Career Trajectories and Occupational Transitions Study

Public Use Data Documentation

December 2021

1. Introduction

This documentation describes the contents of the accompanying Public Use Data (PUD) files. Raw data sources, imputation methods, and other analysis descriptions can be found in the appendices to the *Building Better Pathways* report.

2. Emsi Transitions Data

Data Files

- Stata dataset: This file contains the formatted and labeled data.
- CSV dataset: This file contains unformatted and unlabeled data (please use in conjunction with the codebook for labels and formatting).
- Excel codebook: This file contains an overview of the dataset including number of variables and records, as well as variable information including labels, codes, formats, unweighted frequencies for categorical variables, and unweighted descriptives (min, max, mean, standard deviation, median) for continuous variables.

Dataset Description

The Emsi Transitions dataset includes occupational transitions and their typical associated wage changes from the Occupational Employment and Wage Statistics (OEWS) Survey. The dataset is at the aggregated transition level and contains no individual demographic information, so each record represents all transitions that occurred between that origin and destination occupation. The dataset includes the ranking by frequency of number of transitions that occurred between each origin and destination occupation. More information on the raw Emsi and OEWS data sources and final dataset creation can be found in the *Building Better Pathways* Appendices A, B, and C.

3. CPS/SIPP Transitions Data

Data Files

- Stata dataset: This file contains the formatted and labeled data.
- CSV dataset: This file contains unformatted and unlabeled data (please use in conjunction with the codebook for labels and formatting).
- Excel codebook: This file contains an overview of the dataset including number of variables and records, as well as variable information including labels, codes, formats, unweighted frequencies for categorical variables, and unweighted descriptives (min, max, mean, standard deviation, median) for continuous variables.

Dataset Description

The CPS/SIPP Transitions dataset includes job change data from both the Current Population Survey (CPS) and the Survey of Income and Program Participation (SIPP). Observations are included only if the occupation for the “origin” job was considered mid-level. The SIPP data contains both job and occupational transitions, while the CPS data only contains occupational transitions. This data is at the person-transition level, so each record represents a single transition made and an individual may have multiple transitions. Weights (included in the dataset) should be applied on all analyses to make the sample representative of characteristics

of workers in 2020. More information on the raw CPS and SIPP data sources, weight creation, and final dataset creation can be found in the *Building Better Pathways* Appendices A, B, and C.

4. Three-, Five-, and Ten-Year Trajectories Data

Data Files

- Stata dataset: This file contains the formatted and labeled data.
- CSV dataset: This file contains unformatted and unlabeled data (please use in conjunction with the codebook for labels and formatting).
- Excel codebook: This file contains an overview of the dataset including number of variables and records, as well as variable information including labels, codes, formats, unweighted frequencies for categorical variables, and unweighted descriptives (min, max, mean, standard deviation, median) for continuous variables.

Dataset Description

The Trajectories datasets include career trajectories compiled from longitudinal worker data from the National Longitudinal Survey of Youth (NLSY) and Panel Study of Income Dynamics (PSID). The unit of observation is the "trajectory," meaning that each record begins when an individual takes primary employment in a given occupation for the first time, and tracks month by month employment and education status for the specified dataset duration (three years, five years, or ten years). An individual may have multiple trajectories if their primary employment switched to different "mid-level" occupations during the observed time frame. The three-year dataset contains more trajectories than the five-year dataset or the ten-year dataset since more individuals have three years of data available after starting in a new occupation. Similarly, the five-year dataset contains more trajectories than the ten-year dataset. Weights (included in each dataset) should be applied on all analyses to make the sample representative of workers aged 18-34 in 2020 in terms of demographic characteristics. More information on the raw NLSY and PSID data sources, weight creation, and final dataset creation can be found in the *Building Better Pathways* Appendices A, B, and C.

5. Dashboard Cluster Data

Data Files

- Stata dataset: This file contains the formatted and labeled data.
- CSV dataset: This file contains unformatted and unlabeled data (please use in conjunction with the codebook for labels and formatting).
- Excel codebook: This file contains an overview of the dataset including number of variables and records, as well as variable information including labels, codes, formats, unweighted frequencies for categorical variables, and unweighted descriptives (min, max, mean, standard deviation, median) for continuous variables.

Dataset Description

The dashboard cluster data is compiled from cluster-level regression analysis predictors of wage growth using the ten-year Trajectories datasets. More information on these analyses can be found in the *Building Better Pathways* Appendix C. This data is at the cluster level with some occupational overlap in some of the clusters (see the full Trajectories datasets for occupation-cluster mapping).

6. Dashboard Trajectories Data

Data Files

- Stata dataset: This file contains the formatted and labeled data.
- CSV dataset: This file contains unformatted and unlabeled data (please use in conjunction with the codebook for labels and formatting).
- Excel codebook: This file contains an overview of the dataset including number of variables and records, as well as variable information including labels, codes, formats, unweighted frequencies for categorical variables, and unweighted descriptives (min, max, mean, standard deviation, median) for continuous variables.

Dataset Description

The dashboard trajectories data is compiled from weighted summary statistics of the three-, five-, and ten-year Trajectories datasets. Each record represents a trajectory starting occupation and duration. In other words, each starting occupation has a record with three-year, five-year, and ten-year summary statistics. Additionally, some occupations appear more often if they map to multiple clusters (applies only to occupations that appear in the ECE cluster).[1] Outcome information is suppressed if data for fewer than 40 trajectories were available for the dataset duration for the given occupation. More information on the weighting and underlying Trajectories datasets can be found in the *Building Better Pathways* Appendices A, B, and C.

7. Dashboard Transitions Data

Data Files

- Stata dataset: This file contains the formatted and labeled data.
- CSV dataset: This file contains unformatted and unlabeled data (please use in conjunction with the codebook for labels and formatting).
- Excel codebook: This file contains an overview of the dataset including number of variables and records, as well as variable information including labels, codes, formats, unweighted frequencies for categorical variables, and unweighted descriptives (min, max, mean, standard deviation, median) for continuous variables.

Dataset Description

The dashboard transitions data is compiled from the Emsi Transitions dataset. This data is at the aggregated transition level and shows the top thirty most common destination occupations for each origin occupation (most frequent transitions), with all other destination occupations collapsed into an “Other” record. More information on the Emsi Transitions dataset can be found in the *Building Better Pathways* Appendices A, B, and C.

8. References

The report, *Building Better Pathways: An Analysis of Career Trajectories and Occupational Transitions*, includes additional information on the background and motivation for compiling these data files. The appendices to the report provide more detailed descriptions of the raw data sources, dataset creation, and analysis methods.

[1] Occupations that map to multiple clusters have separate records for each cluster since some summary statistics vary by cluster. In other words, occupations in the ECE cluster that also appear in another cluster appear twice as often as other occupations in the dataset.