T. Kirk White
U.S. Census Bureau
Center for Economic Studies
Thomas.Kirk.White@census.gov
https://sites.google.com/site/tkirkwhite/

February 24, 2023

Dear Members of the Robert Summers Fellowship Committee:

I am writing to apply for a Robert Summers Economic Measurement Fellowship. I have spent much of the past 15 years of my career working on economic measurement both at the Census Bureau and at USDA's Economic Research Service. Most of my self-directed research over this period has focused on the effects of statistical agencies' editing and imputation in economic data. Although most of this work has necessarily focused on U.S. data, in recent research my coauthors and I have shown that differences in data cleaning across national statistical agencies have very large effects on international comparisons of allocative efficiency and measures of productivity dispersion.

In addition to my self-directed research, I have also worked on internal projects creating new imputation models for survey data; redesigning the Census Bureau's Longitudinal Business Database (LBD); adding new content to the Economic Census; creating synthetic business microdata; and (currently) helping to construct the sample design and imputation cells for the Census Bureau's new Annual Integrated Economic Survey (AIES). I provide more details and links to papers and other documentation in the attached summary of my expertise.

In my work on internal projects at the Census Bureau, I have seen many examples of measurement and post-collection processing decisions that affect economic data in ways that many academic economists are unaware of. Academic research economists can benefit from understanding more about what economic statisticians at statistical agencies are doing to data *before* it reaches the statistical agency's website or Research Data Center. As key users of economic data, academic economists can in turn help statistical agencies improve their products by providing feedback on the suitability of the data for particular purposes, and by suggesting data that the agencies should be collecting but are not.

In short, I think the aims of the Robert Summers Fellowship to promote research on economic measurement and to strengthen ties between the academic and practitioner communities are excellent goals. I hope you will agree that my research and work on internal projects at statistical agencies over the past 15 years has made me well-qualified to further these goals.

Sincerely,

T. Kirk White

Summary of Relevant Expertise and Recent Research in Economic Measurement

Effects of Editing and Imputation on Measured Productivity Dispersion, Allocative Efficiency, and Productivity Dynamics (2012-present). Much of my self-directed research over the past 10 years has focused on the effects of the Census Bureau's editing and imputation on economic measurements in the Census of Manufactures (CMF) and the Annual Survey of Manufactures (ASM). Researchers in the FSRDCs have known for decades that *some* of the data in the CMF and ASM are imputed. However, because it was difficult to identify imputed data, until my research with coauthors it was not widely known outside the Census Bureau how *much* of the data is imputed. In White, Reiter and Petrin (2018) we show that over 70% of plants in the mail sample of the 2002 and 2007 CMFs have imputed data for at least one variable needed to calculate revenue Total Factor Productivity (TFPR). The Bureau primarily imputes for missing values using mean-imputation methods which can reduce the true underlying variance of the imputed variables. For 5 variables entering TFPR in 2002 and 2007 we show the dispersion is significantly smaller in the Census mean-imputed versus the Census non-imputed data. We build alternative imputation models using non-parametric Classification and Regression Trees (CART), and show that for 90% of the 473 industries in 2002 and 84% of the 471 industries in 2007 TFPR dispersion increases as we move from Census mean-imputed to Census non-imputed to the CART-imputed data. In the CART-completed data 66% (2002) and 51% (2007) of industries have 75-25 TFPR ratios that are at least 10 log points higher than in the Census-completed data, suggesting U.S. TFPR has more dispersion than has been previously thought.

In subsequent work with Martin Rotemberg, we take advantage of access to the raw reported data for the 2002, 2007, and 2012 CMFs and look at the effects of all types of Census editing and imputation (not just the mean-imputation that was the focus of the WRP paper) on measured allocative efficiency (AE). We compare these estimates to measured AE from India's Annual Survey of Industries. We document that (1) measured allocative efficiency is 4 to 80 times higher in the CMF after the Census Bureau's editing and imputation; (2) India's statistical agency does no imputation and no systematic editing in its annual manufacturing survey; and (3) applying the same data cleaning methodology to raw data from both countries results in higher measured allocative efficiency in *India* than in the U.S. While we do not take these results at face value, they do cast doubt on the highly cited results of Hsieh and Klenow (2008)—which find India's and China's manufacturing productivity would be 40-60% higher if it had the same level of allocative efficiency as the U.S. More broadly, our results highlight the importance of paying attention to differences in statistical agencies' data cleaning when doing international comparisons. In ongoing research with Hang Kim, Martin and I are extending this editing and imputation methods to panel data from the Census Bureau's ASM and India's Annual Survey of Industries.

- **White, T. Kirk**. 2014. Recovering the Item-level Edit and Imputation Flags in the 1977-1997 Censuses of Manufactures. CES Working Paper CES-WP-14-37. https://www2.census.gov/ces/wp/2014/CES-WP-14-37.pdf
- **White, T. Kirk**, Jerome P. Reiter, and Amil Petrin. 2018. Imputation in U.S. Manufacturing Data and Its Implications for Productivity Dispersion, *The Review of Economics and Statistics*. DOI: http://dx.doi.org/10.1162/REST_a_00678
- Rotemberg, Martin and **T. Kirk White**. 2020. Measuring Cross-Country Differences in Misallocation. Center for Economic Studies Working Paper CES-WP-16-60R. https://www2.census.gov/ces/wp/2016/CES-WP-16-50R.pdf

LBD Redesign and Production Implementation Project (2015-2021). With a team of economists, programmers, and subject matter experts, I helped redesign the U.S. Census Bureau's Longitudinal Business Database (LBD) and transition it a production environment. The LBD is used to create the Business Dynamics Statistics (BDS), tabulations describing the entry, exit, expansion, and contraction of businesses. The redesigned LBD and BDS also incorporate information formerly provided by the Statistics of U.S. Businesses program, which produced similar year-to-year measures of employment and establishment flows. In Chow et al. (2021), we describe in detail how the LBD is created from curation of the input administrative data, longitudinal matching, retiming of economic census-year births and deaths, creation of vintage consistent industry codes and noise factors, and the creation and cleaning of each year of LBD data. This documentation is intended to facilitate the proper use and understanding of the data by both researchers with approved projects accessing the LBD microdata and those using the BDS tabulations.

- Chow, Melissa, Teresa C. Fort, Christopher Goetz, Nathan Goldschlag, James Lawrence, Elisabeth Ruth Perlman, Martha Stinson, and **T. Kirk White**. 2021. "Redesigning the Longitudinal Business Database", CES Working Paper Number CES-21-08 https://www.census.gov/library/working-papers/2021/adrm/CES-WP-21-08.html

Economic Census Synthetic Data Research Project (2018-2020)**.** In this 2-year research project with a team of statisticians I helped develop edit-imputation models for 42 industries in 18 sectors of the Economic Census using Dirichlet Process mixture of normals models. The method flexibly models the joint distributions of revenue, payroll, and employment within a given industry and simultaneously ensures that the imputations satisfy the Census Bureau's industry-specific ratio edit rules. For each industry we created 5 implicates (edited and partially imputed datasets) and from each of those implicates we used the same models to create 5 synthetic datasets, with a goal of creating datasets that have high utility while still providing privacy protection. This proof-of-concept research project is described in detail in the project's final report:

- Thompson, Katherine J., Hang Joon Kim, Noah Bassel, Kevin Bembridge, Charles Coleman, Michael Freiman, Maria Garcia, Stephen Kaputa, Steven Riesz, Phyllis Singer, Eric Valentine, **T. Kirk White**, and Daniel Whitehead. 2020. "Final Report: Economic Census Synthetic Data Project Research Team." ADEP Working Paper ADEP-WP-2020-05. https://www.census.gov/content/dam/Census/library/working-papers/2020/econ/ADEP_WP_2020_05Jennyspaper.pdf

Other Economic Measurement-Focused Work.

- Basker, Emek, Randy Becker, Lucia Foster, **T. Kirk White**, and Alice Zawacki. 2019. Addressing Data Gaps: Four New Lines of Inquiry in the 2017 Economic Census. CES Working Paper CES-WP-19-28. https://www2.census.gov/ces/wp/2019/CES-WP-19-28.pdf
- Agricultural Resource Management Survey (ARMS) Imputation Project (2009-2011)**.** Helped develop a new imputation model for the ARMS. See "NASS Imputation" section of the ARMS page of the ERS website (https://www.ers.usda.gov/data-products/arms-farm-financial-and-crop-production-practices/documentation/#dataprocessing). See also https://www.niss.org/research/niss-nass-cross-sector-research-residence-program