Replication instructions

Purpose

Supplementary Material S1 reproduces the manuscript's employment results, historical tests, appendix checks, and four figures. The primary window is 2022Q4–2025Q4. The same 47 states and the District of Columbia are followed from 2006Q1 through 2025Q4.

Run the package

Use Python 3.14.3 with the versions listed in requirements.txt. From the Replication folder, install the dependencies with python -m pip install -r requirements.txt. Then run python run_all.py. Installation needs internet access if those packages are not already available. Once installed, all calculations run offline from the supplied inputs. No AI tool, model account, API key, or paid service is required.

Public data

inputs/qwi contains frozen extracts for all 50 states and the District of Columbia from Census QWI release R2026Q3. inputs/qwi_2025Q4_extension.csv contains the additional 1,200 published observations for the 48 retained areas, five industries, and five age codes. Raw-source inventories and release metadata are retained. The analysis uses private ownership, both sexes combined, unadjusted state geography, and five-digit NAICS 54151, 54161, 54121, 54111, and 54133. Emp counts jobs on the first day of the quarter. A00 measures all ages. A03 measures ages 22–24; older employment sums A04, A05, and A06, ages 25–54. Only positive young and older values with publication status 1 qualify for the balanced panel.

Source reconstruction

Public source archives are at https://lehd.ces.census.gov/data/qwi/R2026Q3/ and have names such as qwi_ca_sa_f_gs_n5_op_u.csv.gz. Definitions are in schema V4.14.0 at https://lehd.ces.census.gov/data/schema/V4.14.0/lehd_public_use_schema.html. Optional scripts reconstruct_inputs.py and extract_q4_extension.py read downloaded archives supplied through --qwi-dir PATH and write to reconstructed_inputs. They do not download files or replace the supplied inputs. Full source archives are much larger than the frozen extracts and are unnecessary for the offline run.

Reported historical tests

Each industry is compared with its own history of annual Q4 log-ratio changes ending in 2007–2019. The recent departure subtracts three times the historical mean annual change from the three-year change. Areas receive equal weight. The circular year-block bootstrap draws whole years with all areas together, uses three-year blocks and 30,000 draws, and includes uncertainty in the historical mean. Seed: 20261008. The two-sided p-value counts simulated errors whose absolute log size reaches the observed departure. Bounds invert that same rule. Each industry's sensitivity checks remain separate. The primary dispersion test retains both the standard-deviation and range statistics, together with the reported industry-removal checks.

Primary calculations and verification files

analysis.py performs the main calculations. independent_audit.py reads the same frozen inputs and recalculates the employment endpoints and primary historical test through a separate implementation. It checks agreement with analysis.py and also creates data for the figures and historical comparisons. The independent_ prefix identifies files produced by that second implementation. These files describe the same study and sample; they are not additional independent studies or separate evidence of statistical significance.

independent_levels.csv contains the employment endpoints and percentage changes used in Table 1. independent_inference.csv contains the primary historical departures, p-values, and intervals used in Table 2. These are checked against table1_levels.csv and the primary rows of table2_tests.csv.

independent_series.csv contains the quarterly employment indexes used in Figure 1. independent_state_departures.csv contains each reporting area's recent ratio change relative to its own historical mean. independent_validation.json records the checks performed by independent_audit.py, including the source-file fingerprints.

The final dispersion analysis is stored in dispersion_test.csv. Its reported industry-removal checks are in dispersion_industry_removal_checks.csv. State-level young and older employment changes used to verify geographic counts are in state_age_changes.csv.

Files supporting the manuscript

Table 1 uses independent_levels.csv and is checked against table1_levels.csv. Table 2 uses independent_inference.csv, checked against the primary rows in table2_tests.csv. Table A1 uses the individual-industry sensitivity rows in table2_tests.csv. Table A2 uses historical_record_comparison.csv. The 2019 baseline and earlier employment comparisons use table_A_2019_base.csv and alternative_explanations_verified_levels.csv. Geography uses q4_geography_counts.csv and independent_state_departures.csv. Figure 1 uses independent_series.csv; Figure 2 uses the 2025Q4 rows of state_historical_windows.csv; Figure 3 uses age_changes.csv; Figure 4 uses rolling_dispersion.csv. Original PNG and vector PDF outputs are in output/figures.

Geographic definitions

Own-history counts compare each area with that area's historical mean. They differ from comparisons with one common industry-wide mean. For consulting, 46 areas fall below their own history. The package reports that measure consistently. Employment sums weight areas by their number of jobs; historical ratio tests weight areas equally.

Sample checks

The sector scores and industry size inputs reproduce Appendix A. Occupational exposure scores use employment weights within industries; the sector ranking is an arithmetic mean of four-digit industry scores. The size check uses a separate 49-area panel. These files document the chosen sample. They do not reconstruct the original task annotations or the entire OEWS weighting process from raw occupation files.

Verification

independent_audit.py separately reconstructs the endpoints and primary historical test from the frozen inputs. q4_extension.py supplies a further Q4 cross-check. verify_geography.py checks the reported area counts and added rows. check_outputs.py tests consistency between the estimates, bounds, history, and figure files. The validation JSON files record the checks. raw_source_spot_checks.json is a retained source-validation record, not an output regenerated by the offline run. SHA256_manifest.csv records file hashes before rerunning.

Interpretation

The tests measure departures from the historical comparison model. They do not identify AI's causal contribution or incorporate all QWI imputation and disclosure uncertainty. The study counts jobs by age and industry; it does not identify individual occupations, graduates, first jobs, or the same people as they age.