MEASURING DATA QUALITY OF GEOSCIENCE DATASETS USING DATA MINING TECHNIQUES[*]

Cuo Cai[*] and Kunqing Xie

Center for Information Science, Peking University, Beijing 100871, China
**Email*: ccai@cis.pku.edu.cn

ABSTRACT

Currently there are many methods of collecting geoscience data, such as station observations, satellite images, sensor networks, etc. All of these data sources from different regions and time intervals are combined in geoscience research activities today. Using a mixture of several different data sources may have benefits but may also lead to severe data quality problems, such as inconsistent data and missing values. There have been efforts to produce more consistent data sets from multiple data sources. However, because of the huge gaps in data quality among the different sources, data quality inequality among different regions and time intervals has still occurred in the resultant data sets. As the construction methods of these data sets are quite complicated, it would be difficult for users to know the data quality of a dataset not to mention the data quality for a specified location or a given time interval. In this paper, the authors address the problem by generating a data quality measure for all regions and time intervals of a dataset. The data quality measure is computed by comparing the constructed datasets and their sources or other relevant data, using data mining techniques. This paper also demonstrates how to handle major quality problems, such as outliers and missing values, by using data mining techniques in the geoscience data, especially in global climate data.

Keywords: Data quality, Geoscience data, Data mining, Data integration

1 INTRODUCTION

The quality of geoscience data is an issue gaining more and more attention today. This issue is important because a) the methods of producing geoscience data have grown more numerous and b) uses for them have also increased. The increase of data available from earth orbiting satellites is large, not to mention the emergence of sensor networks and other data collecting techniques. On the other hand, use of the Geographical Information System (GIS) makes users able to use and combine more forms of data, some of which they may not be familiar with. In order to allow better usage of geoscience data, its quality needs to be studied more deeply, especially when the data is produced by combining a number of different sources. Although combining sources may cause problems with geoscience data quality, it also offers new approaches to solving the problem.

Users combine datasets for two reasons a) to have more attributes available and b) to get more detailed data of an attribute. In both cases the problem of inconsistency is of the greatest interest. When discussing inconsistency, we face both the inconsistency of the data itself and the inconsistency of data quality. Furthermore, the data quality of geoscience data has a nature of its own: the quality varies both spatially and temporally.

[*] This research is sponsored by China National Advanced Science & Technology. (863) Plan under Grant SQ2006AA12Z214047 and China National Advanced Science & Technology. (863) Plan under Grant 2006AA12Z217 and the Beijing Science Foundation under Grant 4062019.

The CRU dataset, produced by the Climatic Research Unit, University of East Anglia, is a typical example of a multi-source geospatial dataset. Efforts have been taken to level out the inconsistency of the data sources, but the variety of data quality in space and time has not been studied thoroughly. The density of data is much higher for some regions and some periods, and data gathered by different agencies may have significant differences in quality.

The rest of the paper is organized as follows. Section 2 introduces the background of geoscience data quality studies. In section 3, we present ways to estimate geoscience data quality variations in both space and time. In section 4, we present our experiments, and section 5 is the conclusion.

2 BACKGROUND

Geoscience data quality is not a new issue. Almost every researcher producing or using data has worked with it in terms of uncertainty estimation, quality control, etc. Data quality as a whole is discussed as spatial data quality by some researchers and is also an issue in geostatistics. Chrisman (1984) was among the first to bring up concerns of the spatial data quality problem. Aalders (2002) considered the process of how data are derived from the real world and divided this process into 2 steps: conceptualization and measurement. The former is the specification of what should be considered the real world and of abstractions of selected objects; whereas the latter is the specification of the measuring methods and the measurement requirements for capturing the data. Fisher (1999) further distinguished three forms of uncertainty that arise during in the process of deriving a spatial data set from the real world: error, vagueness, and ambiguity. Error is the difference between the measured value and the true value. Vagueness arises from poor definitions, and ambiguity arises because of disagreements about the definition of objects in a spatial dataset. Van Oort (2005) proposed a way of using classification to verify the accuracy of land use data derived from satellite images versus data derived from agricultural censuses. He developed logistic regression models to include a number of explanatory variables.

3 VARIATIONS IN QUALITY OF SPATIAL AND TEMPORAL DATA

There are two main approaches in estimating the quality of a geoscience dataset. One is to crosscheck the dataset with other datasets with similar or related attributes. Another is to find data that are inconsistent with the prescribed pattern. The first method is commonly used by the community. Simmons et al. (2004) suggests that simply comparing data in different charts and tables can explain many inconsistencies. This method is reliable because if two datasets describing the same attributes do not agree, at least one of them is biased. However, in many situations, finding a reference data set with similar attributes within the same region and period is very difficult if not impossible. Often the datasets do not have exactly the same attributes, and a simple comparison does not work. This is where methods such as classification and regression are brought in, as described by van Oort (2005). The second method does not require a reference data set but uses a pattern or model of what the data should look like. As this pattern or model can be viewed as an abstraction of real data, this method is a reduced version of the first method. The validity of this method relies on the validity of the pattern. Data that do not conform to the pattern are called outlier data. We use the term outlier in a broader way than it is commonly used. It can be used not only when a single data element does not conform to the pattern, but also when a structure of data elements does not conform.

Most geoscience data are 3-dimensional, two if only spatial, and one if only temporal. Patterns lie in both space and time. In spite of the differing patterns for data of different attributes, spatial autocorrelation and temporal periods always exist. What is more, geoscience data made of combinations have certain other patterns. For example, if one part of the data has the exact same value, it is highly likely to be filled with default or average values because of the lack of source data on that part.

4 EXPERIMENT

The CRU dataset is a 0.5° latitude by 0.5°longitude gridded dataset of monthly terrestrial surface climate for the period 1901-1996. The dataset consists of a suite of 7 climate elements, including precipitation, mean temperature, etc. The spatial coverage extends over all land areas, excluding Antarctica. The dataset was produced by interpolating the climate station data of thousands of stations from a number of institutions, including national meteorological agencies, WMO, and many others. Although the dataset was constructed in a very sophisticated manner, the fact that the density of original data varies both spatially and temporally has affected the quality of the dataset. As we can see in Figure 1, station density is much higher in some counties. Even in 1981, there are still some areas not covered by any stations (shaded areas were covered by stations).

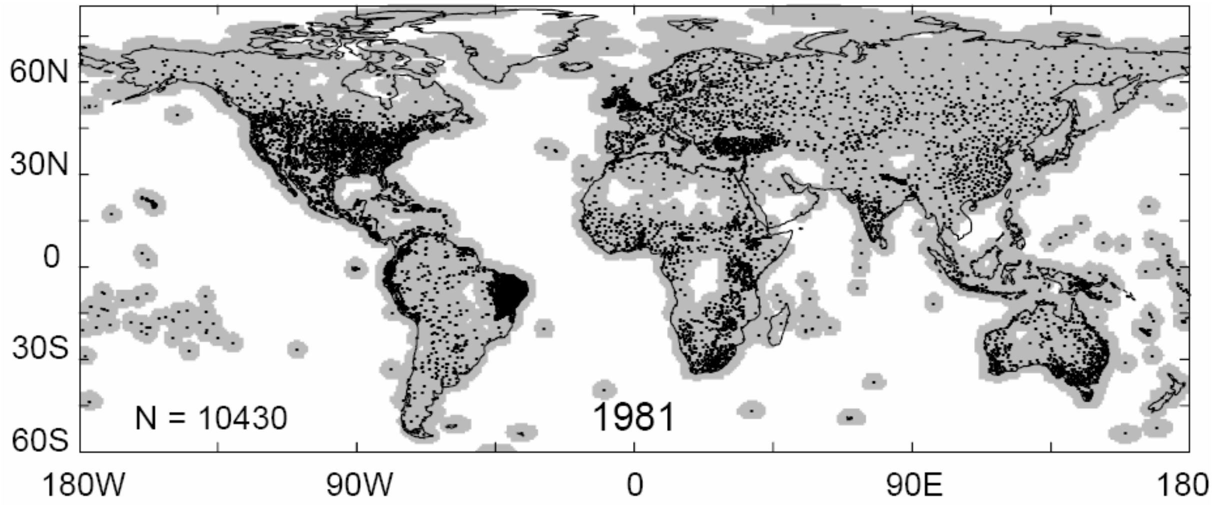

Figure 1. Distribution of precipitation stations in CRU datasets in 1981

Density variation may affect dataset quality in many different ways. Areas of lower density are not necessarily of lower data quality, but they may produce many types of outliers. For example:

a) Some grids may have the same data for many years.
b) Some neighboring grids may have the same data for years.
c) Some grids may have large differences between two periods.

Figure 2 shows global areas in which precipitation has not changed over time. Most of the areas are here because they actually have no precipitation at all; however, areas such as Middle America and islands in the Pacific Ocean are included because there is simply no source data for them.

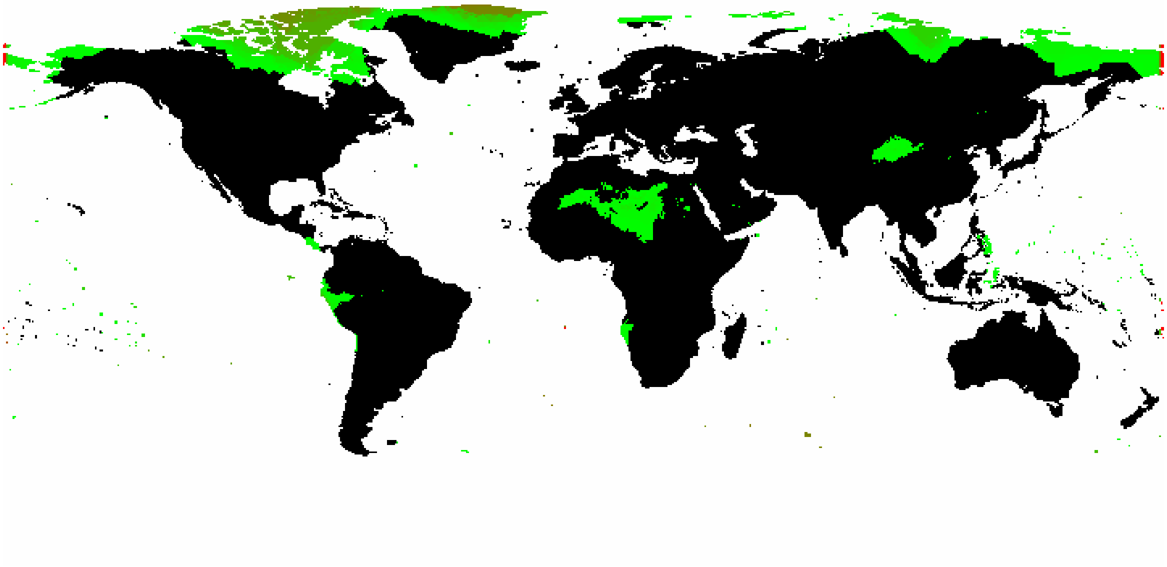

Figure 2. Areas containing non-changing precipitation time series over years.

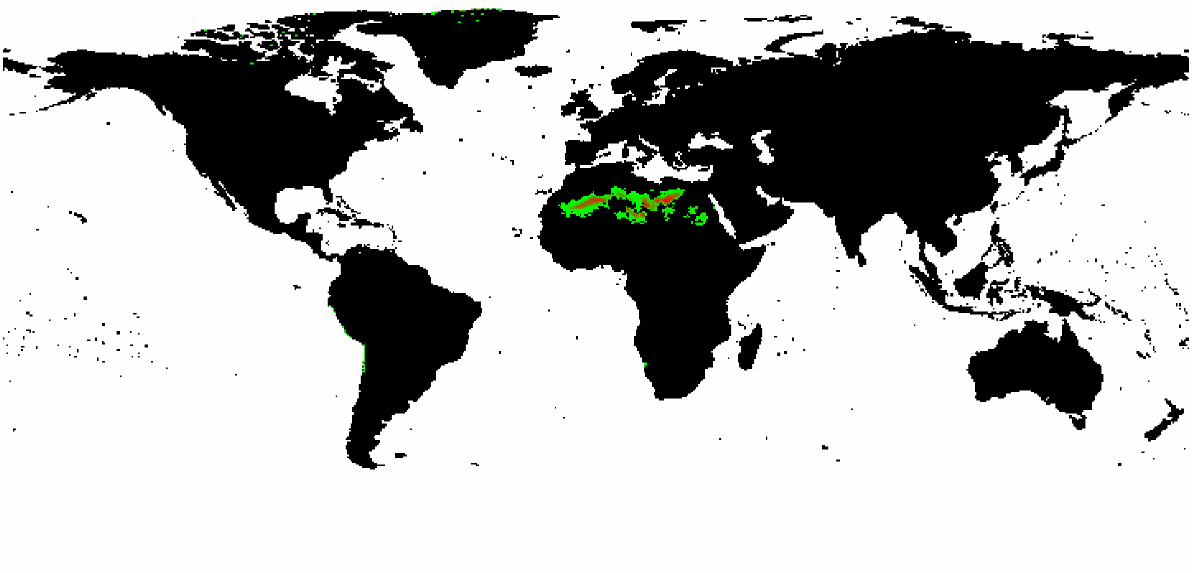

Figure 3. Areas containing the same values in neighboring grids

5 CONCLUSION

While concern is increasing about the quality of geoscience data, there is not enough work being done in estimating this data quality. Geoscience data have a number of properties in common, therefore, many datasets share common patterns. In this paper we have demonstrated methods for estimating data quality using these basic patterns. We discussed the effects of combining and reconstructing data on data quality by presenting a simple experiment.

6 REFERENCES

Aalders, H. (2002) The Registration of Quality in a GIS. In: Shi, W., Fisher P., & Goodchild, M. (eds.) *Spatial Data Quality.* London: Taylor and Francis.

Chrisman, N. (1984) The role of quality information in the long-term functioning of a geographic information system. *Cartographica 21(2-3*): 79-87.

Fisher, P. (1999) Models of uncertainty in spatial data. In: Longley, P., Goodchild, M., Maguire, D., & Rhind, D. (eds.) *Geographical Information Systems: Principles and technical issues 2nd edition, Vol. 1.* New York: John Wiley & Sons.

Goodchild, M. & Jeansoulin, R. (eds.) (1998) *Data Quality in Geographic Information, from Error to Uncertainty.* Paris: Hermes.

Simmons, A., Jones, P., da Costa Bechtold, V., Beljaars, A., Kallberg, P., Saarinen, S. Uppala, S., Viterbo, P., & Wedi, N. (2004) Comparison of trends and low-frequency variability in CRU, ERA-40, and NCEP/NCAR analyses of surface air temperature. *J. Geophys. Res*.: 109.

van Oort, P. (2005) Improving land cover change estimates by accounting for classification errors. *International Journal of Remote Sensing 26(14):* 3009-3024.