Valuing the U.S. Data Economy using Machine Learning and Online Job Postings: Implications for Measured Production and Productivity

Dylan Rassier and José Bayoán Santiago Calderón
Bureau of Economic Analysis

The current global efforts for revising the System of National Accounts have identified valuing and recording data as a high-priority area. The emergence and growth of data-enhanced and data-enabled businesses such as online platforms have contributed to the status of data as a vital asset in modern economies. These ongoing efforts include developing a taxonomy to identify the scope of data assets, measure them, and incorporate them into the national accounts (OECD 2021). Current guidance encourages national statistical offices to explore and share potential methods, estimates, and conceptual bases used in these efforts.

The Bureau of Economic Analysis has applied an unsupervised machine learning algorithm to estimate the labor costs of data-related activities using online job advertisements. Our method augments the traditional labor costs methodology by proxying time-use factors using only the language in job postings text. Using this method, we estimate data-related (nominal) labor costs grew from $100 billion in 2010 to more than $200 billion in 2018, representing an average annual growth rate of 9.7 percent.

The occupations' time-use factors can be decomposed into two components: (1) the average share of time allocated to data-relevant tasks and (2) the share of employees engaging in those job activities. Using Burning Glass Technologies (BGT) job advertisement data, we identify which skills in the BGT taxonomy are "data-related" as relevant to data entry, storage, analysis, or management. We identify occupations with the highest rate of job openings containing "data-related" skills. The top occupations are denoted as "known" data-intensive occupations and serve as "landmark" occupations (e.g., statisticians). A doc2vec model is trained on the job advertisement text for each occupation to obtain a high-dimensional representation of what the occupation-level job postings convey. Using this numerical representation, we can obtain occupation-level pair-wise distances to measure how "close" or similar an occupation is to the "known" data-intensive occupations. The product of the similarity to a landmark measure and the ratio of job openings with identified "data-relevant" activities serves as the proxy of the occupation-level time-use factor. The BGT data also allow us to estimate and adjust for overlap with other intellectual property products currently captured in the U.S. national accounts as capital formation, including own-account software and own-account R&D.

If incorporated into the U.S. national accounts, data as an asset would amount to expanding the production boundary as part of the own-account software and databases category of intellectual property products, which currently excludes the value of the embedded information content (i.e., data). Given the relative size and growth of a complete cost estimate for data as an asset, this paper will build stocks of data as an asset and analyze implications for the U.S. measured production and productivity.

Reference

Organization for Economic Cooperation and Development (OECD). 2021. “An Update on Recording and Measuring Data in the System of National Accounts.” Paper prepared for the Advisory Expert Group on National Accounts.