Data Quality, e-Government and e-Commerce

Fabián Peter
Èapek Jan
Institute of System Engineering and Informatics,
Faculty of Economics and Administration, University of Pardubice, Pardubice

Abstract:
Recent initiatives, such as e-government in the public sector and e-business in the private sector rely more and more on data. Databases, that were previously independent are now linked and inter-connected via networks. The joint use of variable quality data from multiple sources may increase possibility of failure of many applications, threatening the success of e-initiatives.

Introduction

Large amounts of public and private sector data relate to personal and geo-reference data. Unfortunately, to target the services, it is necessary to count with frequent changes of these data or their misclassification. As persons and firms move and change their addresses, names and businesses, reliability of once collected and recorded data decreases.

If we look on the geographic database as a collection of measurements of phenomena on or near the Earth's surface, we can use many widely used methods for description of errors in observations and measurement. To estimate these errors we can use statistical models of uncertainty and its propagation. There are two elementary types of measurement and the errors to which they are subject. To one type belong the measurements, which take form of classes, e.g. in classification of land use or type of building. To other type belong measurements on continuous scales, such as the elevation in digital elevation model.

Targeted applications require precise information. Recent applications, like e-CRM (Customer Relationship Management) in private sector or e-democracy, geographical analysis based decision making, as a part of e-government initiatives in public sector, require high accuracy.

Solution to the recent quality problems must validate the currency of information in addition to correct definition of location and must support ongoing revision and maintenance of the data.

Errors in classification

Let us consider a single observation of data, whose value serves to distinguish an instance of one class from an instance of another, such as in land use classification or in unique identification of objects. Due to the quality of the photograph used for classification or some other reasons, the class may have been assigned falsely. Let us suppose, that there are four classes into which we divide observation/measurement of some phenomenon. Due to the misclassification, apart from true classification (measurement belonging to class A classified truly as belonging to class A), some measurements may be classified as belonging to some other class. The true class may be determined by check, which is usually more accurate, but usually more expensive or time-consuming.

If we put the numbers of true and wrong classifications into a table (Table 1), we get so called *confusion matrix* in which row *i* represents cases, in which item of Class i was classified as belonging to Class j, where the j is the column index of the matrix. Item *i,i* represents the true classification, so ideally only the items on the principal diagonal would have a non-zero values.

Table 1: Example of confusion matrix

Class	**A**	**B**	**C**	**D**	**Total**
A	50	4	0	5	**59**
B	4	20	5	2	**31**
C	3	5	18	2	**28**
D	1	5	8	43	**57**
Total	**58**	**34**	**31**	**52**	**175**

Users and producers of data look at misclassification in different ways. The columns are called the producer’s perspective, since the task of the producers of data is to minimise entries outside the diagonal cells. The rows are called the consumer’s perspective, because they show what the database content really means, i.e. the accuracy of the database’s contents.

The success rate may be expressed by the proportion of entries in diagonal cells to total number of entries, called the per cent correctly classified (PCC). This measure may be sometimes misleading, so the more useful index of success, called kappa index is often used, defined as follows:

$$\boldsymbol{k} = \frac{\sum_{i=1}^{n} c_{i,i} - \sum_{i=1}^{n} C_i . C^i / C}{C - \sum_{i=1}^{n} C_i . C^i / C} \qquad (1)$$

where $c_{i\ j}$ denotes the entry in row i, column j, the C_i denotes the sum over all columns of row i , C^i denotes the sum over all rows of column i , C is the total of all entries and n is the number of classes.

Errors in continuous data

With the data on continuous phenomena, errors do not represent a change of class, but change of value. The observed or measured value ***x’*** differs from the true value ***x*** in δx. The value of δx should be small and may be positive or negative. It is sometimes necessary to distinguish between **precision** and **accuracy,** which are often confused, due to the several \ways of their definition. Those, who are concerned with measurement, often refer to precision as property of the measuring device/instrument. It is demonstrated through the repeated measurements of the same phenomena. Such an instrument is precise, if it provides repeatedly similar measurements, whether or not they are accurate. Different, more commonly used definition of precision defines it as the number of digits representing the phenomena. It is again not related directly to accuracy. Figure 1 illustrates the difference between the meanings of precision and accuracy.

In Figure 1a repeated measurement of the same phenomenon give similar values, so they are precise, even if their distance from the true value represented by cross show bias from the correct value, they are inaccurate. In Figure 1b the precision is the same, but the distance from the true value is lower, i.e. the accuracy is higher.

In the measurement of the continuous phenomena the magnitude of errors may be described by the root mean square error (RMSE). It is used by the US Geological Survey as its primary measure of the accuracy of elevation in digital elevation models [1].

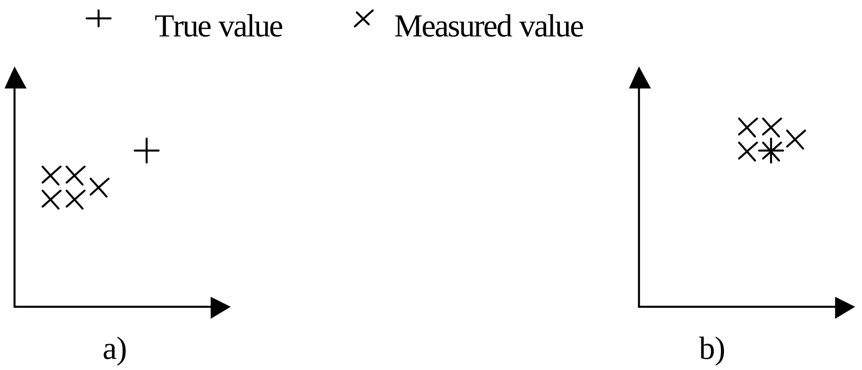

Figure 1: Illustration of difference between **precision** and **accuracy**

RMSE is defined as the square root of the average squared error for all values of *n* observations (2).

$$RMSE = \sqrt{\sum_{i=1}^{n} dx^2 / n} \qquad (2)$$

RMSE captures the magnitude of the average error, but does not show its distribution. It is therefore useful in many cases to show the distribution of errors magnitude. The most common and most important is the Gaussian or Normal distribution. In practice many distributions of error follow the Gaussian distribution. It can serve as a method of prediction of relative abundancies of different magnitudes of error.

Errors in addresses and personal data

The performance of the most up to date information technology infrastructure depends heavily on the quality of data held by the system. Whether the information technology is based on data warehousing, e-CRM, GIS or data mining, the result is the same. If the wrong and incorrect data is used, the results are of limited use or unusable. According to GI News Magazine [2], 52 to 70 percent of all CRM project fail and 92 percent of data warehouses deliver bellow project expectations. The users usually blame the information technology used, but the usual reason of the failure is the data quality fault.

The reason frequently is, that the recent IT applications depend on information extracted from various older (previous century) datasets, prepared for legacy applications. If such legacy data should be relied on, it should be cleaned, corrected and validated.

As local authorities try to use power of the World Wide Web service to deliver round the clock availability and convenience [3], fast delivery, customer focus and personalisation as well as the best online businesses, they are coming against the same data quality barrier.

Good definition of data quality is the “fitness for purpose”. For example the map of Metro in Prague is good for travellers and serves them well with station names and metro lines, bit it would not be suitable for buying a house. This transaction would require precise location data.

Data quality relates to how well data helps to achieve 95 to 100 percent of required objectives.

Fitness for purpose

As data is moved from one application to another it is almost always found to be inaccurate, inconsistent and incomplete. What was good enough for one purpose has to be often cleaned and reworked for another use.

By the turn of the century significant advances in technology and massive reductions of its cost had changed end user definitions of fitness for purpose of data and increased the demands of the users on data quality and level of detail. The affordability of computing power and availability of reliable software has seen an increasing demand for datasets, which are fit for purpose and can support end user work at increasingly detailed level.

Certainly, if e-government and e-commerce are to work, data must be accurate, revised and maintained at desired level of detail. If the analysis tools such as GIS are to deliver evidence-based decision support, they will have to work from that level of detail or use data aggregated from that level, to make it possible to find the data if the evidence is checked.

The shift in focus has served to highlight the technical complexity of what has been so far seen as a simple process- maintaining the currency of property-level information. It has also served to highlight the importance of data standards, which underlie the supply of personal and location data accurate to the desired level of detail. The standards need to reflect the needs of central and local governments and the private sector, all of which use data on people and locations and all of which are involved in e-CRM, data warehousing, advanced analysis and various web-based activities requiring accurate information.

Problems of data usability

As the organisations can no longer rely on the cost of information technology as a barrier to competition, they must increasingly rely for competitiveness on the quality of information they hold. Essentially this means that data is accurate, precise, up to date and well maintained. Unfortunately, personal and locations data can change frequently. If something as basic as person's name or address changes or is misspelled, any initiative concerning the person is likely to fail and any analysis using this data will go wrong. Accepting without question the accuracy of data and taking action based on analysis of such data is an unacceptable risk. All e-government and e-commerce will do is speeding up of transmission of incorrect information to a wider audience.

As research in United Kingdom has shown [2], from three elementary datasets

- a Post Office Address File – all postal delivery points (Post Office),
- a geo-coded address file of all addresses (Ordnance Survey),
- a national land and property gazetteer of all property locations (Improvement and Development Agency –I DeA),

which are cleansed and validated on regular basis, due to the rapid changes on average none exceeds 80-85 per cent accuracy at any time. Their decline in fitness for purpose is proportional to increasing demands for intelligence that enables one-to-one targeting of people and places. Using an 80 per cent accurate address index to validate an 80 per cent accurate client list generates only 64 per cent reliable result. With each data source added inaccuracy increases further.

Of equal importance is to know, how the revisions are made. To be concise, revisions must be in common format or to adhere to common standards.

Elimination of the data problem

First of all, it is necessary to decide, whether the information we want to exploit is fit for purpose. If it is not, then we have to decide the problem. Once the problem is defined, we have to choose the right tools to fix it. WE can choose of several categories of data quality products:

1. Information quality analysis products,
2. Business rule discovery products,
3. Data re-engineering cleansing and transformation products,
4. Information quality defect-prevention products,
5. Metadata management and quality products.

The role of first category products is to extract data, measure qualities such as validity or conformance to business rules, report analysis. The second category deals with the analysis of data to discover patterns and relationships, which define business rules as actually used. Third category serves the purpose to extract standardise, transform, enhance in preparation for data integration or migration to warehouse, CRM, etc. Fourth category products prevent data errors during merging, purging, clean duplicate entries, preventing data errors at point of entry by applying business rules and quality tests. Last category serves the purpose of managing the quality of data about data.

It is important to focus the attention on data, not the hardware or software. Data quality in projects is often ignored and project fails or delivers under plan and expectations. Without the data cleansing, the resources may be targeted in the wrong areas, the campaigns may go wrong, etc.

The perspective

In the long term, it is necessary to develop and maintain national datasets, against which the data can be validated. It seems, that in the future data quality issues will have to be resolved first to create a validated source of information.

Data quality may be defined by different means. As it is very important aspect of GIS use, there have been many attempts to identify its basic dimensions. The US Federal Geographic Data Committee's (FGDC) standards list five components of quality: Attribute accuracy, positional accuracy, logical consistency, completeness and lineage. Definitions of these can be found on the FGDC Web pages [4] or in the text [5].

Data protection, freedom of information and human rights legislation will all have an increasing effect on what can and cannot be done with data about persons and places. Although there will always be tension between the need for information and the need to respect personal rights, legislation will have to find balance and will have to provide possibility to explore the data in the public interest.

Increasingly data resources from the private and public sectors will have to be combined and analysed in order to fight immense problems, such as the crime, rising social costs for unemployment, social care, health care, etc. Such action will require increased cooperation and trust between the public and private sectors and legislative actions, enabling data sharing.

References

[1] Longley, P.; Goodchild, M.; Maguire, D.; Rhind, D.: *Geographic Information Systems and Science*, John Wiley and Sons, ISBN 0 471 89275 0, 2001

[2] McKeon, A.:*What stands in the way of e-government and e-commerce?*, GI News, Vol.2, No. 2, ISSN 1470-1994, 2001

[3] O' Looney, J.: Beyond Maps: *GIS and Decision Making in Local Government*, ESRI Press, ISBN 1 879102 79 X, 2000

[4] *http://www.fdgc.gov* - US Federal Geographic Data Committee's Web pages

[5] Guptill, S.; Morrison, J., L.: *Elements of Spatial Data Quality*, Elsevier, Oxford, 1995

Kontaktní adresa: doc. Ing. Peter Fabián, CSc., prof. Ing. Jan Èapek, CSc.
Institute of System Engineering and Informatics, Faculty of Economics and Administration,
University of Pardubice, Studentská 84, 532 10 Pardubice

tel: 466 036 038, 466 036 510
e-mail: fabi@frdsa.fri.utc.sk, Capek@upce.cz

Recenzoval: doc. RNDr.Bohdan Linda, CSc., vedoucí Ústavu matematiky, FES, UPa