**Summary of Datasets and a Variable Dictionary**

1. aggregate\_establishments\_cumulative\_all.dta
2. aggregate\_establishments\_groups.dta
3. region\_level\_cumulative.dta
4. nworkers\_bins.dta
5. entrants\_only\_15\_5y\_PUMA.dta
6. hours\_decomposed\_jobslt19.dta

This data dictionary is for the data associated with Jardim et al., 2021. Section A describes the datasets for which we seek disclosure. Section B provides detailed information on the geographies used in the Section A datasets.

**Section A: Overview of Datasets**

1. **Dataset 1. aggregate\_establishments\_cumulative\_all.dta**

Data are identified at the level. Wage groups range from less than $12 per hour to less than $25 per hour.[[1]](#footnote-1) The variables included in the dataset are as follows:

* region\_time: unique identifier for PUMA by yearquarter
* region\_wageX: identifier for region and wage group
* cum\_payroll\_flowX : total payroll paid to all jobs with hourly wages *less than $X.00 per hour* in quarter t
* cum\_payroll\_begX : total payroll paid to jobs with hourly wages *less than $X.00 per hour* that existed at the beginning of quarter t
* cum\_hours\_flowX: total hours of jobs with hourly wages *less than $X.00 per hour* in quarter t
* cum\_nworkers\_begX: number of jobs which pay less *than $X.00 per hour* that existed at the beginning of quarter t
* d\_cum\_payroll\_flowX: year-over-year change in total payroll paid to all jobs with hourly wages *less than $X.00 per hour* in quarter t
* d\_cum\_payroll\_begX: year-over-year change in total payroll paid to jobs with hourly wages *less than $X.00 per hour* that existed at the beginning of quarter t
* d\_cum\_hours\_flowX: year-over-year change in total hours of jobs with hourly wages *less than $X.00 per hour* in quarter t
* d\_cum\_nworkers\_begX: year-over-year change in number of jobs which pay less *than $X.00 per hour* that existed at the beginning of quarter t
* cum\_earningsX: cum\_payroll\_begX / cum\_nworkers\_begX
* cum\_mean\_wagerateX: cum\_payroll\_flowX / cum\_hours\_flowX
* d\_cum\_mean\_wagerateX: year-over-year change in cum\_payroll\_flowX / cum\_hours\_flowX
* L4\_cum\_hours\_flowX= total hours of jobs with hourly wages *less than $X.00 per hour* in quarter t, lagged by 4 quarters
* L4\_cum\_nworkers\_begX: number of jobs which pay less *than $X.00 per hour* that existed at the beginning of quarter t, lagged 4 quarters
* yearquarter: variable combining year and quarter
* puma\_id0: Puma identifier in WA state. See Section B.
* year: year
* quarter: quarter
* date\_q: a unique identifier for yearquarter
* region0: Regions defined by WA county. See Section B.
* R: treatment indicator

**Dataset 2. aggregate\_establishments\_groups.dta**

Data identified at the level. The three wage groups in this dataset are binned: 1) for jobs paying between $19-24.99 (X = 20) , 2) jobs paying between $25-39.99 (X = 30) , and 3) for jobs paying $40 or more an hour (X= 40).

* region\_time: unique identifier for PUMA by yearquarter
* region\_wageX: identifier for region and wage group
* bin\_hours\_flowX : total hours among jobs with hourly wages between $XX-XX.XX per hour in quarter t
* bin\_nworkers\_begX: number of jobs which pay between $XX-XX.XXthat existed at the beginning of quarter t
* bin\_mean\_wagerateX : mean wagerate among jobs with hourly wages between $XX-XX.XX per hour in quarter t
* d\_bin\_hours\_flowX : year-over-year change in total hours among jobs with hourly wages between $XX-XX.XX per hour in quarter t
* d\_bin\_nworkers\_begX: year-over-year change in number of jobs which pay between $XX-XX.XXthat existed at the beginning of quarter t
* d\_bin\_mean\_wagerateX: year-over-year change in mean wagerate among jobs with hourly wages between $XX-XX.XX per hour in quarter t
* L4\_bin\_hours\_flowX: total hours among jobs with hourly wages between $XX-XX.XX per hour in quarter t, lagged 4 quarters
* L4\_bin\_nworkers\_begX: number of jobs which pay between $XX-XX.XXthat existed at the beginning of quarter t, lagged 4 quarters
* yearquarter: variable combining year and quarter
* puma\_id0: Puma identifier in WA state. See Section B.
* year: year
* quarter: quarter
* date\_q: a unique identifier for yearquarter
* R: treatment indicator

**Dataset 3. region\_level\_cumulative.dta**

Data are identified at the level. Regions aggregate PUMA variables into geographies loosely defined by county borders. Wage groups range from less than $12 per hour to less than $40 per hour, with an additional variable for >$40 per hour.[[2]](#footnote-2) The variables included in the dataset are as follows:

* region\_time: unique identifier for region by yearquarter
* cum\_payroll\_flowX : total payroll paid to all jobs with hourly wages *less than $X.00 per hour* in quarter t
* cum\_hours\_flowX: total hours of jobs with hourly wages *less than $X.00 per hour* in quarter t
* cum\_nworkers\_begX: number of jobs which pay less *than $X.00 per hour* that existed at the beginning of quarter t
* cum\_mean\_wagerateX = cum\_payroll\_flowX / cum\_hours\_flowX
* d\_cum\_payroll\_flowX : year-over-year change in total payroll paid to all jobs with hourly wages *less than $X.00 per hour* in quarter t
* d\_cum\_hours\_flowX: year-over-year change in total hours of jobs with hourly wages *less than $X.00 per hour* in quarter t
* d\_cum\_nworkers\_begX: year-over-year change in number of jobs which pay less *than $X.00 per hour* that existed at the beginning of quarter t
* d\_cum\_mean\_wagerateX = year-over-year change in cum\_payroll\_flowX / cum\_hours\_flowX
* yearquarter = variable combining year and quarter
* region0: Regions defined by WA county. See Section B.
* year = year
* quarter = quarter
* date\_q: a unique identifier for yearquarter

**Dataset 4. nworkers\_bins.dta**

Data identified at level.

* bin\_nworkers\_begX: number of workers with hourly wages between $X.00 and $X.99 per hour at quarter t where X is between 11 and 39.99. There are two exceptions: the variable bin\_nworkers\_beg11 denotes the number of workers with hourly wages below $11.99 and the variable bin\_nworkers\_beg40 denotes the number of workers with hourly wages above $40 per hour.
* NX: the number of firms in each level.
* yearquarter: variable combining year and quarter
* region0: Regions defined by WA county. See Section B.
* year: year
* quarter : quarter
* date\_q: a unique identifier for yearquarter
* T: treatment indicator

**Dataset 5. entrants\_only\_PUMA.dta**

Data identified at the level for jobs paying < $15 per hour. Variables include:

* quarter : yearquarter
* entXXXXX: Number of entrants in PUMA XXXXX. See Section B.
* seattle\_entrants : Number of entrants in Seattle
* PXXXXX\_MA: moving average of number of entrants in PUMA XXXXX

**Dataset 6. hours\_decomposition.dta**

Data is identified at the level for jobs paying <$19 per hour. Variables include:

* Yearquarter: variable combining year and quarter
* region: Regions defined by WA county. See Section B.
* puma\_id: Puma identifier in WA state. See Section B.
* year=year
* quarter: quarter
* date\_q: date (year, quarter)
* hours\_all: total hours of jobs with hourly wages *less than $19.00 per hour* in quarter t
* L4\_hours\_all: total hours of jobs with hourly wages *less than $19.00 per hour* in quarter t, lagged by 4 quarters
* region\_puma\_id : identifier of region and puma id
* d\_hours\_missing: hours worked by workers who were not observed at baseline
* d\_hours\_hires\_seps: hours worked by newly hired workers with wages <$19 per hour
* d\_hours\_seps : hours of separations paying less than $19 per hour
* d\_hours\_contjobs: hours of workers employed in continuing jobs paying less than $19 per hour
* d\_hours\_up : hours worked by workers whose wages increased to >$19 per hour
* d\_hours\_down: hours worked by workers whose wages that decreased to <$19 per hour
* d\_hours\_all0: year-over-year change in total hours of jobs with hourly wages *le*ss than $19 per hour

**Section B. Geography**

These datasets rely on two levels of geography: regions and PUMAs. PUMAs correspond to [Public Use Microdata Areas in WA state (2010)](https://www.ofm.wa.gov/washington-data-research/population-demographics/gis-data/census-geographic-files/supplemental-information-gis-files/public-use-microdata-area-maps). Regions aggregate PUMAs up and are loosely defined by county lines.

Region = 1 is comprised of the PUMAs that make up Seattle. These include PUMA IDs 11601 , 11602 , 11603, 11604 , 11605

Region =2 is comprised of the PUMAs that make up SeaTac, outside of Seattle. These include PUMA ID 11611 and 11613.[[3]](#footnote-3)

Region = 3 is comprised of King county, outside of Seattle. These include PUMA IDs 11606 , 11607, 11608 , 11609, 11610 , 11611 , 11612 , 11613, 11614 , 11615 , 11616

Region =4 is comprised of Snohomish, Kitsap, Peirce Counties: These include PUMA IDs 11501 , 11502 , 11503 , 11504 , 11505 , 11506 , 11507 , 11701 , 11702 , 11703 , 11704 , 11705 , 11706 , 11801 , 11802

Region =5 is comprised of the rest of WA counties. These include PUMA IDs 10100 , 10200 , 10300 , 10400 , 10501 , 10502 , 10503 , 10504, 10600 , 10701 , 10702 , 10703 , 10800 , 10901 , 10902 , 11000 , 11101 , 11102, , 11103 , 11104 , 11200 , 11300 , 11401 , 11402 , 11900

Region =9 are jobs in firms that did not have precise addresses in the ESD data. These jobs may have had city, state, or zipcode information that ArcGIS used to put these firms into space in Washington (and thus is technically comprised of all PUMAs in Washington), however the PUMA information associated with this region is invalid and not used in analysis. Jobs in region 9 are counted when assessing the total number of jobs in Washington state.

1. The X in the variable name is actually one higher than the $X.00-X.99. So for example, cum\_hours\_beg17 is the sum of all hours worked by workers with hourly wages less than 16.99 in that puma at the beginning of that quarter. [↑](#footnote-ref-1)
2. The X in the variable name is actually one higher than the $X.00-X.99. So for example, cum\_hours\_beg17 is the sum of all hours worked by workers with hourly wages less than 16.99 in that puma at the beginning of that quarter. [↑](#footnote-ref-2)
3. Note PUMA 11613 would also be in Region 2, however there were many yearquarters in which this PUMA had fewer than four firms, and so this PUMA is omitted from the PUMA level analysis. It is included in the Region analysis and aggregated with PUMA 11611, and the combined “PUMA” has more than four firms in all quarters. [↑](#footnote-ref-3)