Detecting Quality Defects in Soil-Sampling Data

Yotam Rotholz
Supervised by Dr. Adir Even
Department of Industrial Engineering and Management, Ben-Gurion University of the Negev

In collaboration with Shay Mey-Tal
Agam Advanced Agronomy, Meggido

Abstract:

The agricultural sector is lagging behind others with respect to data collection and analysis (Stafford, 2000; Khanna et al., 1999; Gandonou et al., 2003). However, the growing awareness to the benefits of precision agriculture (PA) (Silva et al., 2011) promotes growing interest in establishing reliable agricultural data infrastructure. This research deals with the detection of data quality (DQ) defects in such data. The focus is on spatial datasets, which usually stored as layers in Geographical Information System (GIS) (Bandyopadhyay et al., 2009; Lan et al., 2009). The test-case dataset reflects soil-sampling measures over several years from adjacent plots in Jezreel Valley. Such data is used frequently for assessing chemical and physical characteristics of land plots and supporting important decisions accordingly – irrigation, fertilizing, setting plantation plans, etc. DQ defects, such as incompleteness or inaccuracies, are quite common in such data – and are apparent in the sample dataset used for this research.

Research has offered plethora of metrics for assessing DQ from different "dimensions" (completeness, currency, etc.). The reliability dimension, the scope of this research, has rarely been studied, and no relevant quantitative metrics could be found for it. In the agricultural context, reliability is of critical importance – does the dataset reflect reliable data acquisition? Another unique perspective offered by this research, is the observation of DQ both in the context of time gap and the context of spatial distribution.

A key contribution is the development of an analytical dynamic model for data reliability assessment, based on the Cobb-Douglas production function. The model is defined in a generic manner, which may fit DQ handling in different forms of spatial data. The model estimates the reliability of a certain data item, considering defects that may be inferred by

observing the item itself (e.g. violation of the value domain), as well as defects that be detected using corresponding values taken earlier (e.g., by using techniques of statistical process control) and/or in adjacent locations (e.g. by spatial-autocorrelation assessment). The detection is based on the assumption that a data value collected from a location would be similar, but not identical, to precedent or adjacent corresponding values.

This research is conducted as part of a broader research initiative, which aims at supporting the use of agricultural information with cloud-computing. The intention is integrating DQ indicators and other outputs as layers in a cloud-based GIS. A preliminary model evaluation, with real-world samples, showed encouraging results, and contributed insights toward future improvements.

Soil samplings indicate on the fertility state of plots, which support farmers in decisions of fertilizing type and rate. Such decisions are quite cardinal and complex, as they involve a large number of factors to be considered: on the one hand, excessive fertilization comes at a high cost, without necessarily contributing to greater crops yield. On the other hand, under-fertilization might cause substantial degradation in crops yield and major revenues loss for the farmers.

Currently, the model is still underdoing development and enhancement. In parallel, we have collected and organized large-scale sample of real-world data regarding factors that may affect fertility. This data includes fertilizers type and rate, crop type and magnitudes, and the fertility-ingredients rates consumption per crop. The intention is to split the sample into a training set for estimation of model parameters versus a test set for validating model performance using the fertility data collected, where a potential use of model output is to predict fertility rates per plot. The ultimate goal of this study is to make some contribution toward a solution that would increase farmers' trust in their data sources and reduce their uncertainty regarding the reliability of the soil samplings and, by that, take better decisions.

References

Bandyopadhyay, S., Jaiswal, R., Hegde, V., & Jayaraman, V. (2009). Assessment of land suitability potentials for agriculture using a remote sensing and GIS based approach. *International Journal of Remote Sensing, 30*(4), 879-895.

Gandonou, J., Dillon, C. R., Kanakasabai, M., & Shearer, S. (2003). Precision agriculture,

whole field farming and irrigation practices: A production risk analysis. Paper presented at the *Southern Agricultural Economics Association Meeting (SAEA), Mobile, Alabama.*

Khanna, M., Epouhe, O. F., & Hornbaker, R. (1999). Site-specific crop management: Adoption patterns and incentives. *Review of Agricultural Economics, 21*(2), 455-472.

Lan, Y., Zhang, S., Li, W., Hoffmann, W., & Ma, C. (2009). Variable rate fertilization for maize and its effects based on the site-specific soil fertility and yield. *Agricultural Engineering International: CIGR Journal*

Silva, C. B., de Moraes, Márcia Azanha Ferraz Dias, & Molin, J. P. (2011). Adoption and use of precision agriculture technologies in the sugarcane industry of são paulo state, brazil. *Precision Agriculture, 12*(1), 67-81.

Stafford, J. V. (2000). Implementing precision agriculture in the 21st century. *Journal of Agricultural Engineering Research, 76*(3), 267-275.