Revisions to the LEHD Establishment Imputation Procedure and Applications to Administrative Jobs Frame

by

Lee Tucker
U.S. Census Bureau

Moises Yi
U.S. Census Bureau

Filip Babalievsky
U.S. Census Bureau

Hubert P. Janicki
U.S. Census Bureau

Stephen R. Tibbets
U.S. Census Bureau

Lawrence Warren
U.S. Census Bureau

CES 24-51 **September 2024**

The research program of the Center for Economic Studies (CES) produces a wide range of economic analyses to improve the statistical programs of the U.S. Census Bureau. Many of these analyses take the form of CES research papers. The papers have not undergone the review accorded Census Bureau publications and no endorsement should be inferred. Any opinions and conclusions expressed herein are those of the author(s) and do not represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. Republication in whole or part must be cleared with the authors.

Papers in the series are posted on www.census.gov/library/working-papers/series/ces-wp.html. For information about the series, contact Sean Wang, Editor, Discussion Papers, U.S. Census Bureau, Center for Economic Studies, 4600 Silver Hill Road, Washington, DC 20233, CES.Working.Papers@census.gov.

Abstract

The Census Bureau is developing a “jobs frame” to provide detailed job-level employment data across the U.S. through linked administrative records such as unemployment insurance and IRS W-2 filings. This working paper summarizes the research conducted by the jobs frame development team on modifying and extending the LEHD Unit-to-Worker (U2W) imputation procedure for the jobs frame prototype. It provides a conceptual overview of the U2W imputation method, highlighting key challenges and tradeoffs in its current application. The paper then presents four imputation methodologies and evaluates their performance in areas such as establishment assignment accuracy, establishment size matching, and job separation rates. The results show that all methodologies perform similarly in assigning workers to the correct establishment. Non-spell-based methodologies excel in matching establishment sizes, while spell-based methodologies perform better in accurately tracking separation rates.

[*] Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau. All results have been reviewed to ensure no confidential information is disclosed (DMS Project 7515812: CBDRB-FY24-0247).

1 Introduction

As part of the Census Bureau's transformation program, the Bureau is currently producing a set of prototype frames. Each frame is intended to provide a comprehensive portrait of one type of population or economic activity, and each does so by incorporating information from a range of administrative and survey sources. Collectively, these frames will eventually be able to provide a view of population activity that is broader in scope than any single product in the Federal statistical system.

Among these frames will be a comprehensive measure of employment (the "jobs frame"). The jobs frame will provide detailed, job-level information on the universe of administratively recorded employment in the United States, longitudinally on an annual basis. To do this, it will link administrative records from several sources including unemployment insurance (UI) records and IRS W2 tax filings. A prototype of the jobs frame is scheduled to be complete at the end of Fiscal Year 2024. Characteristics of each job in the frame will include the specific employer, total earnings, and the location and industry of employment; additional worker-level, employer-level, and geography-level information will be obtainable through linkages to other administrative frames.

The closest current analogue to the jobs frame is provided by the Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program. Like the prototype jobs frame under development, LEHD uses administrative UI records to report job-level earnings. LEHD also reports information on the industry and geography of employment, and it does so by imputing a specific establishment of employment for each employee within each employing firm. The parameters of this imputation are derived from a statistical model trained on a single state (Minnesota) which has provided establishment of employment as part of its UI records. Establishment-level information on location and total employment from the Quarterly Census of Employment and Wages (QCEW) are then used to assign each remaining worker in all other states to an establishment. From that establishment, the industry and geography of employment are drawn. Within the LEHD infrastructure, this methodology is referred to as the Unit-to-Worker (U2W) imputation.[2] For an overview of the U2W imputation procedure, see (Abowd, et al. 2009).

This technical paper summarizes the research of the jobs frame development team as it pertains to modifying and extending the U2W imputation procedure to the jobs frame prototype product. The paper proceeds as follows. Section 2 provides a conceptual outline of the imputation method, a brief overview of the construction of the training and validation framework, and a description of several key problems and tradeoffs faced in using this method. Section 3 described the four primary methodologies that we compare in the results section of this paper, and Section 4 describes the measures used to evaluate the performance of each methodology. Section 5 shows the empirical results.

[2] Throughout this technical note, we will use the terms "Unit-to-worker" or "U2W" to refer to the existing LEHD imputation procedure.

2 Conceptual Outline and Training/Validation Framework

Both the existing and revised unit-to-workers (U2W) imputations are based on a classic conditional discrete choice framework (McFadden 1973). In this framework, each worker has already been matched to a firm, but the worker must make a choice of which establishment within the firm they will work for, and the choice of establishment is solely made by workers.[3] For each candidate establishment, the worker has a utility function that takes into account the cost of commuting to that establishment, any establishment-level characteristics that all workers value, and an idiosyncratic term that will lead workers to prefer different establishments.

Under standard parametric assumptions about the idiosyncratic term, the probability that worker i at time t chooses establishment j from candidate set $R(it)$ can be written as:

Equation 1

$$p_{ijt} = \frac{e^{\alpha_{jt} + \beta_t X_{ijt}}}{\sum_{k \in R(it)} e^{\alpha_{kt} + \beta_t X_{ikt}}}$$

Each of the candidate models shown in this technical note uses a variation on this basic discrete choice framework. They vary in their construction of covariates X_{it} , in their construction of candidate establishment set $R(it)$, or both. They also vary in their estimation and imputation methods. Each specific candidate model is described in Section 3.

The overall procedure can be characterized as taking place in two steps, which we will refer to as the estimation step and the imputation step. In the estimation step, a conditional logit estimator (or similar estimator) is used to estimate the distance parameters of the model $\hat{\beta}_t$ on an estimation sample of workers whose true establishment of employment is known. The set of establishment shifters $\hat{\alpha}_{jt}$ may also be estimated at the same time, or as is the case for the existing U2W imputation, a proxy covariate may be used. In the imputation step, the corresponding impact of distance $\tilde{\beta}_t$ is fixed as $\hat{\beta}_t$, and establishment shifters $\tilde{\alpha}_{jt}$ for establishments in the imputation sample are either constructed by using a proxy covariate or estimated by maximum likelihood.[4] Then, since commute distance to each candidate establishment is known, we can use Equation 1 to generate the conditional probabilities $\tilde{p}_{ijt} \equiv$

[3] In practice, we might believe that an individual's choice of firm and of within-firm establishment are made simultaneously, or that a given job offer might not permit employment in all establishments within the firm. However, our administrative records sources already provide the precise firm of employment, and do not provide any information about possible establishments of employment within the firm at any given time. So, this discrete choice framework aligns closely with the information we have on hand.

[4] The existing U2W methodology introduces some Bayesian features into this strategy to better reflect the notion that $\hat{\beta}_t$ from the estimation sample is a Bayesian prior for $\tilde{\beta}_t$, the impact of distance on the imputation sample. That is, an imputation value of $\tilde{\beta}_t$ is drawn from a multivariate normal distribution with means $\hat{\beta}_t$ and variance-covariance matrix as provided by the estimation of $\hat{\beta}_t$. Similarly, since the proxy establishment shifters $\tilde{\alpha}_{jt}$ are a Bayesian prior for the true establishment effects on the imputation sample, they are drawn from a Dirichlet distribution with parameters corresponding to the percentage of month 1 employment within each firm that is attributable to each establishment. These Bayesian assumptions are conceptually correct, though they have minimal impact on the ultimate performance of the imputation procedure. For simplicity and tractability, we have dispensed with the previous Bayesian assumptions made in the U2W procedure for this imputation exercise.

$\Pr\left(j = J(i,t) \middle| X_{ijt}, \tilde{\alpha}_{jt}, \tilde{\beta}_t\right)$. The CDF generated by these probabilities is then used to choose a unique imputed establishment for each job.

Construction of the Training and Validation Framework

To evaluate the performance of our candidate model methods, we split the data from Minnesota into training and validation datasets. Validating our model on Minnesota data allows us to compare the individual establishment match rates against their true values, while still permitting the evaluation of more aggregate measures of model performance.

Sampling into training and validation datasets proceeds by complete firms. For each firm (SEIN), a single uniform draw was taken, and based on this draw, 50% of firms were assigned to each of the training and validation samples. The training and validation datasets consist of all employment in the associated firms, respectively. There are two main reasons for conducting this sampling at the firm level. The first is that several of our candidate specifications use a standard conditional logit specification in which they estimate the establishment shifter parameters $\hat{\alpha}_{jt}$ directly. Although those parameters estimated on the training dataset are not used for the validation step, these models have the most statistical power when they are estimated on the entire employment of a firm. The second reason for sampling at the firm level is that several of our candidate models impute establishment at the spell-level rather than on a quarter-by-quarter basis. Sampling at the firm level ensures that every worker's entire job spell history is included in the same file.

For all the results described in this note, the year 2015 was chosen as a representative year for evaluating the performance of the framework, and so estimation and imputation were only performed on jobs in the four quarters of 2015. However, all years of Minnesota LEHD data were processed in preparation for this exercise. So, where applicable, imputation methods define job spells and establishment existence spells based on the full set of observed data in all years.

Conditional Logit Specification: Attenuation Bias, Computational Concerns, and Data Availability

It is a known feature of logistic regression models that they exhibit attenuation bias in the event of model misspecification. An omitted covariate that is relevant (correlated with the dependent variable) will always be a source of attenuation bias in the parameter of interest, even if the omitted covariate is orthogonal to the covariate of interest (Mood 2010). In this context, this means that an omitted variable could lead our estimates of the impact of commute distance on establishment choice to be biased toward zero, and this could lead to an over-abundance of long commutes relative to true commuting patterns. Previous research comparing commuting patterns to those reported by respondents to the American Community Survey have shown a substantially higher rate of cross-county commuting in LEHD, most of which is seemingly attributable to the existing establishment imputation (Green, Kutzbach and Vilhuber 2017). Moreover, after finding longer commute distances on even a matched sample where the ACS-reported work location corresponds to a candidate establishment, Green, Kutzbach and Vilhuber conclude that "the U2W impute probabilities for distant establishments may be biased upwards, indicating a potential misspecification of the model or inappropriate constraints for selecting candidate establishments for a job." This finding suggests that attenuation bias may be a substantial problem with the existing U2W imputation model.

One nice feature of the McFadden conditional logit framework is that the establishment shifters α_{jt} control flexibly for *any* factors or policies that make some establishments more attractive to employees than others, regardless of their nature, as long as the impact of those factors on employees' utility is homogeneous. This would seem to greatly mitigate any concerns of attenuation bias. However, it turns out that even small simplifications made to the conditional logit framework are enough to reintroduce the attenuation bias problem, even if the model is otherwise appropriately specified. In particular, any imputation model that does not estimate $\hat{\alpha}_{jt}$ simultaneously to $\hat{\beta}_t$ will exhibit such a bias, and this unfortunately includes the existing U2W imputation model. The reason is that—assuming that our discrete choice framework is correct—the true value of α_{jt} depends in part on an average of all employees' values of $\beta_t X_{ijt}$; even if the difference between a proxy variable $\tilde{\alpha}_{jt}$ and α_{jt} is uncorrelated with X_{ijt}, it will be a source of attenuation bias simply because it is correlated with individuals' establishment choices. The case of this specific issue as it applies to the existing U2W imputation is shown in Appendix B.

In contrast, any attempt to estimate $\hat{\alpha}_{jt}$ alongside $\hat{\beta}_t$ also has two serious potential problems. The first is purely computational; estimating a conditional logit model with many parameters (on the order of the number of establishments in one state or even one industry in one state) is very computationally intensive. However, estimating a single set of distance parameters $\hat{\beta}_t$ does not permit splitting the sample to reduce the number of establishment shifters $\hat{\alpha}_{jt}$ that must be estimated at one time. At the time of initial LEHD development over 20 years ago, estimating a conditional logit model with unconstrained establishment shifters was likely impractical to the point of infeasibility. Fortunately, recent advancements in the computation of high dimensional fixed effects have greatly reduced this concern. The likelihood function used for conditional logit regression has an equivalence relation to that of the Poisson regression, and this allows us to estimate both $\hat{\alpha}_{jt}$ and $\hat{\beta}_t$ simultaneously using Poisson pseudo-maximum likelihood with high-dimensional fixed effects using a standard estimator package (Guimaraes, Figueirdo and Woodward 2003, Correia, Guimaraes and Zylkin, Fast Poisson estimation with high-dimensional fixed effects 2020).

The second problem pertains to data availability at the time of imputation. The standard conditional logit is a binary discrete choice model; the outcome variable is simply an indicator of whether a worker was in fact employed at a particular establishment. However, outside of Minnesota, these establishment choices are never actually observed; if they were, we would have no need for imputation. So, any method that imputes establishment assignment needs to either use a pre-defined proxy variable for $\tilde{\alpha}_{jt}$ (as U2W has historically done), or it needs to estimate $\tilde{\alpha}_{jt}$ without a discrete choice outcome variable.

Our solution to the imputation-step problem is to first calculate the percentage of each firm's employment that is at each establishment, and then to use this percentage as the outcome variable in place of the unobservable discrete choice outcome when imputing. Although the use of a non-binary outcome variable violates the assumptions of conditional logit regression, it can readily be estimated using Poisson pseudo-maximum likelihood, and it maintains a congruence between the likelihood function used in estimation and the ones subsequently used in imputation. Moreover, this strategy incorporates the full set of information on employment available in this context, since the outcome variable is in essence the unconditional probability that each worker in the firm will choose any particular establishment.

Throughout the latter sections of this technical note, we will refer to the strategy outlined above as a "full conditional logit" strategy, to be contrasted with the "proxy variable" strategy. To be clear, at the time of estimation, this is in fact a classic McFadden conditional logit framework. It is only at the time of imputation that our "full conditional logit" approach differs, and only in response to the underlying data availability constraint.

Spell-Based Imputation

The existing U2W imputation is a spell-based imputation methodology. That is, a worker's establishment is imputed once per job spell, rather than once per time period, and the imputation for that job spell is then attached to all time periods within the spell. Imputing at the level of the spell ensures that workers' establishment choice is longitudinally consistent; establishment changes within the duration of the job spell are not permitted by construction. It also has the potential to reduce the computational burden of the procedure by limiting the number of imputations per job spell to one. As we will see in Section 3, we have considered a mixture of spell-based and non-spell-based (cross-sectional) imputation methodologies for the job frame establishment impute.

A spell-based imputation method requires at least two non-trivial design decisions to be addressed. The first decision is of which covariates to include in X, and if a proxy variable based on employment is used, how that variable should be constructed. Since workers relocate, commute distance is not constant over the course of a job spell. In our spell-based methodologies, we follow the existing U2W procedure in only using commute distance and employment information from the last quarter of the job spell when imputing probabilities.

The second design decision is in exactly how a job spell is defined, and how the set of candidate establishments $R(it)$ is defined for worker i's job spell ending in time t. Broadly speaking, a job spell is a continuous period of employment for a worker within a firm, and the set of candidate establishments is the set of establishments that have existed for the entirety of that time frame. However, in one of the candidate methodologies described in Section 3, we consider an alternative spell definition that makes use of time variation in place of residence. Further refinements to the definition of a job spell could also be considered in the future.

Although spell-based imputation methods have several advantages, one key disadvantage of them is that we cannot observe the full set of covariate information for job spells that are ongoing (i.e. current job holders as of the last period of available data). In the existing U2W imputation procedure, as new periods of data become available, establishments are re-imputed for ongoing job spells so that they can incorporate the newly available data. In practice, a full history of establishment imputes are re-produced with each LEHD data vintage. This means that results produced from an analysis of one data vintage often cannot be recreated using a later data vintage. It also has implications for data storage, as each data vintage must include the full establishment imputation history of each worker, not just an imputation for the latest time periods.

Although we considered some alternatives to purely cross-sectional and purely spell-based imputation, so far these have been found to perform poorly. All four of the methodologies described in Section 3 are either fully cross-sectional in nature, or spell-based. The other alternatives we considered are described briefly in Appendix C.

3 Candidate Methodologies

A substantial number of model tweaks were examined at various points during this project. Not all of these tweaks made it through the full training and validation exercise, although many did. Many other minor modifications were found to have no meaningful impact on the performance of the imputation and were eventually set aside. In the interest of brevity, we have identified four primary candidate empirical methodologies, each of which highlights some of the fundamental tradeoffs that we are making in choosing an ultimate imputation strategy. Since these empirical methodologies differ on a number of dimensions, they are labeled as methodologies 1 through 4. This section describes the basic features of each methodology, and outlines the key contrasts between them that will be the source of the performance differences shown in Section 5.

With one substantial exception, **Methodology 1** adheres most closely to the strategic decision of the existing Unit-to-Worker (U2W) imputation, as outlined in Abowd et al. (2009). In lieu of a classic conditional logit estimation, it makes use of a proxy variable, a normalized ratio of month 1 establishment log employment from QCEW. The impact of commute distance is estimated using linear splines, with knots at 25, 50, and 100 miles of distance. Moreover, as in the existing U2W imputation, an entirely separate regression is used to train a model on three size classes of multi-establishment firms (less than 100 workers, 100-499 workers, and 500 or more workers).

Where methodology 1 differs meaningfully from the existing U2W methodology is that it is fully cross-sectional in nature, not spell-based. In other words, each worker is given a separate establishment impute in each quarter of their employment, with no longitudinal constraints whatsoever. The candidate establishment set $R(it)$ for each worker also has no longitudinal constraints; it is simply the full set of establishments that report positive employment in month 1 of that quarter.

Methodology 2 is also fully cross-sectional. It incorporates three major changes relative to methodology 1, each of which will also be incorporated into methodologies 3 and 4.

- First, in lieu of training the model with three splines of commute distance, the model is trained with a single distance parameter; the log of commute distance.[5]
- Second, instead of training three regressions on firm size class subsets of the data, this model is trained separately on two-digit NAICS industry sectors of firms.[6]

[5] To avoid dropping observations where the commute distance was zero, we used the log of 1+commute distance, measured in miles. A primary reason for modifying the commute distance parameter is that it ensures that the coefficient is identified off of workers in all firms. In a “full conditional logit” framework such as this one, firms with no long commuters could have no variation in longer distance spline covariates, leading to parameter instability in small industry-sector training samples.

[6] Industry is an establishment-level characteristic in the LEHD infrastructure on which this methodology is tested, but for the conditional logit estimator to work, all candidate establishments within each firm must be included in the same estimation sample. For the purposes of grouping firms by industry, we calculated the employment-weighted all-time modal 2-digit NAICS sector for each firm. Sectors in Manufacturing (31-33), Retail Trade (44-45), and Transportation and Warehousing (48-49) were grouped together, and due to their small size, sectors 21, 22, 51, and 53 were grouped together as other industries. A penalized regression analysis of firm-level commute patterns was unable to identify any additional suitable industry aggregations, nor any other relevant characteristics on which firms might be stratified to improve model performance. However, we hypothesize that

- Finally, and most importantly in terms of the empirical performance of the model, methodology 2 uses a "full conditional logit" specification. That is, the establishment shifter parameters $\hat{\alpha}_{jt}$ are estimated simultaneously to $\hat{\beta}_t$ on the training dataset. Then, at imputation time, $\hat{\alpha}_{jt}$ are again estimated for the establishments in the validation dataset before using those estimates to impute choice probabilities. Conceptually, this change should reduce the potential for attenutation bias in our estimated $\hat{\beta}_t$. More generally, an improved estimation of establishment-level factors should improve our ability to match establishment sizes by capturing amenities or other forces that lead some establishments within a firm to be more popular than others.

Methodology 3 incorporates the same changes as described for methodology 2, except that it is also a spell-based imputation model. A job spell, in this context, is the continuous set of quarters in which a worker is employed by a specific firm.[7] The set of candidate establishments for a spell is the set of the firm's establishments that have existed continuously for at least the entire duration of that spell. Like all spell-based imputations we have implemented, this methodology imputes once per job spell, using the covariate information from only the last quarter in the spell, and then applying the resulting impute to all prior quarters in the spell. Relative to Methodology 2, imposing spell-based restrictions could lead to lower performance in matching establishment sizes. At the same time, such restrictions could lead to improvements in the longitudinal consistency of imputations, as well as reduced computation time. The quantitative importance of these is explored in Section 5.

Methodology 4 is identical to methodology 3, except in how we have defined the spells. Specifically, this methodology defines spells not only based on continuous employment within the firm, but also based on a lack of large residential changes. That is, whenever a person makes a residential move of more than 25 miles, we break their job spell and thus permit the possibility that they may have change establishments corresponding to their move.[8] An establishment is a candidate for imputation in a residentially-defined spell if it has existed for the full time of that spell, even if it did not exist during a previous quarter of the worker's continuous employment with the firm.

Since methodologies 3 and 4 are very similar, it will not be surpising that the differences in performance between them are small. The key reason for the distinction between these two methodologies is to consider whether additional information on residence may improve our ability to match aggregate establishment separation rates.

4 Measures Evaluated

Given the complexity of this imputation framework, there are many potential ways in which one could evaluate the quality of the model fit. An ideal model specification successfully matches individuals to

training the model separately by industry may allow this methodology to better accommodate differential industry trends in commuting patterns driven by the unequal rise in remote work.

[7] Because of occasional establishment reporting issues in the administrative data, we make an allowance in defining job spells for one-quarter periods in which the worker's reported establishment is unreliable due to a known reporting issue.

[8] The Residence Candidate File from which place of residence is obtained is an annual file. Whenever we observed a large residential move between years t and $t + 1$, we smooth aggregate establishment separation behavior over the quarters of the year by choosing at random a quarter between Q3 of year t and Q2 of year $t + 1$ in which to break the spell and therefore allow for a potential establishment change. See Appendix D for more details.

their true establishment a very high percentage of the time. If an individual match rate near 100% were possible, then there would be little need for other evaluation metrics. However, two specifications with identical individual match rates may differ dramatically in how well they capture establishment sizes, changes in establishment employment over time, separation rates, geographic patterns, or other phenomena of interest.[9]

This section outlines the metrics that we use, which are the presented for several candidate models in the tables below.

(Probabilistic) individual successful match rate:

Per the discrete choice framework, each worker is assigned a probability of employment in each candidate establishment in their choice set $R(it)$. These probabilities are used to construct a cumulative density function, and an establishment is imputed by taking a uniform random draw against the CDF. These individual draws can be compared against the true establishment $J(i, t)$, but to minimize issues related to small samples, we instead use the average of the predicted probabilities directly. That is, our primary individual match rate statistic is:

$$P_t = \frac{1}{N_t} \sum_i \sum_{j \in R(it)} p_{ijt} \cdot \mathbb{I}(j = J(i, t))$$

We use similar statistics for evaluating performance on employer characteristics. For example, the probability that we impute an individual to the correct industry (NAICS sector or subsector), or to the correct geographic unit is measured similarly to the above.

Establishment-level size performance:

Another important measure of our imputation model's performance is how well it matches establishment sizes. An imputation method may match establishment sizes very well even if its individual match rate is relatively low.

Our primary measure of establishment-level size performance is the mean squared difference between imputed establishment employment measures and the true ones. These differences can be considered model errors, so we call this measure Mean Squared Error (MSE), defined as:

$$MSE_{jt} = \sum_j \frac{p_{jt}}{\sum_j p_{jt}} \left(\hat{p}_{jt} - p_{jt}\right)^2$$

Where p_{jt} is the true employment size, $\hat{p}_{jt}$ is our imputed prediction, and $\frac{p_{jt}}{\sum_j p_{jt}}$ is the establishment employment weight.[10] We calculate an MSE of the full set of firms. In Appendix A, we also perform this analysis for the subset of firms with the highest levels of structural and employment change.

[9] For a simple thought exercise, consider a simple imputation that imputes teachers to employment in one of two nearby, equally-sized schools. A model that places each teacher in exactly the wrong school will have a 0% individual match rate, but it will do a much better job of replicating the size of the schools than a model that places all teachers in one school (50% match rate).

[10] $\hat{p}_{jt}$ can be computed by adding up all individual predicted probabilities for an establishment, or by adding up the workers imputed to each establishment. Both approaches yield similar results, discussed below. Alternatively, the

MSEs are straightforward summary statistics, but they are limited in the information they can convey. For example, a high MSE might reflect widespread errors of small magnitude, or alternatively it might capture a small number of very large errors. MSEs are also difficult to interpret in economic terms. An alternative approach is to look at the entire distribution of differences between the imputed establishment sizes and the true ones (the model "errors", defined as $(\hat{p}_{jt} - p_{jt})$). In this approach we look at the magnitudes of the errors in the model at different points in the error distribution. While there is no single summary statistic like the MSE, looking at the full distribution yields important insights which will be discussed in the results section.

Longitudinal Consistency

A third measure of performance focuses on how well each model matches establishment level separation rates. We do this for two reasons. First, separation rates constitute an important downstream data product (QWI). Second, accurate separation rates would imply longitudinal consistency in job spells subject to imputation (as such, this margin would be of great importance for data users interested in tracking workers in jobs over time).

We calculate each establishment separation rate $Seprate_{jt}$ as the fraction of workers in establishment j who were present in such establishment at time t, but are not present in the establishment j in the subsequent quarter (t+1).[11] As before, this measure can be calculated both using workers' true employment spells and spells based on our models predictions.

Computation Time

Our final empirical consideration is computation time. As described in Section 2, computation time is a substantial problem for conditional logit models that estimate many parameters. Since the chosen imputation method will ultimately need to be applied to a national-level dataset on an annual basis, choosing a method that is practicable is important.

We have tracked the time needed used to train each model on one half of the quarterly Minnesota LEHD data for 2015, and also to impute establishments on the other half of the data. Since our training and validation analysis has been performed in Stata, these times reflect the computation times for the specific procedures used to train the models and impute establishments only. They do not include the time needed to build the datasets or identify the candidate establishment sets.[12] Although the precise computation time may differ when our chosen specification is ultimately implemented, we believe that

MSE can be computed using establishment employment shares (within a firm) instead of employment levels. This approach yields similar results as well.

[11] Specifically, we use the formula:

$$Seprate_{jt} = \frac{\sum_j \mathbb{I}\big(j = J(i,t)\big) * \mathbb{I}(j \neq J(i,t+1))}{\sum_j \mathbb{I}\big(j = J(i,t)\big)}$$

Where $\mathbb{I}\big(j = J(i,t)\big)$ is an indicator for worker i being employed in establishment j at time t. Note that this measure captures workers moving to: different firms, different establishments within the same firm, and workers moving to non-employment. For tractability purposes, we exclude workers with multiple jobs in a given quarter (1.5%) of observations. We also only compute this measure for the first three quarters of 2015 (since this measure requires a follow-up quarter in the computation).

[12] The time required to define establishment candidate sets is a function of the spell definition used, but it does not otherwise vary with the estimation method.

this information gives a rough sense of the differences in computational demands of the four methodologies. Time to impute establishments is a more important measure than time to train the model, since the ultimate imputation sample will be much larger than the training sample. However, we report computation time results for both steps of the process.

5 Results

Table 1 shows the average of the predicted probabilities for each model. Column 1 shows the mean probability for the correct establishment, while columns 2-4 show the mean probabilities for the correct industry (NAICS 2, 4 and 6) or to the correct county or tract (columns 5 through 6). A comparison across rows shows that all models perform similarly across all measures. Column 1 shows that individual match rates for all models are all within a narrow 0.2 percentage point interval (49.3%-49.5%). As the level of aggregation increases (e.g. going from 6 to 2 digit NAICS), match rates increase mechanically but they do so uniformly across models.

Appendix Figure 1 provides some context for why we see such a narrow range of performance in terms of individual match. This set of kernel density plots show the distribution of probabilities that each methodology assigns to the true establishment where each worker was employed. The methodologies perform similarly in terms of individual match because on average they assign similar probabilities to the true establishment. Moreover, the distribution of assigned probabilities is highly bimodal. This means that roughly half of workers will be imputed to their true establishments nearly all of the time, while most of the remaining workers have a very low probability of being imputed correctly, regardless of the methodology used. These latter workers may be, for example, workers in firms with many establishments that have a similar size and commute distance, or workers who commute to a farther establishment when a much closer establishments exist.

Appendix Table 1 provides evidence on the subject by showing firm and commute distance statistics for records on different parts of the distribution of Appendix Figure 1. Records are split into terciles based on the Methodology 1 assigned probabilities to the true establishment.[13] Tercile 1 includes cases with the lowest assigned probabilities (i.e. worse performance) ranging from 0 to 18.4%, while Tercile 3 includes the right tail of Appendix Figure 1 (probabilities ranging from 80.5 to 100%, which correspond to good individual match performance). Column 2 shows the number of candidate establishments is negatively correlated with match performance. Records in Tercile 1 have on average 63 candidate establishments, while those in Tercile 3 have only 11.8 candidate establishments on average. Column 3 shows the mean firm size across quartiles, but in this case there is no clear correlation with match performance: imputations with the best performance (Tercile 3) involve firms that are smaller than in all other quartiles, but the second best performing group (Tercile 2) features the largest firms. Column 4 shows mean commuting distance (in miles), while column 5 shows commuting distance relative to the mean distance to all candidate establishments.[14] In both cases, it does not appear that distance plays a major role in explaining differences in assigned probabilities to the true establishment. The evidence in this table suggests that large number of candidate establishments can explain why some records are

[13] Results based on the other three methodologies are very similar given the similar distribution of assigned probabilities shown in Appendix Figure 1.

[14] Negative values of this measure indicate that workers are employed in establishments that are closer in distance than the average establishment within their firm.

difficult to impute to the correct establishment. Without additional sources of information, methodological choices that only tweak the parameters of the model are unlikely to improve individual match rates substantially.

Table 2 shows the MSEs for establishment size for each model. Each column shows the MSEs based on different imputes. Column 1 MSE is based on employment sizes calculated by adding up all imputed workers, column (2) is based on employment sizes calculated by adding up all predicted probabilities for each establishment. Column 3 shows MSEs based on predicted establishment employment shares (within a firm). All MSE measures are computed based on squared differences making a direct interpretation of their magnitudes difficult.[15] For our purposes, what matters is their relative performance. Comparing the different models in each row, we can see that Methodology 1 showcases the best performance, with MSEs that are about 1.5% to 10% lower than the other models. In Appendix A, we also show two tables of results for the subgroups of firms with either the largest structural changes over time, or the largest employment changes over time, with mixed performance across the different methodologies.

In Table 3 we move beyond MSEs and look at the entire distribution of differences between the imputed establishment sizes and the true ones (the model "errors", defined as $(\hat{p}_{jt} - p_{jt})$). Positive errors indicate the model is over-imputing workers to a particular establishment. Negative errors indicate the model is assigning less workers to an establishment than it should. Positive and negative errors would cancel each other if we were to take their mean, hence the need to look at the entire distribution. In table 3, we start by looking at the errors as a percentage of the establishment employment. Column 3 shows that the median error is very small (the discrepancy is below 0.01% in all models). However, there is significant variation across the distribution. Methodology 2 has the best performance in this case, with a 25th percentile error of -4.6% and over imputing by 5.9% at the 75th percentile. This range is significantly smaller than in other models. For example, at the 25th percentile Methodology 3 has a 17.1% undercount, while at the 75th percentile the overcount is 7.9%. The problem is larger at the tails of the distribution. Looking at column 1, the results imply for the worst undercounts (at the 10th percentile) Methodology 2 assigns 18.3% fewer workers to an establishment, compared to 49.1% fewer workers in Model's 3 assignment. Some of this large percentages might arise from errors in small establishments with under/over assignments that are small in absolute terms.

To complement this analysis, we show an analogous set of results in Table 4 using levels of the errors. As such, Table 4 results can be directly interpreted as the number of workers that each of our models is over/under assigning to each establishment. Using this method will give more weight to large establishments (since they are more likely to experience larger errors, even if they have the same error rate as smaller establishments). Reassuringly, we see that in absolute terms all models perform similarly, with interquartile error ranges always within -2 to 2. This means that in all our models, most establishment sizes are off by less than 2 workers. The large percentage errors in Table 3 are likely due to small establishments. The median establishment has 10 workers, so a being off by 1 worker amounts to a 10% error. Looking at the tails of the distribution, we find larger differences across models. Methodology 2 has the best performance at the 10th and 90th percentile with an undercount of 3.6

[15] Another implication of using squared differences is that these measures are heavily influenced by outliers. This feature is useful in cases where large differences are not desirable, as in our case.

workers, and an overcount of 3.7. This compares to under/over counts of 7.2 and 5.4 workers for Methodology 3.

Table 5 compares separation rates across our 4 models relative to the true separation rate. Column 1 shows the average separation rate (across establishments), while column 2 shows the separation rate for the median establishment. The first row presents the true values, indicating that the average separation rate in our sample is 15.9%, and that the median establishment has a separation rate of 9.4%. Models 1 and 2, which are not spell-based, perform badly along this margin. These models' imputation result in mean separation rates of 68.6% and 66.7%, respectively, vastly overestimating the actual separation rate. Their performance for the median establishment is worse, as these models imply a median separation rate of 82.1% and 76.9% (vs. a true median separation rate of 9.3%). The overestimation of the separation rates is not surprising, given that these models do not impose any type of longitudinal consistency in terms of spells. Even if these models were to assign workers to the correct establishments most of the time, workers with long spells are bound to have draws in which they are assigned to the incorrect establishment. This would significantly inflate the separation rates of all establishments in the candidate set.[16] In contrast, Models 3 and 4, which do impose spell-based restrictions, perform fairly well. Their imputations imply separations rates of 14.1% and 14.7%, which are much closer to the true separation rate of 15.9%. This improved performance is expected, given that these models explicitly impose longitudinal consistency in the spells.[17]

Lastly, Table 6 reports the computation time needed to train each model methodology on the training dataset, and then to perform one impute on the validation dataset. There is, unsurprisingly, wide variation in the computational requirements of these methodologies. Methodology 2 is consistently the slowest method, because in both training and imputation it requires that establishment shifter parameters be estimated for every establishment-quarter. It takes about four times as long to impute as Methodology 1 takes by using a proxy variable for $\tilde{\alpha}_{jt}$. However, the performance penalty from using a full conditional logit estimator is much smaller among the spell-based methodologies. This is because a spell-based framework makes a much smaller number of imputes, and because imputation can be performed industry-by-industry (or even firm by firm) to limit the number of establishment shifter parameters $\tilde{\alpha}_{jt}$ that must be estimated at one time. Methodologies 3 and 4 take about 40% longer to impute establishment for all workers than Methodology 1.

To sum up, our measures show that all models perform similarly in terms of individual match rates (assigning workers to the correct establishments). In terms of matching establishment sizes, Methodology 2 shows a better performance. However, when it comes to matching separation rates, the spell-based models (Models 3 and 4) perform better. This pattern follows from each the model's features. All models target matching the true establishment size distribution. Models 1 and 2 do so without any constraints (by not imposing longitudinal consistency). This means they perform better

[16] For a concrete example, take a worker with an employment spell of 10 quarters in the same establishment in a two-establishment firm. Even if Methodology 1 were to give this worker a 90% imputation probability to the correct establishment, such worker will on average be assigned to the other establishment at least once (over the 10 quarters). This would artificially inflate the separation rate of correct establishment. Moreover, this would also inflate the separation rate of incorrect establishment when the worker is imputed back into the correct establishment.

[17] The slight underestimation of Models 3 and 4 separation rate is due to the fact that they impose no mobility of workers across establishments within a firm.

when it comes to establishment sizes, but the tradeoff is worse performance in terms of separation rates. Models 3 and 4 impose longitudinal consistency. This means they perform better along this margin (separation rates), but the tradeoff is lower performance in establishment size accuracy.

The relative weight each performance measure should be given is subjective and will depend on the end-use of the data. For products which rely heavily on accurate establishment sizes, non-spell-based models feature better performance (in particular when it comes to outliers). For products which rely on accurate worker employment spells and accurate job transitions, spell-based models have a clear advantage.

Bibliography

Abowd, John M., Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Woodcock Simon. 2009. "The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators." In *Producer dynamics: New evidence from micro data*, 149-230. University of Chicago Press.

Correia, Sergio, Paulo Guimaraes, and Thomas Zylkin. 2020. "Fast Poisson estimation with high-dimensional fixed effects." *The Stata Journal* 20 (1): 95-115.

—. 2019. "Verifying the existence of maximum likelihood estimates for generalized linear models."

Green, Andrew, Mark J. Kutzbach, and Lars Vilhuber. 2017. "Two perspectives on commuting: A comparison of home to work flows across job-linked survey and administrative files."

Guimaraes, Paulo, Octavio Figueirdo, and Douglas Woodward. 2003. "A tractable approach to the firm location decision problem." *Review of Economics and Statistics* 85 (1): 201-204.

McFadden, Daniel. 1973. "Conditional logit analysis of qualitative choice behavior." *Frontiers in econometrics* 105-142.

Mood, Carina. 2010. "Logistic regression: Why we cannot do what we think we can do, and what we can do about it." *European sociological review* 26 (1): 67-82.

Tables

Table 1: Individual Match Rates

	Individual	County	Tract	NAICS 4d	NAICS 2d
Methodology 1	49.3%	77.1%	61.3%	82.4%	89.6%
Methodology 2	49.3%	76.5%	61.1%	82.5%	89.6%
Methodology 3	49.5%	76.3%	61.1%	83.1%	90.0%
Methodology 4	49.3%	76.1%	60.9%	83.0%	90.0%

Table 2: Establishment Size Performance (MSEs)

	Direct Imputation	Probabilistic	Shares
Methodology 1	67,760	64,160	0.0035
Methodology 2	71,920	72,900	0.0031
Methodology 3	74,550	68,920	0.0050
Methodology 4	68,110	67,420	0.0049

Table 3: Establishment Size Error Distribution (in % of true size)

Percentile	10th	25th	50th	75th	90th
Methodology 1	-33	-12.1	0	8.3	22.9
Methodology 2	-18.3	-4.6	0	5.9	15.1
Methodology 3	-49.1	-17.1	0	7.9	21.6
Methodology 4	-46.9	-16.4	0	7.8	21

Table 4: Establishment Size Error Distribution (in levels)

Percentile	10th	25th	50th	75th	90th
Methodology 1	-5.26	-1.27	0	1.26	5.42
Methodology 2	-3.59	-0.64	0	0.88	3.67
Methodology 3	-7.24	-1.67	0	1.18	5.4
Methodology 4	-7.02	-1.63	0	1.15	5.29

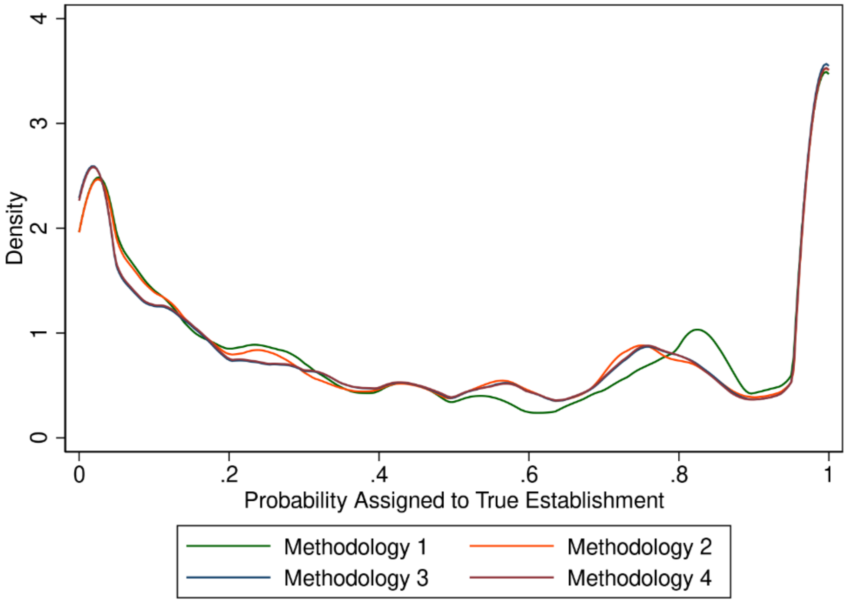
Table 5: Separation Rates

	Mean	Median
True	15.9%	9.3%
Methodology 1	68.6%	82.1%
Methodology 2	66.7%	76.9%
Methodology 3	14.1%	8.5%
Methodology 4	14.7%	9.3%

Table 6: Computation Time in Seconds of the Training and Imputation Procedures

	Training	Imputation
Methodology 1	966	992
Methodology 2	2573	4112
Methodology 3	229	1367
Methodology 4	220	1329

Appendix Tables and Figures



Appendix Figure 1: Kernel Densities of Probability Assigned to True Establishment

Appendix Table 1: Firm, Establishment and Commuting distance by Tercile of Probability Assigned to True Establishment (Methodology 1)

	Probability Range	Mean number of candidate Establishments	Mean SEIN size	Mean commuting distance	Mean commuting distance relative to average of all candidates
Tercile 1	0-0.181	63.1	8922	48.05	-14.66
Tercile 2	0.181-0.805	40.3	9298	45.89	-33.05
Tercile 3	0.805-1	11.8	7051	40.55	-31.95

Appendix A: Results on MSEs for firms with large employment or structural changes

The MSE results shown in Table 2 are constructed at the establishment level for the entire set of firms in the validation dataset. However, it may also be of interest to consider how these methodologies perform specifically when applied to firms that are changing quickly over time.

In Appendix Table A1, we show results for the subgroup of firms with the largest structural changes over time. To measure structural change, we start by computing a measure of firm-level structure: the Hirschman-Herfindahl index (HHI) in within-firm establishment employment shares ($HHI_{ft} = \sum_j p_{jt}{}^2$. Because this index is constructed in employment shares, its range is $[0,1]$. A large change in this measure over time ($HHI_{ft} - HHI_{f,t-1}$) represents big structural changes in a firm's establishment composition. We identify the firms whose absolute change in HHI exceeds the 90th percentile for our dataset. Then, we compare the MSEs for this subgroup of firms. Additionally, in Column 4 we evaluate each model's capacity to handle these large structural adjustments by looking at changes in HHI based on the model-imputed shares and the true changes in HHI. Specifically, our measure is $DD_{HHI_{ft}} = \left(\widehat{HHI}_{ft} - \widehat{HHI}_{f,t-1}\right) - \left(HHI_{ft} - HHI_{f,t-1}\right)$. High values of this measure indicate our models overstate firm's true structural changes, low (negative) values would suggest our models understate structural changes. The results are largely consistent with those in Table 2, with Methodology 1 having the best performance for most of the measures.

In Appendix Table A2, we repeats the analysis, but focusing on firms with the largest absolute change in employment. Here, we have included only firms whose absolute symmetric difference rate of firm growth exceeds the 90th percentile among firms in our dataset. The results are mixed in this case, with Methodology 3 and 4 performing better with three of the measures (and Methodology 2 performing best with one MSE measure).

Appendix Table A1: Establishment Size Performance (MSEs) for Firms with High Structural Change

	MSE			
	Direct Imputation	Probabilistic	Shares	$DD_{HHI_{ft}}$
Methodology 1	30,140	30,780	0.0057	0.0002
Methodology 2	35,510	31,390	0.0052	0.005
Methodology 3	55,140	55,090	0.0058	0.0014
Methodology 4	53,320	53,730	0.0057	0.0052

Note: Sample of firms at the 90th percentile of absolute structural change (as measured by the change over time in true HHI ($HHI_{ft} - HHI_{f,t-1}$). 90th percentile corresponds to a change in HHI of at least 0.025, with the median for the group being 0.05). $DD_{HHI_{ft}} = \left(\widehat{HHI}_{ft} - \widehat{HHI}_{f,t-1}\right) - \left(HHI_{ft} - HHI_{f,t-1}\right)$ measures model performance in terms of capturing structural changes in a firm. High values of this measure indicate our models overstate firm's true structural changes, low (negative) values would suggest our models understate structural changes. $DD_{HHI_{ft}}$ is computed at the firm level, MSEs are computed at the establishment level (for comparison purposes with Table 2).

Appendix Table A2: Establishment Size Performance (MSEs) for Firms with High Absolute Employment Change

	MSE			
	Direct Imputation	Probabilistic	Shares	$DD_{HHI_{ft}}$
Methodology 1	72,030	67,340	0.0028	0.0075
Methodology 2	83,760	67,460	0.0026	-0.0077
Methodology 3	60,450	55,340	0.003	-0.0045
Methodology 4	54,670	54,120	0.0029	-0.0066

Note: Sample of firms at the 90th percentile of absolute employment chan. 90th percentile corresponds to an absolute employment change of at least 23%. $DD_{HHI_{ft}} = \left(\widehat{HHI}_{ft} - \widehat{HHI}_{f,t-1}\right) - \left(HHI_{ft} - HHI_{f,t-1}\right)$ measures model performance in terms of capturing structural changes in a firm. High values of this measure indicate our models overstate firm's true structural changes, low (negative) values would suggest our models understate structural changes. $DD_{HHI_{ft}}$ is computed at the firm level, MSEs are computed at the establishment level (for comparison purposes with Table 2).

Appendix B: Attenuation Bias from Conditional Logit Using non-estimated $\tilde{\alpha}$

The existing U2W imputation procedure uses a normalized ratio of log establishment employment as the proxy variable $\tilde{\alpha}_{jt}$. The normalization turns out to be irrelevant (it is canceled out as part of the ratio built into the probability statement), but this appendix demonstrates the basic issue. Subscripts for choice sets r have been omitted for clarity.

$$\tilde{\alpha}_{jt} = \log(n_{jt}) = \log\left(\sum_i p_{ijt}\right)$$

Plugging in the probability statement yields:

$$\tilde{\alpha}_{jt} = log\left(\sum_i \frac{exp\{\alpha_{jt} + x'_{ijt}\beta_t\}}{\sum_k exp\{\alpha_{kt} + x'_{ikt}\beta_t\}}\right)$$

$$\tilde{\alpha}_{jt} = log\left(e^{\alpha_{jt}} \sum_i \frac{exp\{x'_{ijt}\beta_t\}}{\sum_k exp\{\alpha_{kt} + x'_{ikt}\beta_t\}}\right)$$

And so, the exponentiated gap between the true establishment shifter parameter and our proxy is:

$$e^{\tilde{\alpha}_{jt}-\alpha_j} = \sum_i \frac{exp\{x'_{ijt}\beta_t\}}{\sum_k exp\{\alpha_{k_t} + x'_{ikt}\beta_t\}}$$

A sufficient condition for attenuation bias to exist is that the gap in establishment shifter is relevant, which is true if the covariance between this gap and each individual's outcome of interest is non-zero. This will be true any time that there is variation in commuting distance, even if all workers' locations and choices are independent, because:

$$Cov\left(e^{\tilde{\alpha}_{jt}-\alpha_j}, p_{ijt}\right) =$$

$$Cov\left(\frac{exp\{x'_{ijt}\beta_t\}}{\sum_k exp\{\alpha_{k_t} + x'_{ikt}\beta_t\}}, \exp\{\alpha_{jt}\}\frac{exp\{x'_{ijt}\beta_t\}}{\sum_k exp\{\alpha_{k_t} + x'_{ikt}\beta_t\}}\right) + Cov\left(\sum_{-i}\frac{exp\{x'_{ijt}\beta_t\}}{\sum_k exp\{\alpha_{k_t} + x'_{ikt}\beta_t\}}, p_{ijt}\right)$$

By independence, the latter term is zero, but the former term still yields:

$$Cov\left(e^{\tilde{\alpha}_{jt}-\alpha_j}, p_{ijt}\right) = \mathrm{e}^{\alpha_{jt}} \cdot Var\left(\frac{exp\{x'_{ijt}\beta_t\}}{\sum_k exp\{\alpha_{k_t} + x'_{ikt}\beta_t\}}\right) > 0$$

Appendix C: Other Model Concepts Evaluated

Part of the purpose of this technical note is to guide future researchers who might be interested in updating or revising such an establishment imputation in the future. Given the many ways in which an imputation model can be tweaked, it may be helpful to know some of the other concepts we tried that either had little impact, or that simply didn't work as expected.

Censoring commute distances

One concern we had was that our model parameters on commute distance could be heavily influenced by the existence of either extremely long or extremely short commutes. Commutes beyond a certain threshold are likely to be infeasibly long, and so any distance variation within them might be immaterial. Similarly, once a commute is below a certain short distance threshold, differences in distance might cease to matter, or might be dwarfed by unobservables such as the specific details of road layouts, traffic patterns, or transit lines. We tried a variation on our model in which residential distances were bottom-coded at 5 miles and top-coded at either 200 or 300 miles. These modifications had minimal impact on model coefficients or performance.

Forward iteration

As described in Section 2, spell-based methods require ongoing establishments to be re-imputed for ongoing job spells as new information becomes available, while cross-sectional methods may perform poorly in terms of longitudinal consistency of workers' establishments. One alternative we considered was to iterate forward through time periods, using the probability distribution formed from model training in one time period as an "expected probability" that would inform estimation and imputation in the subsequent time period. That is, we fit a discrete choice model in each time period corresponding to the probability statement:

$$p_{ijrt} = \frac{e^{\alpha_{jt}+\beta_t X_{ijrt}+\gamma_t p_{ijr,t-1}}}{\sum_{k\in R(it)} e^{\alpha_{kt}+\beta_t X_{ikrt}+\gamma_t p_{ikr,t-1}}}$$

Since the prior probability distribution in the first estimation period is always unknown (defined as 0), this exercise was performed over a multi-year period. Variations on this strategy also incorporated interaction terms that might better handle the cases of new establishments and new jobs where there is no meaningful prior on the probability distribution. Overall, the structure is vaguely analogous to the structure of a linear dynamic panel regression, where the prior period's outcome variable is used as a covariate in regressions.

Our expectation was that the coefficients γ_t would be strongly positive and stable over time, reflecting a positive correlation between the conditional probabilities one would form by looking only cross-sectionally. However, the lagged probability coefficients γ_t turned out to be unstable over time, and were often small or even negative in some periods. We suspect that the non-linear nature of the underlying discrete choice model may contribute to its instability when lagged probabilities are included. However, given the time constraints of producing an imputation for the new job frame, we have dropped this line of model methodologies from consideration.

Using employment change after the last period of the spell:

One idea we had was that spell-based imputation might perform better if it incorporates information on what happens immediately after the spell is concluded. For example, workers whose last quarter at a firm is also the quarter before an establishment shrinks or disappears could be much more likely to have been employed by that establishment for their entire job spell. While this idea is conceptually attractive, it presents challenges when estimating a full conditional logit equation. The flexible establishment-time shifters $\hat{\alpha}_{jt}$ already account for all establishment-quarter level variation, including variation in establishments' future sizes. We experimented with model specifications that included only establishment-year shifters, and then included covariates for future establishment level employment changes. However, we found that these models did not perform any better than the standard spell-based conditional logit with quarterly establishment shifters.

Appendix D: Additional Sample Construction Details

The construction of samples for establishment imputation requires several design decisions which are not immediately obvious. This appendix provides an overview of some of the key design decisions in constructing the training and validation datasets for this exercise. In some cases, these methods match those of the existing U2W imputation exactly, while in other cases there are substantial differences. Where possible, a comparison is provided to existing methods used in U2W.

Removed Observations Due to Establishment Reporting Issues

Although Minnesota UI data report work establishments in general, these data do not always appear to be accurately reported. For example, from QCEW-reported establishment employment, it is sometimes evident that a firm's UI reporting assigns all employees into one establishment even though their ES-202 reporting indicates workers spread across establishments. Sometimes these issues arise in only one quarter, while in other cases it appears that a firm's UI establishment reporting is never reliable. We exclude firm-quarters from both the training and validation datasets where we believe that establishment reporting issues are likely to be present, so that these observations cannot impact our estimated model parameters or selection of methodology.

Our method for identifying and excluding firm-quarters with establishment reporting issues makes use of the same criteria for identifying these issues as the current U2W methodology. That is, we construct the HHI (sum of squared shares) of reported establishment employment from UI records and also from QCEW month 1 employment. In cases where UI reporting is substantially structurally different (more or less concentrated in establishments) than QCEW employment, we flag the entire firm's reported employment as potentially bad for that quarter.

Because a single quarter of bad establishment reporting by a firm could lead all worker spells to be broken, we keep worker observations where only a single quarter of reporting is bad, and we assume that their reported establishment is the same as it was in the preceding quarter.

Employment in Multiple Establishments in the Same Quarter

There is no requirement in Minnesota that an employee be reported in only one establishment within a quarter. However, any establishment imputation will impute workers to only a single establishment in each period. So, when constructing the training and validation datasets, workers are automatically assigned to the establishment in each quarter at which the largest portion of their earnings are reported.

Defining Job Spells for Methodologies 3 and 4

In general, a job spell is a series of contiguous quarters of employment (positive earnings) for a worker at a firm. The procedure to identify job spells and perform the spell-based imputation is as follows:

- The input dataset (the EHF_MN file) is sorted by person (PIK) and firm (SEIN), and then the necessary information for the person's entire job history at that firm is read into a series of arrays, so that the remaining processing can occur by iterating through quarters of the data once all the job history information has been read in.
- Three variables are constructed: a variable (*min_spell_qtime*) that will report the first quarter of each job spell, a firm-level job tenure variable (*tenure_sein*), and a binary separation flag variable (*last_sep*) that indicates that a quarter is the last quarter in the spell. In the first quarter of observed positive wages (beginning of the job spell), *min_spell_qtime* is set to the value of the current quarter, and *tenure_sein* is set to 1. The value of *min_spell_qtime* is retained as the algorithm iterates through time, and the value of *tenure_sein* increments for as long as positive earnings are observed at the firm. If earnings are zero in any quarter, then this causes the tenure clock to reset, which leads *min_spell_qtime* to be updated the next time that positive earnings are observed.
- In each quarter q, the algorithm also looks ahead in the array one quarter to quarter $q+1$. If no earnings are reported in $q+1$, then the spell separation flag *last_sep* is set to 1.
- Spell-based methodologies 3 and 4 operate by only training and imputing on observations for which either *last_sep*=1*,* or which occur in the last quarter of the training and validation datasets. After imputation has taken place, the generated probabilities and imputations are applied to all prior quarters of the same spell by looking for observations with the same value of *min_spell_qtime*.

Defining Candidate Establishment Sets for Methodologies 3 and 4

The procedure for defining candidate establishment sets for spell-based imputation is conceptually similar to the procedure for defining job spells above.

- As above, the procedure loads all observations from the LEHD's ECF_MN_SEINUNIT file for each establishment before processing, and it then iterates through time as it outputs information on each establishment-quarter.
- In place of a variable that reports the first quarter of positive earnings, a variable is constructed (*min_unit_qtime_spell*) that reports the first quarter in which the establishment is observed in the ECF. The value of this variable is retained for as long as the establishment continues to be reported contiguously. If the establishment is not observed for a quarter, then in the next quarter of observation, the value of *min_unit_qtime_spell* is replaced with the current quarter.
- This file of establishment-quarter level information is eventually merged to the job-establishment-quarter level dataset that is then split into training and validation datasets. For any job spell, the set of candidate establishments can be identified from the last quarter of the spell as any establishment for which *min_unit_qtime_spell <= min_spell_qtime*.

This method of identifying eligible candidates is generally more restrictive than the approach used by the production U2W, which identifies eligible candidates based on the earliest appearance of the establishment irrespective of gaps. In addition, allowances are made for some establishments that have predecessors or SEINs that change from single unit reporting on the QCEW to multiple units. Also, job spells are broken when significant changes to firm structure are found. The rule used to identify these breaks is that (roughly) 15% of SEIN employment is at establishments that are born or die in a quarter.

Spell Breaks Due to Residential Relocation in Methodology 4

In Methodology 4, we modify the spell-based imputation methods of Methodology 3 to account for large residential changes. We do this through the following procedure, which occurs simultaneously to the general spell-defining procedure outlined above. As with previously described procedure, a variable is constructed that reports the first quarter of each job spell, and a binary variable is constructed that indicates the last quarter of a job spell. As sample construction proceeds through time for each job, the value of the initial spell quarter is retained unless the spell is considered to have ended, at which point it is set to the first quarter of the new spell. In addition:

- To each job-quarter record in year t, we attach the place of residence information for years t, $t-1$, and $t+1$. Place of residence is obtained from the LEHD's ICF_MN_addresses file. Place of residence in year t is required to calculate establishment commute distances, and so workers with no known place of residence in t are already excluded from both the training and validation sets.
- Distance in miles is calculated between places of residence in years $t-1$ and t, and between years t and $t+1$. If the place of residence in $t-1$ or $t+1$ is unknown, then the distance is set to 0 (i.e. it is assumed that the worker did not move unless we observe otherwise).
- In cases where the worker makes a residential move of 25 miles or more, the spell is automatically broken. Subsequent earnings are treated as a new job spell for establishment imputation purposes, even if earnings have been reported continuously.
- Since place of residence is only known on an annual basis, we choose at random a quarter between quarter 3 of year t and quarter 2 of year $t+1$ to be the last quarter of the spell whenever a residential move requires that a spell be broken between t and $t+1$.

Handling of Imputation Edge Cases

- In a few cases, the standard Stata prediction procedures that we use to construct predicted probabilities yield predicted probabilities that do not sum up to 1 across all candidate establishments in the candidate set for a job. In these cases, the predicted probabilities for the job's candidate establishments are rescaled to produce a valid CDF for imputation.
- In spell-based methodologies 3 and 4, there are occasionally spells for which no eligible establishments exist, because no establishments have been in existence for the entire job spell. In these cases, all establishments that exist in the last quarter are used as candidates, and each candidate is assigned the unconditional predicted probability (the percentage of the firm's month 1 employment that is in that establishment).
- The Poisson pseudo-maximum likelihood estimator occasionally drops observations to avoid collinearity issues that can lead fixed effects not to be identifiable. This can lead to jobs for which no predicted probabilities are available, because all candidate establishment observations for the job have been dropped. In these cases, each establishment that exists in the imputation quarter is assigned the unconditional predicted probability (the percentage of the firm's month 1 employment that is in that establishment).

Additional References on the Existing Unit-to-Worker Imputation

In addition to Abowd et al. (2009), several internal documents and analyses have examined the performance of the existing U2W imputation methodology, and have described its handling of data reporting issues and other edge cases. These include:

- “QWI update: Improvements to multi-establishment imputation” (2012).
- “Correcting for Unreliable Reporting of Establishment on Wage Data” (2015).

Although these analyses are not publicly available, they may be made available to internal Census Bureau users upon request.