Trust Indicator for Decisions Based on Geospatial Data

Ivanildo Barbosa[1,2] and Marco A. Casanova[2]

[1]*Seção de Ensino de Engenharia Cartográfica*
Instituto Militar de Engenharia (IME)
Praça General Tibúrcio, 80 – CEP 22290-270 – Rio de Janeiro – Brasil

[2]*Departamento de Informática*
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)
Rua Marquês de São Vicente, 225 - CEP 22451-900 – Rio de Janeiro – Brasil
{ibarbosa,casanova}@inf.puc-rio.br

Abstract A large family of real-world applications are influenced by geospatial features and known relationships between them. Also, a very large volume of geospatial data is currently available from different sources, prepared for different applications, and with different levels of uncertainty. This paper presents a brief analysis about spatial, thematic, technical and temporal aspects of uncertainty and about how they influence the reliability of decisions based on such datasets. It also proposes an indicator to quantify the inherent reliability of such data, based on their provenance, completeness, spatial coverage and data lifetime. In particular, the indicator is applicable in the context of planning applications for which geospatial data are relevant in order to rank or discard available geospatial datasets.

Keywords: data reliability, geospatial uncertainty

1 Introduction

In the past decades, geospatial data producers considerably improved methods to acquire, process and distribute geospatial data (vectors, images, aerial photographs, thematic maps, etc) to the different kinds of users, who depend on such data to their decision making processes. Government agencies and collaborative initiatives offer online data that can be accessed through traditional user interfaces or through Web services.

However the best geospatial datasets will never show complete fidelity to the reality wherever and whenever users need to use them. Maps are designed to bring a controlled level of uncertainty, considered irrelevant for some kinds of users and applications. Despite the fact that data providers adopt the most reliable methods and use the most precise and accurate data acquisition platforms, some uncertainty will remain in geospatial data.

A wide variety of planning applications depends on geospatial data, either static or time-varying, and some of them demands low tolerance for uncertainties. Therefore, users must assess the geospatial data sources in order to select those best suited to their applications (according to the related knowledge base) [1,2].

This paper proposes an approach to evaluate the reliability of geospatial datasets in the context of planning applications that takes into account *spatial*, *thematic*, *temporal* and *technical* aspects.

The paper is organized as follows. Section 2 summarizes the standard quality indicators for geospatial data [3]. Section 3 analyzes the sources of indeterminacy and proposes alternative quality indicators for geospatial data. Section 4 applies the concepts proposed to plan routes for off-road vehicles based on geospatial data from several sources. Section 5 summarizes the approach and discusses future research lines to refine present results.

2 Quality indicators for geospatial datasets

Geospatial data are typically produced to meet the requirements of a given set of applications. The production processes are guided by well-defined specifications to provide a controlled level of accuracy and uncertainty which is adequate for the set of applications in question. Hence, different geospatial datasets covering the same area may have different characteristics.

When geospatial data were represented as printed maps, the only way to evaluate their accuracy was by comparing the represented coordinates of a number of geographic features with their real coordinates measured over the terrain. The differences, as well as their standard deviation, should be less than specified thresholds [4]. Today, geospatial data are digitally represented as vectors, matrices (images or coverages), lists of coordinates and databases, which demand the adoption of proper criteria to assess quality [3].

Quality references are relevant metadata so that the ISO 19115 standard [5] defines a package to deal this issue. In the Data Quality package, one may store the reports of measurements procedures (described in ISO 19114 [6]) and the description of process steps, and the respective data sources used, to create the dataset (also known as *lineage*).

The current specification to assess the quality of a geospatial dataset evaluates its *completeness*, *logical consistency*, *positional accuracy*, *thematic accuracy* and *temporal accuracy*.

Completeness indicates the omission or excess (*commission*) of geographic features, attributes and relationships in the dataset over its declared geographical extents. It may influence query results by improperly accepting or reject features. For raster data, pixels usually do not have null values. However, some applications consider a specific pixel value to representing the absence of value.

Logical consistency provides information about the adherence to rules related to the data structure, attributes and relationships. Such rules allow matching the attributes provided with the list of required attributes and to verify if the provided data is in conformance with the defined domains and formats.

Accuracy reflects preoccupations with spatial, thematic and temporal issues. The specified data quality reports point to absolute values and conformance to some specification, demanding further analysis.

3 Criteria for reliability

Although there are specifications to assess quality, the results obtained are not enough to assign any reliability index to data. Metadata about identification and spatial reference are also necessary to analyze the usability of a dataset. In this section, we propose a set of criteria to assess reliability for geospatial data, which better matches the requirements of planning applications, among others. The

proposed criteria are *spatial coverage*, *data completeness*, *provenance* and *lifetime*.

3.1 Spatial Coverage

Spatial coverage indicator is proposed to assess reliability for geospatial data by analyzing spatial aspects. The first aspect of spatial coverage points to evaluate *how the dataset extents cover the area of interest for planning*: fully, partially or not at all. Both the dataset and the area of interest extents are predefined by, respectively, the dataset design and the application specification. Ideally, the geospatial datasets must therefore cover an area that contains the area of interest, as otherwise the planning process may be affected by lack of available data. The polygons used do define the extents of both the dataset and the area of interest may be compared to compute the overlapping area among them and return the percentage of the area of interest covered by the dataset. The example in Figure 1 illustrates how datasets (dashed boxes) cover the area of interest (continuous line polygon). No dataset covers the whole area of interest (continuous line), so the reliability for each dataset would be lower than 100%. Even if the datasets were merged, the reliability of the whole dataset would be lower than 100%.

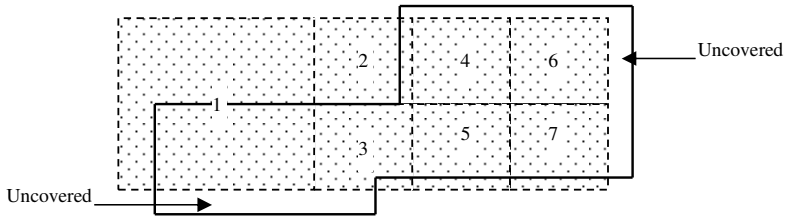

Figure 1 Example of spatial coverage: dashed boxes represent geospatial datasets that cover partially the area of interest represented by the continuous line bounded polygon

When partial coverage occurs, it is suggested to divide the planning area into *covered* from *uncovered areas*. Using the example of Figure 1, the original area would be divided in nine parts: seven areas covered by the respective datasets (100% covered) and the two remaining two not covered.

The total overlap of the areas does not provide completely reliable information. The analysis discussed before considered only the geographical extents of data. However, it is necessary to ensure that all the existing features are represented, by measuring *completeness (omission* and *commission)* [6] or using the respective metadata, when available. It aims at indicating whether all the

geometries relative to features are represented in the extents. Despite the fact that attributes are not fulfilled, these features exist and may be identified even by visual analysis.

After checking the integrity of the representations, the next approach is to analyze the individual geometries stored in the dataset. Data producers align their methodologies and materials in order to achieve a precision coherent with a predefined scale, called *equivalent scale* (usually defined by its denominator [5]). However, the exhibition scale may be controlled by the user and generalization rules must be applied to simplify the representation [7, 8]. Generalization for vector data restricts the types of visualized features, enhances relevant features or types of features in a given context, displaces or omits some features according to their both dimensions and specified precision, and simplifies some features representations. Larger equivalent scales (denominators) imply less detailed geometries. Equivalent scale also impacts the criteria to evaluate omission because some features may be not represented in some scale ranges for datasets considered complete.

Considering raster data (pictures, images and grids), a pixel represents a regularly sized portion of the terrain, either based on the signals captured by the sensors or by transforming vector to raster. Details smaller than the area covered by one pixel are ignored. Therefore, the equivalent scale must be compatible with the pixel size.

In order to assess the reliability of a dataset, it is necessary to compare its *equivalent scale* to a specified reference value, called here *proper equivalent scale for application* (PESA). In order to integrate vector and raster data and to facilitate the understanding of practical consequence of the concept, the *spatial resolution* (the size, in meters, of the smallest detail represented at the respective scale) may be used instead of the equivalent scale value.

3.2 Data Completeness

Completeness, in this context, is related to the *thematic* integrity of the representation of the existing features, that is, how comprehensive are the non-geometric attributes in a database and also in the raster data. These attributes describe the feature and an analysis of their values is relevant to select geographic features according some specific condition. Queries in incomplete databases will

accept or reject mistaken features due to match (or mismatch) attributes and conditions.

Despite the fact that the notion of completeness described in [3] and [5] partially merges the concepts of spatial coverage and data completeness, we distinguish these issues in this paper. While an *uncovered area* misses geometries (and respective attributes), an *incomplete area* misses only some attributes values, although there are geometries for every features.

Null values indicate lack of information in table registries and raster representations. Furthermore, although not explicitly indicated, default values are used to replace absent of data, thereby becoming an indicator for lack of data – field values in vector data and pixels in raster data.

Absent data may be estimated by mining available data, as reported for example by Pearson [9]. However, the estimation methods and models may also embed uncertainty, so it is recommended to distinguish the reliability of the measured (or observed) values from the estimated values.

Figure 2 presents a proposed ranking for reliability based on data completeness. It is suggested to assign values between 0 and 1, proportional to the percentage of fulfilled (or estimated) values. The assigned reliability values will be higher when the dataset satisfies the conditions at the top of the figure. On other hand, lower values will be assigned if the conditions at the bottom of the figure hold. The heights of intermediary boxes illustrate the differences between the ranges of values for reliability at each of the conditions shown and are intended to be qualitative – out of scale. The blank area represents the upper limits for reliability in some cases and is out of scale (in order to fit the text).

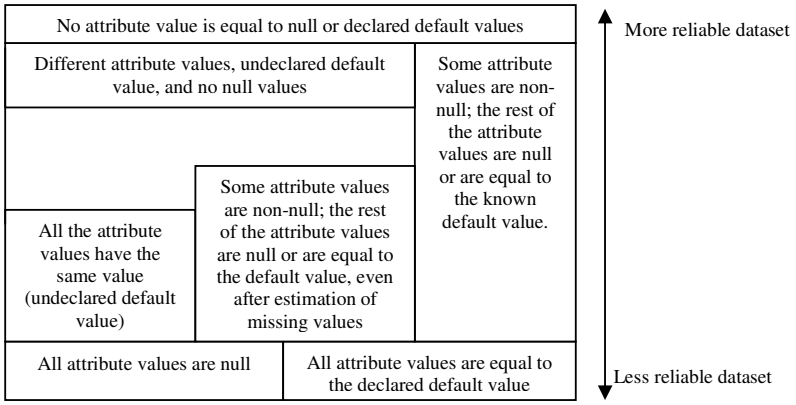

Figure 2 Relative values for reliability based on the data completeness criterion

3.3 Provenance

The usual concept of provenance is related to the source or history of some product (see [10] for a survey). In general, different entities may produce geospatial datasets aiming at achieving different levels of accuracy for different applications. It implies in more (or less) severe specifications, more (or less) accurate equipments and methods, and so on. At the context of geospatial data quality, provenance is related to the lineage.

As a simple example, consider the use of portable GPS receivers by non-corporative users to acquire geospatial data (receiving, grouping and labeling waypoints and tracks) and to publish the acquired data on the Web. A typical use is to georeference features not represented on conventional maps. On other hand, such data is defined only by geometry, with no post-processing to minimize GPS errors [11, 12]. The reliability of such datasets tends to be lower than those created with more accurate methods and equipments.

In general, geospatial data producers should abide to specifications that define methods, equipments, precisions and contents when publishing their datasets. On one extreme of the (trusted) provenance scale, we may classify government agencies that deploy Spatial Data Infrastructures (SDI) adopting standards for files and Web services. Data users rely on such standards and on the reputation of producers to assess data quality. On the other extreme, we may include companies that provide datasets for specific purposes. In this case, users will perhaps depend on some methodology to assess data quality with respect to provenance. We may include in this second category academic institutions that produce data for customized applications, according to standards published by national or international organizations. However, datasets thereby produced may be useful only for the purpose they were created, due to their particularities.

To ensure the adherence of data characteristics to the specifications, it would be recommended that a certification be issued by competent institutions to warrant the usability of that dataset at that particular application – here called *warranty of conformance to application specifications* (WCAS). However, this certification is frequently ensured by the customer himself empirically. In some cases, the producer has an independent auditing department to evaluate this adherence according to predefined legal or technical limits. Therefore, assigning trust certification to producers will warrant the usability of the datasets they

produce. Both options involve legal discussions about the certifier authorities and technical issues about criteria for certifying datasets and producers.

To summarize, the quality of the data in this context depends on the application, on the reputation of the data provider (academic, industrial and government), and on the ability of the user to assess and warrant the datasets. Therefore, the definition of absolute weight values to different data providers is not a simple task. So, we propose a ranking analogous to that proposed in section 3.2, illustrated in Figure 3. The assigned reliability values will be higher when the dataset satisfies conditions at the top of the figure. On other hand, lower values will be assigned in conditions at the bottom of the figure.

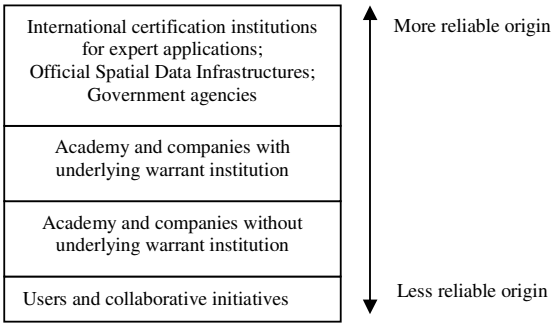

Figure 3 Relative values for reliability based on the provenance criterion

3.4 Lifetime

"*Maps are like milk: their information is perishable, and it is wise to check the date*" [8]. This statement reflects the caution of users about any kind of data and motivates the discussion about dataset reliability after some elapsed time.

The first approach to assess the lifetime for a dataset depends on its age, defined as the time elapsed between the acquisition of the data (certified by the data quality reports of temporal accuracy [6]) and its use by the application. In most planning activities, more recent datasets are preferred to older ones. However, even the most recent datasets may be outdated because geographic features types represented in the dataset change at different rates. In some cases, the representation of each geographic feature has its own distinct indicator. However, it is suggested to consider a single lifetime value for the whole geospatial dataset.

Furthermore, to come up with a reasonable estimation for the rate of change of the data may be quite difficult, although some reasonable

approximations are feasible. For example, natural feature types, such as physiographic and hydrographic ones, present slower changes, usually caused by natural phenomena, such as geological movements and long-term weather events. Man-made feature types tend to change faster (including physiographic, hydrographic and vegetation features), as a result for example of expanding populated areas or increasing infrastructure (transport systems, energy production and distribution, *etc.*).

It is therefore necessary to introduce the concept of *safe age for data application* (SADA), meaning the maximum time interval after date of creation (or last update) the dataset may be considered unchanged. It depends on the application, the equivalent scale, the feature type and the potential of human influences.

Figure 4 illustrates the relationship between the time related aspects mentioned above. The edges of the cube (adapted to facilitate visualization) represent *human occupation* – amount and distribution of fixed population at the area of interest, *economic activities* – indication of land use and the consequent potential to change features, and data *lifetime*. The surface represents the threshold for acceptable values. It is not flat because the relationship among the concepts is not linear, demanding further research to model it. However, the relative relationships about lifetime are preserved: less human occupation and less economic activities imply higher lifetime values for the dataset, and vice-versa. Statistic data about socioeconomic aspects may be retrieved from governmental institutions responsible for census.

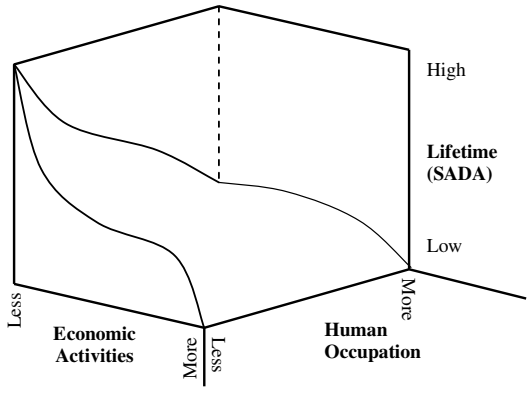

Figure 4 Relationship between lifetime, dataset age and socioeconomic aspects

In order to improve the reliability of geospatial datasets according to this criterion, it is necessary to create policies for checking and for updating feature types with low values for SADA. Even not changing the features, the age will be counted from the last verification.

4 Combining the indicators

After establishing strategies to assess each quality indicator, we must indicate how to compose them to create a unique quality indicator. There are two main approaches to deal it: the first one considers the geometric mean of normalized indicators for lifetime (*L*), spatial coverage (*S*), data completeness (*D*) and provenance (*P*), while the second one defines a qualitative classification to indicate application profiles based on fuzzy concepts. The geometric mean was chosen instead the arithmetic mean because it is indicated to handle rates. Hence, the reliability for some datasets may be assigned as 0, meaning that dataset offers no reliability. A fuzzy approach may provide some reliability for rejected datasets and will be useful when no available datasets meet the specifications and the user may choose a less imperfect dataset. This paper will deal the first approach, suggesting the second one for further discussions.

The first step is to locate both the extents of the dataset and of the area of interest, dividing the whole area to define areas fully covered by the datasets (as illustrated in Figure 1). Planning in uncovered areas must be avoided due the total lack of information.

Spatial resolution (indicator *R*) does not affect only accuracy, but also simplifies or deletes some features representations. Therefore, it is recommended to reject ($R = 0$) datasets with spatial resolution lower (coarse) than the specified one. For finer values, the spatial resolution indicator value *R* is assigned as 1.

In areas covered by datasets with proper resolution, the spatial completeness indicator will define spatial coverage indicator *S*.

The criterion to assign the indicator for provenance *P* may be simplified, by considering the existence of warrant certification (assigning value 1 to *P*). The absence of such certification assigns a partial value to P (0.5, for example).

After computing how long a dataset is valid for a specific application, the assigned value for lifetime indicator *L* will be 1, if the specified SADA is larger than the dataset age. Otherwise, the assigned value will be 0.

Hence, only provenance and data completeness provide values other than 0 and 1 for their respective indicators, *P* and *D*, belonging to the interval [0,1]. The proposal to compute an overall indicator *I* for an individual dataset is presented in (1).

$$I = L \cdot \sqrt[3]{P \cdot S \cdot D} \tag{1}$$

This definition of *I* assumes that all indicators have the same relevance. However, specific indicators may have different influences for some applications. In this case, the definition of *I* may be modified by assigning weights to balance the influence of the indicators. However, indicators will be not accurate enough to demand variations.

When the application requires data from multiple datasets, we would compute a separate indicator I_i for each dataset. The final indicator might be defined as a geometric mean of individual indicators.

$$I = \sqrt[n]{\prod_{i=1}^{n} I_i} \tag{2}$$

5 Conclusions

This paper proposed an approach to evaluate reliability indicators for geospatial data in the context of planning applications. The approach did not question thematic, temporal and positional accuracies measurement [6], but it rather relied on metadata in data quality package (*provenance* and *data completeness*), in identification package (*spatial coverage* and *spatial resolution*, or *equivalent scale*), and external data based on socioeconomic factors (*lifetime*). It means that some datasets should be replaced due to incompleteness (of attribute values), obsolescence and inadequate scale (insufficient level of spatial details). Further studies may deal with the cases where all datasets were rejected, using fuzzy criteria to compose an indicator for "best" fitting.

The use of concepts of *WCAS*, *PESA* and *SADA* aims at supporting the ranking process to select or discard geospatial datasets based on their reliability values. However, further studies are necessary to refine both criteria and threshold values to rank datasets reliability, either isolated or combined. In these cases, one

might rely on expert knowledge to obtain more meaningful indicators for the evaluated datasets face to the target application.

References

1. Russell S, Norvig P (1995) Artificial Intelligence: a modern approach, Prentice Hall, New Jersey.

2. Taştan, H, Altan, MO (1999) Spatial Data Quality, In: Proceedings of 3rd Turkish-German Joint Geodetic Days.
http://www.hgk.msb.gov.tr/haritalar_projeler/bildiriler/cbs/makale(pdf)/cbs_tek_bil3.pdf . Accessed 20 July 2011.

3. International Organization for Standardization (2002) ISO 19113: Geographic information -- Quality principles.

4. Brasil (1984) Decreto Nº 89817. Estabelece as Instruções Reguladoras das Normas Técnicas da Cartografia Nacional.

5. International Organization for Standardization (2003) ISO 19115: Geographic information -- Metadata.

6. International Organization for Standardization (2003) ISO 19114: Geographic information -- Quality evaluation procedures.

7. Mackaness WA, Ruas A, Sarjakoski LT (2007) Generalisation of Geographic Information: Cartographic Modelling and Applications. Elsevier, Amsterdam.

8. Monmonier, M (1996) How to Lie with Maps. 2nd Edition. University Of Chicago Press, Chicago.

9. Pearson RK (2005) Mining Imperfect Data: Dealing with Contamination and Incomplete Records. SIAM, Philadelphia.

10. Marins ALA (2008) Modelos Conceituais para Proveniência, Dissertation, Pontifical Catholic University of Rio de Janeiro.

11. Monico JFG (2008) Posicionamento pelo GNSS – Descrição, Fundamentos e Aplicações. UNESP, Presidente Prudente.

12. Gopi, S (2005) Global Positioning System: Principles and Applications. Tata McGraw-Hill Education, New Delhi.