A primary objective of the QWI is to provide employment, job and worker flow, and wage measures at a very fine level of geographic (place-of-work) and industry detail. The structure of the administrative data received by LEHD from state partners, however, poses a challenge to achieving this goal. QWI measures are primarily based on the processing of UI wage records which report, with the exception of Minnesota, only the employing firm (SEIN) of workers. The QCEW micro-data, however, are comprised of establishment-level records which provide the level of geographic and industry detail needed to produce the QWI. For firms operating only one establishment, the attachment of establishment-level characteristics is trivial. However, approximately 30 to 40 percent of state-level employment is concentrated in firms that operate more than one establishment. For these multi-establishment firms, the SEIN on workers' wage records identifies the employing firm in the QCEW data, though, not the employing establishment.

In order to attach establishment-level characteristics to workers of multi-establishment firms, a probability model for employment location and imputation was developed. The model explains establishment-of-employment using two key characteristics available in the LEHD data: 1) distance between place-of-work and place-of-residence and 2) the distribution of employment across establishments of multi-establishment firms. The model is estimated using data from Minnesota, where both the firm (SEIN) and establishment identifiers appear on a worker's UI wage record. Then, parameters from this estimation are used to multiply impute establishment-of-employment for workers in the data from other states. Emerging from this process is an output file, called the unit-to-worker (U2W), containing ten imputed establishments for each worker of a multi-establishment firm. These implicates are then used in the downstream processing of the QWI.

The U2W process relies on information from each of the four infrastructure files – ECF, GAL, EHF, and ICF – as well as the auxiliary SPF file. Within the ECF, the universe of multi-establishment firms is identified. For these firms, the ECF also provides establishment-level employment, date-of-birth, and location (which is acquired from the GAL). The SPF contains information on predecessor relationships which may lead to the revision of date-of-birth implied by the ECF. Finally, individual work histories in the EHF in conjunction with place-of-residence information stored in the ICF provide the necessary worker information needed to estimate and apply the imputation model.

0.1 A Probability Model for Employment Location

0.1.1 Definitions

Let $i = 1, ..., I$ index workers, $j = 1, ..., J$ index firms (SEINs), and $t = 1, ..., T$ index time (quarters). Let R_{jt} denote the number of active establishments at firm j in quarter t, let $\mathfrak{R} = \max_{j,t} R_{jt}$, and $r = 1, ..., \mathfrak{R}$ index establishments. Note the index r is nested within j. Let N_{jrt} denote the quarter t employment of establishment r in firm j. Finally, if worker i was employed at firm j in t, denote by y_{ijt} the establishment at which she was employed.

Let $\mathcal{J}_t$ denote the set of firms active in quarter t, let $\mathcal{I}_{jt}$ denote the set of individuals employed at firm j in quarter t, let $\mathcal{R}_{jt}$ denote the set of active ($N_{jrt} > 0$) establishments at firm j in t, and let $\mathcal{R}^i_{jt} \subset \mathcal{R}_{jt}$ denote the set of active establishments that are feasible for worker i. Feasibility is defined as follows. An establishment $r \in \mathcal{R}^i_{jt}$ if $N_{jrs} > 0$ for every quarter s that i was employed at j.

0.1.2 The Probability Model

Let $p_{ijrt} = \Pr(y_{ijt} = r)$. At the core of the model is the probability statement:

$$p_{ijrt} = \frac{e^{\alpha_{jrt} + x'_{ijrt}\beta}}{\sum_{s \in \mathcal{R}^i_{jt}} e^{\alpha_{jst} + x'_{ijst}\beta}} \tag{1}$$

where α_{jrt} is a establishment- and quarter-specific effect, x_{ijrt} is a time-varying vector of characteristics of the worker and establishment, and β measures the effect of characteristics on the probability of being employed at a particular establishment. In the current implementation, x_{ijrt} is a linear spline in the (great-circle) distance between worker i's residence and the physical location of establishment r. The spline has knots at 25, 50, and 100 miles.

Using (1), the following likelihood is defined

$$p(y|\alpha, \beta, x) = \prod_{t=1}^{T} \prod_{j \in \mathcal{J}_t} \prod_{i \in \mathcal{I}_{jt}} \prod_{r \in \mathcal{R}^i_{jt}} (p_{ijrt})^{d_{ijrt}} \tag{2}$$

where

$$d_{ijrt} = \begin{cases} 1 & \text{if } y_{ijt} = r \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

and where y is the appropriately-dimensioned vector of the outcome variables y_{ijt}, α is the appropriately-dimensioned vector of the α_{jrt}, and x is the appropriately-dimensioned matrix of characteristics x_{ijrt}. For α_{jrt}, a hierarchical Bayesian model based on employment counts N_{jrt} is specified.

The object of interest is the joint posterior distribution of α and β. A uniform prior on β, $p(\beta) \propto 1$ is assumed. The characterization of $p(\alpha, \beta|x, y, N)$ is based on the factorization

$$\begin{aligned} p(\alpha, \beta|x, y, N) &= p(\alpha|N)\, p(\beta|\alpha, x, y) \\ &\propto p(\alpha|N)\, p(\beta)\, p(y|\alpha, \beta, x) \\ &\propto p(\alpha|N)\, p(y|\alpha, \beta, x). \end{aligned} \tag{4}$$

Thus the joint posterior (4) is completely characterized by the posterior of α in (**??**) and the likelihood of y in (2). Note (2) and (4) assume that the employment counts N affect employment location y only through the parameters α.

Is the reference here needed? Refers to p(a—N)

0.1.3 Estimation

The joint posterior $p\left(\alpha,\beta|x,y,N\right)$ is approximated at the posterior mode. In particular, we estimate the posterior mode of $p\left(\beta|\alpha,x,y\right)$ evaluated at the posterior mode of α. From these we compute the posterior modal values of the α_{jrt} using (**??**) , then, maximize the log posterior density

Refernce to (a-def) not resolved

Reference to (log-posterior) not resolved

$$\log p\left(\beta|\alpha,x,y\right) \propto \sum_{t=1}^{T}\sum_{j\in\mathcal{J}_t}\sum_{i\in\mathcal{I}_{jt}}\sum_{r\in\mathcal{R}^i_{jt}} d_{ijrt}\left(\alpha_{jrt}+x'_{ijrt}\beta-\log\left(\sum_{s\in\mathcal{R}^i_{jt}} e^{\alpha_{jst}+x'_{ijst}\beta}\right)\right) \tag{5}$$

which is evaluated at the posterior modal values of the α_{jrt}, using a modified Newton-Raphson method. The mode-finding exercise is based on the gradient and Hessian of (5). In practice, (**??**) is estimated for three firm employment size classes: 1-100 employees, 101-500 employees, and greater than 500 employees, using data for Minnesota.

0.2 Imputing Place of Work

After estimating the probability model using Minnesota data, the estimated parameters are applied in the imputation process for other states. A brief outline of the imputation method, as it relates to the probability model previously discussed, is provided in this section. Emphasis is placed on not only the imputation process itself, but also the preparation of input data.

0.2.1 Sketch of Imputation Method

Ignoring temporal considerations, 10 implicates are generated as follows. First, using the mean and variance of β estimated from the Minnesota data, we take 10 draws of β from the normal approximation (at the mode) to $p\left(\beta|\alpha,x,y\right)$. Next, using QCEW employment counts for the establishments, we compute 10 values of α_{jt} based on the hierarchical model for these parameters. Note these are draws from the exact posterior distribution of the α_{jrt}. The drawn values of α and β are used to draw 10 imputed values of place of work from the normal approximation to the posterior predictive distribution

$$p\left(\tilde{y}|x,y\right)=\int\int p\left(\tilde{y}|\alpha,\beta,x,y\right)p\left(\alpha|N\right)p\left(\beta|\alpha,x,y\right)d\alpha d\beta. \tag{6}$$

0.2.2 Implementation

Establishment Data Using state-level micro-data, the set of firms (SEINs) that ever operate more that one establishment in a given quarter are identified; these SEINs represent the set of ever-multi-establishment firms defined above as the set $\mathcal{J}_t$. For each of these firms, its establishment-level records are identified. For each establishment, latitude and longitude coordinates, which emerge from GAL processing, parent firm (SEIN) employment, and QCEW first month employment[1] for the entire history of the establishment are retained. Those establishments with positive first-month employment in a given quarter characterize $\mathcal{R}_{jt}$, the set of all active establishments. An establishment date-of-birth is identified and, in most cases, is the first quarter in the QCEW time series in which the establishment has positive first-month employment. For some firms, predecessor relationships are identified in the SPF; in those instances, the establishment date-of-birth is adjusted to coincided with that of the predecessor's.

Worker Data The EHF provides the earnings histories for employees of the ever-multi-establishment firms. For each in-scope job (a worker-firm pair), one observation is generated for the *end* of each job spell, where a job spell is defined as a continuum of quarters of positive earnings for worker at a particular firm during which there are no more than 3 consecutive periods of non-positive earnings[2]. The start-date of the job history is identified as the first quarter of positive earnings; the end-date is the last date of positive earnings[3]. These job spells characterize the set $\mathcal{I}_{jt}$

Candidates Once the universe of establishments and workers is identified, data are combined and a priori restrictions and feasibility assumptions are imposed. For each quarter of the date series, the history of every job spell that *ends in that quarter* is compared to the history of *every* active (in terms of QCEW first month employment) establishment of the employing firm (SEIN). The start date of the job spell is compared to the birth date of each establishment. Establishments that were born after the start of a job spell are immediately discarded from the set of candidate establishments. The remaining establishments constitute the set $\mathcal{R}^i_{jt} \subset \mathcal{R}_{jt}$ for a job spell (worker) at a given firm[4].

Given the structure of the pairing of job spells with candidate establishments, it is clear that within job spell changes of establishment are ruled-out. An establishment is imputed once for each job spell[5], thereby creating no false labor market transitions.

[1] In rare instances where no QCEW employment is available, an alternative employment measure based on UI wage record counts may be used.

[2] A new hire is defined in the QWI as a worker who acceeds to a firm in the current period but was not employed by the same firm in any of the 4 previous periods. A new job spell is created if, for example, a worker leaves a firm for 4 or more quarters and is subsequently re-employed by the same firm.

[3] By definition, an end-date for a job spell is not assigned in cases where a quarter of positive earnings at a firm is succeeded by fewer than 4 quarters of non-employment and subsequent re-employment by the same firm.

[4] The sample of UI wage and QCEW data chosen for processing of the QWI is such that the start and end dates are the same. Birth and death dates of establishments are, more precisely, the dates associated with the beginning and ending of employment activity observed in the data. The same is true for the dates assigned to the job spells.

[5] More specifically, an establishment is imputed to a job spell only once within each implicate.

Imputation and Output Data Once the input data are organized, a set of 10 imputed establishment identifiers are generated for each job spell ending in every quarter for which both QCEW and UI wage records exist. For each quarter, implicate, and size class, $s = 1, 2, 3$, the parameters on the linear spline in distance between place-of-work and place-of-residence $\hat{\beta}^s$ are sampled from the normal approximation of the posterior predictive distribution of β^s conditional on Minnesota (MN)

$$p(\beta^s|\alpha_{MN}, x_{MN}, y_{MN}) \tag{7}$$

The draws from this distribution vary across implicates, but not across time, firms, and individuals.

Next, for each firm j at time t, a set of $\hat{\alpha}_{jrt}$ are drawn from

$$p\left(\alpha_{ST}|N_{ST}\right) \tag{8}$$

which are based on the QCEW first-month employment totals (N_{jrt}) for all candidate establishments $r_{jt} \subset \mathcal{R}_{jt}$at firm j within the state (ST) being processed. The initial draws of $\hat{\alpha}_{jrt}$ from this distribution vary across time and firms but not across job spells.

Combining (7) and (8) yields

$$\begin{aligned} & p\left(\alpha_{ST}|N_{ST}\right) p(\beta^s|\alpha_{MN}, x_{MN}, y_{MN}) \\ \approx \; & p\left(\alpha_{ST}|N_{ST}\right) p(\beta^s|\alpha_{ST}, x_{ST}, y_{ST}) \\ = \; & p\left(\alpha_{ST}, \beta_{ST}|x_{ST}, y_{ST}, N_{ST}\right) \end{aligned} \tag{9}$$

an approximation of the joint posterior distribution of α and β^s (4) conditional on data from the state being processed.

The draws $\hat{\beta}^s$ and $\hat{\alpha}_{jrt}$ conjunction with the establishment, firm, and job spell data are used to construct the p_{ijrt} in (1) for all candidate establishments $r \in \mathcal{R}^i_{jt}$. For each job spell and candidate establishment combination, the $\hat{\beta}^s$ are applied to the calculated distance between place-of-residence (of the worker holding the job spell) and the location of the establishment, where the choice of $\hat{\beta}^s$depends on the size class of the establishment's parent firm. For each combination an $\hat{\alpha}_{jrt}$ is drawn which is based primarily on the size (in terms of employment) of the establishment relative to other active establishments at the parent firm. In conjunction, these determine the conditional probability p_{ijrt} of a candidate establishment's assignment to a given job spell. Finally, from this distribution of probabilities is drawn an establishment of employment.

Emerging from the imputation process is a data file containing a set of 10 imputed establishment identifiers for each job spell. In a minority of cases, the model fails to impute an establishment to a job spell. This is often due to unanticipated idiosyncrasies in the underlying administrative data. Furthermore, across states, the proportion of these failures relative to successful imputation is well under 0.5%. For these job spells, a dummy establishment identifier is assigned and in downstream processing, the employment-weighted modal firm-level characteristics are used.