

Abstract for United Nations 2009 Statistical Commission Session on Innovations in Official Statistics

Innovative Uses of Data Mining Techniques in the Production of Official Statistics
Jaki McCarthy, Dale Atkinson, Et al

Data mining techniques are used to find patterns, classify records, and extract information from large data sets. These techniques, often used in the private sector for market research, fraud detection, and customer relationship management, can also be used by statistical agencies to analyze their large survey data sets. While large datasets are common in many statistical agencies, data mining techniques have not been widely used to improve the production of official statistics. However, innovative applications of these techniques can be very effectively used in this regard through improving survey data, processing and estimation.

There are many potential applications of data mining techniques with large survey data sets. This paper will discuss data mining applications already underway in NASS, as well as several proposed future data mining projects. NASS has used classification trees to model survey reporting errors, identify records most likely to be population members (for NASS, agricultural operators), predict survey non-respondents, and to create non-response weighting groups. In the future, cluster analyses may be used to suggest possible subgroups of respondents for which data collection may be tailored. Association analyses may have potential applications in designing data editing systems. These and other potential applications of the broad range of data mining techniques for improvements in data collection and processing will be discussed.