIS*, edited by M. F. Goodchild, B. Parks, and L. Steyart (New York: Oxford University Press), pp. 363-371.

CHILD, M. F., and PROCTOR, J., 1997, Scale in a Digital Geographic World. *Geographical & Environmental Modelling.* **1**, 1, 5-23.

ILL, S. and Morriosn, J. L., 1995, *Elements of Spatial Data Quality* (Tarrytown, NY: Elsevier Science).

ELINK, G.B.M., 1993, *Error Propagation in Quantitative Spatial Modeling* (Utrecht: Drukkerij Elinkwijk).

ELINK, G.B.M., 1998, *Error Propagation in Environmental Modelling with GIS* (London: Taylor & Francis).

211 1998, URL: http://www.statkart.no/isotc211/

ONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, 1994, Federal Information Processing Standard Publication 173. (Spatial Data Transfer Standard Part 1.Version 1.1) U.S Department of Commerce.

ONS, S., 1996, Current Approaches to Handling Imperfect Information in Data and Knowledge Bases. *IEEE Transactions on Knowledge and data Engineering.* **8**, 3, 353-372.

ET, D., SMITH, B., and BROGAARD, B., 1999, *The Ontology of Fields*, Report of a Specialist Meeting held under the auspices of the Varenius Project.