Title: **Generating Industrial Classification Codes from Public Data**

Authors: John Cuffe and Sudip Bhattacharjee

Keywords: Public data, Administrative records, NAICS codes, Machine Learning, Natural Language Processing.

Abstract

The North American Industrial Classification System (NAICS) Codes are a critical component in the U.S. economic measurement infrastructure; in this paper, we propose an improvement to the generation of these codes that could improve the quality of these codes and the efficiency of the generation process. The NAICS codes serve as a basis for survey frames and published economic statistics. In the current state, multiple statistical agencies and bureaus generate their own codes (e.g. Census Bureau, Bureau of Labor Statistics (BLS), and Social Security Administration) which can introduce inconsistencies across datasets housed at different agencies. For example, the business list comparison project undertaken by BLS and the Census Bureau found differences in classification even for single-unit establishments (Elvery et al. (2006)) No indisputable ground-truth dataset exists to adjudicate between differences between alternative sources, and methods for generating codes for new businesses rely heavily on manually-maintained dictionaries and models with onerous statistical assumptions.

We propose that modern web-scraping and machine learning techniques can solve many of these issues, providing more accurate NAICS codes for lower costs while also reducing burden on businesses to identify their industry on survey forms. When these processes are repeatable and adopted by agencies, they would improve consistency over time and datasets. Specifically, we combine data from the Census Bureau's Business Register (BR), IRS Employment Identification Number (Form SS4) application forms, and the Google Places API to gather information such as user reviews and website text on over 100,000 single-unit establishments present in the BR between 2012 and 2016. We use Natural Language Processing techniques to generate numerical representations of the information contained in public user reviews and website text for these businesses, and then use this information to generate a machine learning model for predicting 2-digit NAICS codes. Our approach shows substantial improvement over current methodologies for generating NAICS codes, and demonstrates the possibility of combining government or administrative data records with rich textual data to improve data quality and generate new statistical products.