8 Modeling Data Quality with Possibility Distributions

Gerhard Navratil

CONTENTS

8.1 INTRODUCTION

The amount of available data has increased with the development of new technologies. The availability of data and the capability of processing more data than before have led to new applications like online route planning or visualizations in landscape planning and architecture. The outcome of the application depends on the adequate selection of datasets. However, data quality varies with the source. Data quality descriptions have been defined to cope with that problem (Guptill and Morrison, 1995).

The automation of data processing requires the automatic handling of data quality. Fitness for use describes if a specific dataset is suitable for a specific task (Chrisman, 1984). Automatic preselection of the datasets can simplify the selection process for the user but requires automatic determination of the fitness for use. Byrom (2003) pointed out the necessity for the discussion of user needs. Grum and Vasseur (2004) and Pontikakis and Frank (2004) presented approaches to define the user needs and to use them to specify the fitness for use. A first step in that process is the suitable description of data quality.

Datasets are often collected over long periods. Graphs of street networks, for example, are determined once and updated periodically to reflect changes in the network. Since the quality of the determination may change due to improved technology, the quality of the data varies within the dataset. A worst-case scenario may be used for providing a number for the quality. However, this solution may give a wrong impression if a small part of the data is of significantly worse quality than the rest of the dataset. Quality should thus not be described by a single value.

The use of fuzzy numbers is a solution for varying data quality. Fuzzy numbers specify a range for the value. This allows showing the range of quality within the dataset. Fuzzy numbers are defined using distributions. Probability distributions specify probabilities for the different possible values whereas possibility distributions only describe the possibility of the outcome. In this chapter I discuss an approach using possibility distributions. I use the Austrian cadastre as an example to specify possibility distributions. Different aspects of quality are modeled with different distributions. I then assume user requirements and present a method to compare the data quality with the user requirements. This allows assessing the fitness for use.

8.2 DATA QUALITY

How can we describe data quality? ISO 19113 "Quality Principles" (ISO 19113, 2002) defines the framework for a quality model. The data quality elements are completeness, positional accuracy, temporal accuracy, logical consistency, and thematic accuracy. Each of these elements describes a specific aspect of geographic data. Determination of positional accuracy provides an example where the use of precise numbers is not sufficient for complex datasets. The other elements can be treated in a similar way.

The determination of data quality must consider the creation process. The creation process is influenced by three different aspects, which influence the quality of the resulting data (Navratil and Frank, 2005): technology, legality, and usability.

Technological possibilities limit the achievable quality. In general, there is a maximum quality as well as a minimum quality. The usual method to create a terrain model, for example, is either laser scanning or aerial photogrammetry. In both cases the quality of the terrain model depends on the expenditure. Lower flight height will improve the quality of the model. Different measurement equipment will result in different precisions. The precision will not be arbitrarily high since there are always small changes in the terrain, e.g., footsteps, which should not influence the terrain model. Reduction of quality is limited, too. Less than a single height point in a terrain model is useless.

Laws have an impact on the quality of available datasets, too (Navratil, 2004). Laws may prohibit the use of data with higher quality than specified due to data protection laws or security reasons. In contrast to the technological limitations, legal influences are not "hard." It is possible to specify the maximum technical quality of a distance measure. This is not true for laws. The Austrian law stipulates a minimum precision of 15 cm for boundary points in the coordinate-based cadastre. Since it is difficult to prove the quality of coordinates, a point with a precision of 20 cm may be accepted, too. Thus legal rules on data quality can be seen as guidelines to develop technical solutions. It is then assumed that the results of the process meet these rules.

Usability may affect data quality. Data used more often may have higher quality since they produce more revenue and thus more money is available for collecting additional data. Nautical maps, for example, have higher quality in the areas where it is needed. Users only need details in coastal areas where the danger of hitting the ground is high. Thus more money is spent on mapping coastal areas than on mapping the ocean.

8.3 IMPRECISE NUMBERS

Many real-world situations cannot be described precisely. Statistics on the number of cars waiting at a red traffic light seems to be a simple task, but the definition of a "waiting car" is difficult. A stopped car is definitely waiting, but how about a car rolling slowly toward the traffic light? What is the maximum speed that a rolling car may have to be labeled "waiting"? Questions like that led to the development of mathematics with imprecise numbers.

Reasoning can be defined as testing the correspondence of a specified hypothesis with given statements. The statements can be data stored in a database and the hypothesis is a query on these data. A typical example is a database containing the heights of persons and the question of whether a specific person is ""tall." Four different situations can be determined (Dubois and Prade, 1988a):

- Both the data in the database and the definition of "tall" are crisp. The entry in the database for the person could be 1.7 m and "tall" is defined as ">1.65 m." This leads to traditional, two-valued logic.
- The data are vague and the definition of "tall" is crisp. Here the definition for "tall" is the same as above but the entry in the database is expressed with a possibility distribution. This leads to possibility theory as published by Zadeh (1978; 1979) and expanded by Dubois and Prade (1988b).
- The data are crisp and the definition of " tall" is vague. The entry in the database could be 1.7 m but the concept of "tall" is uncertain. This leads to many-valued logic.
- Both the data and the definition of "tall" are vague. This leads to fuzzy logic (Zadeh, 1975).

Which of these types of logic shall we use for modeling data quality? Data describe the world. Since the world changes, the data must change, too. Thus the data acquisition is a continuous process. Data quality parameters shall describe the quality of this data. It will not be possible to use a crisp description because the quality will vary throughout the dataset, and this variation should be reflected by the data quality description. Thus we deal with uncertain data.

The questions are crisp or can be made crisp. Users have two different questions:

- I need a dataset with a specific quality. Is it available?
- There is a dataset with a specific quality. Can I use it for the purpose at hand?

Both questions are crisp. In the first case, there may be several parameters for the data quality. All of these parameters must be fulfilled. Thus a dataset either fulfills the quality specification or it does not. This gives a crisp answer to the question. The second question is more complex. Again data quality issues must be considered, but in addition a cost-benefit analysis is necessary. According to Krek (2002), the value of a dataset emerges from better decisions. The value can be compared to the costs of acquisition and processing of the data. The dataset is applicable if the costs are lower than the benefits and there is no other possible outcome than using or not using the dataset. Thus both questions are crisp and we must use possibility theory.

8.4 POSSIBILITY DISTRIBUTIONS

A discussion of processes requires a method to describe the outcome of the processes. Possibility distributions (Zadeh, 1978) are such a method. In general, the use of fuzzy methods is suitable for the results of precise observation processes, and they can be used for statistical analysis (Viertl, 2006). Viertl uses probability distributions, which assign probabilities to each possible outcome. Determination of probabilities requires detailed knowledge. Possibility distributions avoid that problem. Possibility distributions only specify the possibility of the result: The value 0 shows impossibility and 1 shows possibility. Values between 0 and 1 provide information on the plausibility of the outcome. Thus, a result with value 0.4 is possible but less plausible than a result with 0.8. However, a result with 0.8 is not twice as probable as a result with 0.4.

The use of a set Θ of mutually exclusive and exhaustive possibilities is the most common way to express propositions (Wilson, 2002). A possibility distribution π assigns a value of possibility to each element of the set. If there is an element with value 1, then the function is said to be normalized:

$$\pi: \Theta \rightarrow [0,1]$$

8.5 QUALITY OF CADASTRAL DATA

The Austrian cadastral data are used as an example for a large dataset collected over an extended period. The dataset includes parcel identifiers, parcel boundaries, and current land use. Details on the Austrian cadastral system can be found in different publications (e.g., Twaroch and Muggenhuber, 1997). An important aspect is the definition of boundary. Whereas evidence in reality (like boundary marks, fences, or walls) defines the boundary in the traditional Austrian cadastre, the new, coordinate-based system uses coordinates to specify the position of the boundary. This change allows the creation of datasets reflecting reality since the data provide the legal basis for the boundaries.

The elements of data quality as listed in Section 8.2 must be defined in order to specify the quality of the Austrian cadastre. Positional accuracy connects to the elements defining the boundary lines. The Austrian cadastre uses boundary points to define the boundary. Thus the positional accuracy of the boundary points stipulates the positional accuracy of the dataset.

8.6 MODELING DATA QUALITY WITH POSSIBILITY DISTRIBUTIONS

8.6.1 Technological Influence

Positional accuracy for cadastral boundaries depends on the accuracy of boundary points, which depends on the precision of the point determination and the point definition itself. Thus the accuracy of the points will be used in the following discussion. Modern technical solutions for point determination use GPS and high-precision measurement equipment. This results in a standard deviation of 1–5 cm for the points based, e.g., on Helmert's definition, $\sigma_H^2 = \sigma_x^2 + \sigma_y^2$. This can be reached if the whole dataset is remeasured to eliminate influences of outdated measurement methods. Reduction of quality is possible, e.g., by using cheaper equipment. The lower limits are reached if the topology described by the dataset is influenced by random deviations. These effects may start with an accuracy of approximately 1 m and the dataset will become unusable in large parts of Austria with an accuracy of 10 m. Figure 8.1 shows the possible positional accuracy for boundary points.

8.6.2 Legal Influence

The positional accuracy of boundaries depends on the cadastral system used, the coordinate-based cadastre or the traditional cadastre. The traditional cadastre allows adverse possession. A person acquires ownership of land by using the land for 30 years in the belief that the person is the lawful owner. This is only detected during boundary reconstruction or in cases of disputes. Thus parts of the dataset

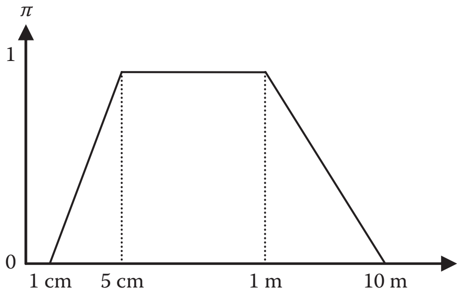

FIGURE 8.1 Possibility distribution for technological influence on positional accuracy.

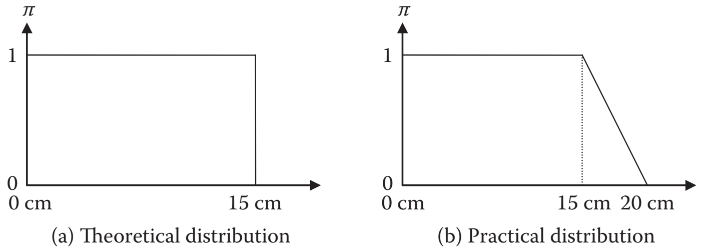

FIGURE 8.2 Possibility distribution for legal influence on positional accuracy in the coordinate-based system.

will not be describing the correct boundaries, and even points with high internal accuracy may be incorrect. Therefore, the overall accuracy is low but it is impossible to specify precise numbers. An estimate of the percentage of affected points cannot be provided, but it seems plausible that the number is not high because many boundaries are fixed by walls or fences. The possibility distribution will be similar to Figure 8.2b, but the values will be in the range of meters.

Accuracy is better defined for boundary points in the coordinate-based cadastre. The decree for surveying (Austrian Ministry for Economics, 1994) stipulates a minimum positional accuracy of 15 cm for boundary points. This value determines the standard deviation for the boundary points. Thus, theoretically, the possibility distribution for the positional accuracy looks like the one in Figure 8.2a. This rule is strict, as the law disregards statistical measures like standard deviation for decision making (Twaroch, 2005). However, it is difficult to control the actual accuracy of a boundary point. The existence of points with lower accuracies is possible. This is modeled in Figure 8.2b. Accuracies of less than 20 cm should not be possible since they should have been detected.

8.7 MODELING USER NEEDS

Two different groups of users of cadastral data are considered:

- Users of the boundary itself: Owners of land need data on their parcel and the neighboring parcels with high accuracy.
- Users of the positional reference in general: The cadastre is the only large-scale map available for the whole area of Austria, and thus it is often used to provide spatial reference.

These two groups have different requirements. The differences will show in the possibility distributions. In contrast to the technological and legal influences, the possibility distributions are not based on the specifications of the dataset but on the intended application. The possibility distribution shows if it is possible to use the dataset for the specific application.

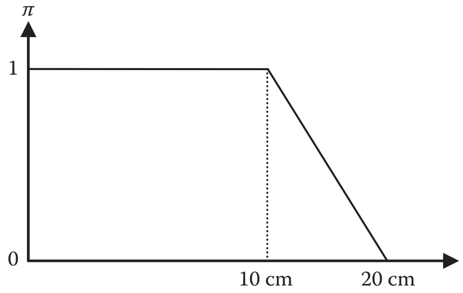

FIGURE 8.3 Positional accuracy for users of boundary.

8.7.1 Users of the Boundary

Positional accuracy is important for land owners. Land owners want to use their land, e.g., by creating a building. In Austria buildings must comply with legal rules specifying, for example, the maximum building height or the distance from the parcel boundary. The last point requires high positional accuracy to fit the strategies of courts. Thus, although an accuracy of 20 cm may be sufficient for some tasks of land owners, most tasks require a positional accuracy of at maximum 10 cm (compare Figure 8.3).

8.7.2 Users of the Spatial Reference

Spatial reference has limited demands for positional accuracy. Assuming a scale of 1:10.000 and accuracy on the map of 1/10 mm, then the accuracy of the points should be 1 m. Higher mapping accuracy leads to higher accuracy demands, but accuracy better than 0.5 m is not needed for positional reference. The lower limit of accuracy depends on the type of visualization. Accuracy of less than 10 m in builtup areas may result in less plausible datasets because it will not be possible to determine on which side of a street a point is (compare Figure 8.4).

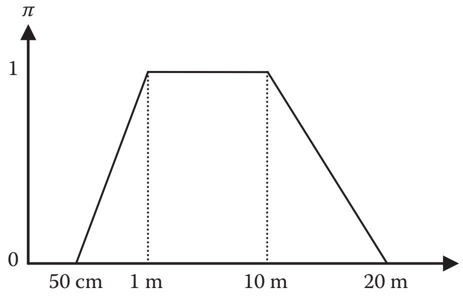

FIGURE 8.4 Positional accuracy for users of boundary.

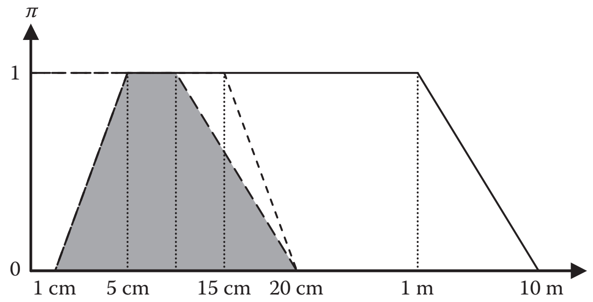

FIGURE 8.5 Combination of possibility distributions for users of the boundary.

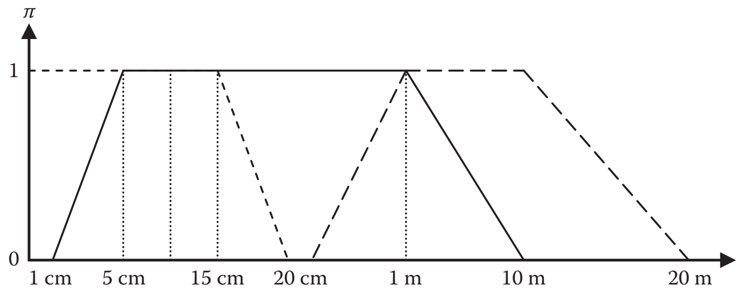

FIGURE 8.6 Combination of possibility distributions for users of the spatial reference.

8.8 COMBINATION OF POSSIBILITY DISTRIBUTIONS

In many cases data quality must meet several conditions. These conditions can be combined by a logical “and”-relation. The minimum-function provides this for possibility distributions. The “or”-relation would lead to the maximum-function (Viertl, 2006; Viertl and Hareter, 2006). Figure 8.5 shows the combination of the possibility distributions for positional accuracy. The gray area marks the overlap of the possibilities. Figure 8.6 shows the same combination for users of the spatial reference. This combination has no solution.

The example shows that the technical solutions and legal rules for cadastral systems do meet the demands of owners of land. Other types of use have different demands and thus the possibility distribution is different. Users who need spatial reference only require a different technical solution and different legal rules.

8.9 CONCLUSIONS

As we have seen, it is possible to model the influences on data quality with possibility distributions. It was possible to specify all necessary possibility distributions. The combination of influences produced a result that can be verified by practical experience. The method thus can be used to assess the correspondence of the influences on data quality.

Left for future investigation is the application for dataset selection. The chapter showed how to model possibility distributions for influences on data quality. It showed a simple method of combination. A general method will be needed to create the possibility distribution for more general examples. These distributions might require more sophisticated methods of combination.

REFERENCES

Austrian Ministry for Economics, 1994. Verordnung des Bundesministers für wirtschaftliche Angelegenheiten über Vermessung und Pläne (VermV). BGBl.Nr. 562/1994.

Byrom, G. M., 2003. Data Quality and Spatial Cognition: The Perspective of a National Mapping Agency. In: *International Symposium on Spatial Data Quality*, The Hong Kong Polytechnic University, pp. 465–473.

Chrisman, N. R., 1984. The Role of Quality Information in the Long-Term Functioning of a Geographical Information System. *Cartographica* 21: 79–87.

Dubois, D. and H. Prade, 1988a. *An Introduction to Possiblistic and Fuzzy Logics. Non-Standard Logics for Automated Reasoning*. P. Smets, E. H. Hamdani, D. Dubois and H. Prade, Eds., London, Academic Press Limited, pp. 287–326.

AU: Please supply title of chapter in the book.

Dubois, D. and H. Prade, 1988b. *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. New York, NY, Plenum Press.

Grum, E. and B. Vasseure, 2004. How to Select the Best Dataset for a Task? In: *International Symposium on Spatial Data Quality*, Vienna University of Technology, pp. 197–206.

Guptill, S. C. and J. L. Morrison, Eds., 1995. *Elements of Spatial Data Quality*. New York, NY, Elsevier Science, on behalf of the International Cartographic Association.

AU: Correct location of publisher?

ISO 19113, 2002. Geographic Information—Quality Principles.

Krek, A., 2002. An Agent-Based Model for Quantifying the Economic Value of Geographic Information. PhD thesis, Vienna University of Technology.

Navratil, G., 2004. How Laws affect Data Quality. In: *International Symposium on Spatial Data Quality*, Vienna University of Technology, pp. 37–47.

Navratil, G. and A. U. Frank, 2005. Influences Affecting Data Quality. In: *International Symposium on Spatial Data Quality*, Peking.

Pontikakis, E. and A. U. Frank, 2004. Basic Spatial Data according to User's Needs-Aspects of Data Quality. In: *International Symposium on Spatial Data Quality*, Vienna University of Technology, pp. 13–21.

Twaroch, C., 2005. Richter kennen keine Toleranz. In: *Intern. Geodätische Woche*, Obergurgl, Wichmann.

Twaroch, C. and G. Muggenhuber, 1997. Evolution of Land Registration and Cadastre. In: *Joint European Conference on Geographic Information*.

AU: Location of conference?

Viertl, R., 2006. Fuzzy Models for Precision Measurements. In: *Proceedings 5th MATHMOD*, Vienna, ARGESIM / ASIM.

Viertl, R. and D. Hareter, 2006. *Beschreibung und Analyse unscharfer Information*. Vienna, Springer.

Wilson, N., 2002. A Survey of Numerical Uncertainty Formalisms, with Reference to GIS Applications. Annex 21.1 to REV!GIS Year 2 Task 1.1 deliverable.

Zadeh, L. A., 1975. Fuzzy Logic and approximate Reasoning. *Synthese* 30: 407–428.

Zadeh, L. A., 1978. Fuzzy Sets as a Basis for a Theory of Possibility. *Fuzzy Sets and Systems* 1: 3–28.

Zadeh, L. A., 1979. A Theory of Approximate Reasoning. Machine Intelligence, Vol. 9. J. E. Hayes, D. Michie and L. I. Mikulich, Eds. New York, Elsevier, pp. 149–194.

AU: Please supply title of chapter in book.