25th Meeting of the Wiesbaden Group on Business Registers
- International Roundtable on Business Survey Frames

Tokyo, 8 – 11 November 2016

Michael E. Kornbau
U.S. Census Bureau
Session No. 5

Technology

Automating Processes for the U.S. Census Bureau Business Register

Abstract

The U.S. Census Bureau has replaced or improved upon clerical operations for Business Register data using a combination of machine learning and rules-based automation processes. It also has taken an approach to modify data collection instruments to improve and enhance the automated processes. The most noteworthy example is the automation of assigning six-digit industry classifications to new businesses applying for an Employer Identification Number (EIN), achieving an 80 percent automated coding rate. The process includes automated dictionary creation based upon manually-coded records using business name and description; a logistic regression model to assign a probability score to potential codes; a web-based instrument to collect business data from applicants in a way that complements and improves upon the automated coding; and a quality control process to monitor coding quality and to update coding dictionaries with new terms. The quality goal is to maintain at least comparable quality to 100 percent manual coding. Cost savings were initially estimated to be $1 million per year in 2004. The Census Bureau uses a similar, but simpler, process for assigning legal form of organization (LFO) based on business name, when LFO cannot be assigned through other means, and for assigning industry classifications based on written tax form descriptions. More recent work in automating manual processes include text analytics of various write-in fields that uses unsupervised learning clustering techniques, and the evaluation of incoming administrative record data using a common approach for each source combined with the use of standards to identify unusual data. This paper will provide a summary of these efforts, some of the challenges encountered to implement them, the benefits, and future endeavors to reduce manual processing of Business Register data.