A Methodology for Determining the Data Quality of Geographical Information Systems

Sheyeen Liew

School of Mechanical Engineering

Western Power Corporation

Abstract

There is a general perception that data quality is poor at Western Power. However, an objective quantitative measure is necessary as a first step to improving data quality. This seminar paper will discuss the methods for determining the quality of data in geographical information systems. Data quality consists of components that are mainly concerned that an object is there (completeness), at the right time to be useful (timeliness), in the right position (positional accuracy) and connected correctly to other assets (logical consistency).

Positional accuracy metrics look at measuring, estimating and inferring data quality. Measurement of positional accuracy requires a data set of higher accuracy with which to compare the test data set. If this is not available, then it is possible to examine the process that the data undergoes as it progresses into the system. Each stage in the process can introduce an error which propagates to the end data set. Measuring the error at each stage gives an estimate of the error of the end data set. The third method uses data rules such as the requirement that electrical poles be x metres apart to infer positional accuracy.

Completeness measures involve comparing the test data set to its specifications. Errors of omission (specified data not in test data) and commission (excess data) can be averaged to give a completeness measure.

Timeliness measures give an indication of how up-to-date data is for the task at hand. For example, the elapsed time between an event occurring and it being recorded may be compared against a reference level to give an indication of timeliness.

1.0 Introduction

Obtaining an objective measure of data quality is the first step to improving data quality. This seminar paper will present a method for determining the data quality of WPC's GIS. It is not concerned with actually developing methods of assessing error; it is simply discussing the application of the current literature to WPC's GIS. There is more research to be done on the effect of political, institutional and other human factors on data quality that is not presented in this paper.

This seminar paper will first provide the background and context of the research, and then describe the theory and how it is applied to determine data quality, and finish off with a summary and conclusion.

1.1 Background & Context

This section will discuss the information system that will be investigated, including its significance to WPC's operations. The application of basic concepts of data quality to WPC will be briefly covered to give the reader an understanding of the theory to be developed in the next section.

In the mid-1980's, WPC developed an in-house GIS known as DFIS to manage their electrical distribution network. DFIS is a computer-based model of Western Power's electrical networks. Put simply, it provides the capability to visualize assets in a database geographically. For example, searching under the street name 'Mounts Bay Road' in 'Nedlands' brings up a map of 'Mounts Bay Road' similar to that found on a street map, with the electrical network overlaid on top as in Figure 1. DFIS stores spatial (location) data; the storage of thematic data (information about an object's attributes) is handled by other systems.

DFIS provides information to approximately twenty-one other information systems, for applications such as load flow calculations, fault management and switching strategies. Therefore, it is essential that the DFIS database contain data of high quality in order for these dependent information systems to operate correctly.

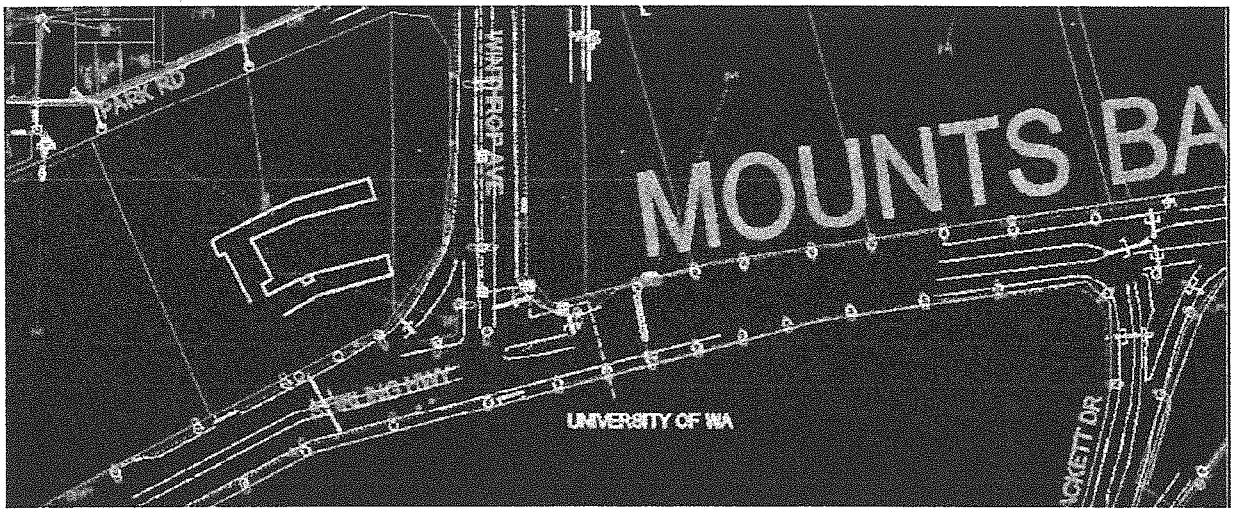

Figure 1: Screenshot of DFIS showing streets and the electrical network overlaid as they are located in the real world

There are several quality components of spatial data, which include positional accuracy, completeness, timeliness and logical consistency. The basic concern is that the object (or asset) is there (completeness), at the right time to be useful (timeliness), in the right position (positional accuracy) and connected correctly to other assets (logical consistency). Due to time constraints, this project looks only at the first three components, as logical consistency is checked automatically by DFIS.

2.0 Theoretical Development

Positional Accuracy

The majority of the positional accuracy models compare the untested data with data of higher known accuracy. For example, Goodchild & Hunter (1997) developed a variation of Langaas and Tveite's (1999) buffer overlay statistics method that applies a buffer of width x around the reference boundary (of known quality) and intersects it with the boundary to be tested, as seen in Figure 2. A cumulative probability distribution may be obtained by varying the buffer width and determining the proportion of the tested source length that lies within the buffer. For example, the 95th percentile gives the distance within which 95% of the length of the tested source lies.

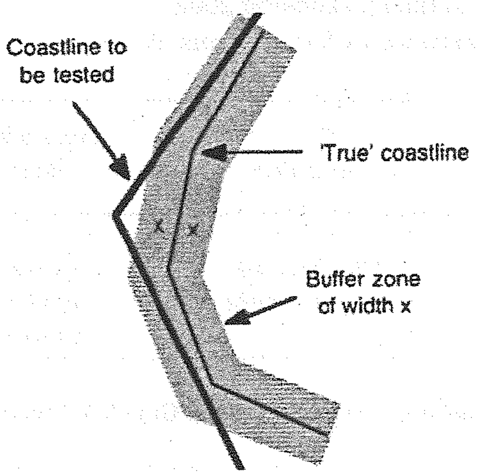

Figure 2: Buffer-overlay-statistics method for determining positional accuracy (from Goodchild & Hunter 1997)

Another method for estimating positional accuracy is error propagation (FGDC 1995). This operates on the assumption that errors are introduced at each stage of the cycle of collection, input and analysis of data, and that they propagate to each subsequent stage in the process. Error propagation is an interesting tool to empirically assess the effect of human factors on the quality of data.

Yet another method that is applicable to Western Power's DFIS is that of data rules. Data rules defines what is allowed or supposed to happen in a data set. They are commonly applied to non-geographical information and GIS attribute information, as they are easier to apply when data is in textual form; however, it is also possible to apply data rules to spatial data. For example, it is a safety rule that wooden poles are at most x metres apart to prevent electric conductors from clashing and arcing (which can result in fires and outages). Therefore, a quick and dirty check can use the assumption that poles are at most (x + T) metres apart, where T = tolerance to allow for obstruction which prevent a pole from being placed exactly x metres apart, for example due to a tree. An example metric would be to calculate the proportion of poles that violate this rule over the poles that comply with this rule.

Completeness

The measures for completeness are basic. A simple count of what is in the test data compared to its specifications (Cai et al. 2003) can be used to determine the proportion of data that is superfluous (error of commission) or missing (error of omission). The challenge lies in determining the area of the electrical grid in Western Australia and which assets to examine for completeness. The proposed sample areas are those that are informally known to be more complete, such as new land developments, examining the most critical assets that are critical to the operation of the network, such as transformers.

Timeliness

The measures for timeliness are also rudimentary. For example (Moellering 1985, 1987):

- The moment of last update
- Rate of change of assets per unit of time
- Temporal validity: summations of the number of data that is out of date, or the number of data that is valid, or not yet valid (ISO 1993:3534-1)
- Validity of data with respect to time (Jakobsson 2002)
- Measurements can be made between different events, defined as:

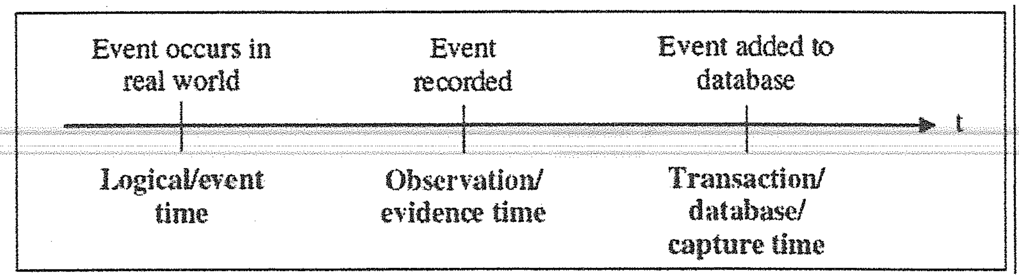
Event occurs in real world
Event recorded
Event added to database
t
Logical/event time
Observation/ evidence time
Transaction/ database/ capture time

Figure 3: Timeliness terminology (from Guptill & Morrison 1995)

Measuring the time elapsed between events is useful in examining the delay in the existing processes and the impacts on the business. A time limit may be set for the time allowed between steps in a process; having an actual measure of this time allows one to observe whether this limit has been breached, or the limits can be reduced.

Several quality component models and measures have been discussed. This includes the buffer-overlay-statistics method, error propagation and the data rule approach to determining positional accuracy. A simple summation can be used to determine completeness, and the time between events for timeliness.

3.0 Conclusion

The effect of poor data quality is gaining increasing awareness in Western Power. It is a costly problem that is difficult to observe, but easy to discern in the problems encountered. Initial steps for improvement require determining the quality of data, and identifying areas that require improvement. Models and measures developed in the literature for determining data quality will be applied to Western Power's GIS, DFIS. Several components of spatial data quality, positional accuracy, completeness and timeliness can be applied. Positional accuracy measures involve comparing the test data set to one of higher, known quality, estimating error through error propagation, and using data rules. Completeness measures involve a summation of the errors of omission and commission of a test data set compared to its specifications. Timeliness measures entail determining the time between different events such as the time elapsed between an event being recorded and it being in the database and ready for use. The results of applying these measures to the DFIS data set and the recommendations for metrics for WPC to apply will be discussed in further detail during the conference.

4.0 References

Cai, Y., Shankaranaryanan, G., Ziad, M. 2003, Evaluating completeness of an information product. Ninth Americas Conference on Information Systems: 2274 - 2281.

Eckerson W. W. (2002). *Data Quality and the Bottom Line*, [Online], The Data Warehousing Institute. Available from: <www.dw-institute.com> [21 February 2005].

Federal Geographic Data Committee (FGDC) 1997, 'Content Standards for Digital Geospatial Metadata Workbook' in *Coordinate Accuracy Reporting for Geospatial Data Sets of Mixed Lineage*, ed D Hansen, 1997 ESRI (Environmental Systems Research Institute Inc.) User Conference.

Goodchild, M. F., Hunter, G.J. 1997, 'A simple positional accuracy measure for linear features', *International Journalof Geographical Information Science*, vol. 11, no. 3, pp. 299-306.

Guptill, S. C., Morrison, J. L. (eds.) 1995, *Elements of Spatial Data Quality*, Elsevier Science Limited, United Kingdom.

ISO 1993, 3534-1: 1993 Statistics – Vocabulary and Symbols Part I: Probability and General Statistical Terms, International Organisation for Standardisation, Geneva, Switzerland.

Jakobsson A 2002, 'Data Quality and Quality Management – Examples of Quality Evaluation Procedures and Quality Management on European National Mapping Agencies', in *Spatial Data Quality*, eds. W Shi et al., Taylor and Francis, pp. 216-229.

Langaas S, Tveite H 1999, 'An accuracy assessment method for geographical line data sets based on buffering', *International Journal of Geographical Information Science*, vol. 13, no. 1, pp. 24-47.

Moellering H 1985, Digital *Cartographic Data Standards: An Interim Proposed Standard*, Report #6, ACSM 1985, National Committee for Digital Cartographic Data Standards, Ohio State University, Numerical Cartography Laboratory, USA.

Moellering H 1987, *A Draft Proposed Standard for Digital Cartographic Data*, Report #8, ACSM 1987, National Committee for Digital Cartographic Data Standards, Ohio State University, Numerical Cartography Laboratory, USA.