Replication files of "Technological Waves, Knowledge Diffusion, and Local Growth" by Enrico Berkes, Ruben Gaetani, and Martí Mestieri

Datasets provided in the package

These folders contain the datasets that constitute the starting point of the empirical analysis and quantitative model. All the other datasets are created by the codes described in the "**Instructions for Data Preparation and Analysis**" section below.

- "**dsets\cz_data\population**": This subfolder contains **csv** files of CZ-level population for each decade in the sample, created following the procedure described in Online Appendix B

- "**dsets\cz_data\cz_state_region_division.dta**": This file associates each CZ to a Census region, division, and state

- "**dsets\cz_data\cz_state_region_division_withfips.dta**": This file associates each CZ to a Census region, division, and state, and includes State FIPS codes

- "**dsets\cz_data\panel_hc_summary_cpp.dta**": This file contains the harmonized ranking of human capital at the CZ-decade level, created following the procedure described in Online Appendix B

- "**dsets\cz_data\income**": This folder contains CZ-level income data for 1950-2010 and income by city-sector for 1990 and 2010, as described in Online Appendix B

- "**dsets\geo_crosswalks**": This folder contains the area-based crosswalks from State Economic Areas (in 1950), County Groups (in 1970), and PUMAs (in 1990 and 2010) to 1990 CZs, as described in Online Appendix B

- "**dsets\ipums_data_industries**": This folder contains the data extracted from the NHGIS and IPUMS to construct the CZ-decade-level measures of industry composition. Accessed July 2021.
    - "**industry_yyy_to_counties90.csv**", for yyy between 1870 and 1930, contains employment count by industry and county, with associated CZ.
    - "**usa_00036.dat**", "**usa_00037.dat**", "**usa_00038.dat**", "**usa_00039.dat**", contains the IPUMS extracts that are used to construct employment count by industry in 1950, 1970, 1990, and 2010 (in the code "**codes\work_with_ipums_industries_cpp.do**", see point 1 in Part 1 below)

- "**dsets\migration":**
    - **"cz_bilateral_distance.dta**": This file contains the geographical distance in km for each pair of CZs
    - **"cz_bilateral_distance_for_gravity_cpp.dta**": This file contains the geographical distance in km for each pair of CZs in the final sample
    - "**migration_cz_origin_destination_prob_cpp_onlyyoung.dta**": This dataset contains the probability of lifetime migration for "young adults" (age 20-40) obtained from the 1990 IPUMS (destination given origin)

    - "**migration_cz_origin_given_dest_cpp_onlyold.dta**": This dataset contains the probability of lifetime migration for "old adults" (age 41-60) obtained from the 1990 IPUMS (origin given destination)

- Patents data: These datasets were created starting from the patent-level data described in Berkes (2018)
    - "**dsets\patents_data\cz_decade_class_patents_cpp.dta**": This dataset contains the patent count by CZ, decade, and class-group
    - "**model\dsets_for_model\patents_data**":
        - "**cz_decade_class_patents_cpp.csv**": Patent count by CZ, decade, and class-group
        - "**citprob_classes_1950_1970.csv**": Bilateral citation probabilities across class-groups, 1940-1979
        - "**citprob_classes_1990_2010.csv**": Bilateral citation probabilities across class-groups, 1980 onwards
        - "**citprob_classes_geq1950.csv**": Bilateral citation probabilities across class-groups, 1940 onwards
        - "**dataset_for_gravity_knowflow.dta**": Bilateral citation probabilities across commuting zones and class-groups, as described in Appendix E.3
        - "**macroipc_list.csv**": List of class-group in Appendix Table A.1

- "**dsets\temp_datasets_todelete**": This folder is initially empty, but will be populated with some temporary datasets created by the codes

Computational Requirements

For the empirical analysis (except for producing figures) we used Stata 15.1 with the packages "**reghdfe**" and "**ppmlhdfe**" installed (these packages can be installed via the command "**ssc install [*name of the package*]**")

To produce the figures and to run the model, we used Matlab R2018b

We expect the code to run with other versions of Stata and Matlab compatible with the packages listed above.

Computer specifications used to run the codes: Windows 10 Enterprise, 64GB RAM

Instructions for Data Preparation and Analysis

To generate the tables, figures, and results in the paper, the following codes must be run in the following order:

Part 1 - Codes that generate tables and figures in the empirical analysis

1. **"codes\empirical_analysis.do":** generates the results in Table 1 and Appendix Table A.3 (exported to the folder "**tables**") and Table A.2 (unformatted, lines 263 in the code). The current directory must be set to the one of the deposit. Running time: about 5min

- **NOTE:** This code also generates other datasets that are used as inputs for other codes, so it is important to execute this code first before moving to the following steps

- This code also calls the code "**codes\work_with_ipums_industries_cpp.do**", constructing the local industry shocks

Part 2 - Codes that generate tables and figures for the quantitative model, and other figures in the paper

1. "**model\gravity_codes**": This folder contains two **.do** files that run the regressions of Appendix Table E.5 and Table E.6, and generate other files that will be used as inputs in the quantitative analysis. The current directory must be set to the one of the deposit.

    - **"estimate_gravity_migration.do":** Runs the regression of Table E.5. Running time: about 1min
    - **"estimate_gravity_knowflow.do"**: Runs the regression of Table E.6. Running time: about 25min
        - **NOTE:** This step is computationally intensive and can be skipped without creating problems for the remainder of the replication (the resulting files are already included in the replication package).

2. The Matlab codes in the folder "**model**" must be run *in the following order*. The working directory should be set to "**model**" (that is, the same directory where the codes are located)

    - "**calibr_ss_1890.m**": This code computes the full model between the initial steady state in 1890 and the last period in 2010. Running time: about 1h41min
    - "**lom_lambda.m**": This code computes the path of lambda and alpha, and produces Appendix Figure E.14. Running time: about 2min
    - "**counter.m**": Executes the main counterfactuals and saves the data (tabulated results and figures will be produced in later codes). It also produces the file that underlies the map in Figure 4 ("**model\results_model\map_2010.csv**"). Running time: about 4min
    - "**show_identification.m**": Computes the model under alternative values of theta, to illustrate the identification, and generates Appendix Figure A.7. Running time: about 12h45min
        - **NOTE:** This step is computationally intensive and can be skipped without creating problems for the remainder of the replication.
    - "**simul_diversification.m**": Executes the simulations underlying the results described in Section 5.2
        - **NOTE:** This step is computationally intensive and can be skipped without creating problems for the remainder of the replication (the resulting files are already included in the replication package).
    - "**future_scenarios_declining_frictions.m**": Executes the counterfactuals described in Section 5.3.1 and produces the data underlying the maps in Appendix Figure A.11 ("**model\results_model\future_scenarios_declining_frictions_results.csv**") Running time: about 1min

- "**future_scenarios.m**": Executes the counterfactuals described in Section 5.3.2 and produces the data underlying the maps in Appendix Figure A.12 and A.13. ("**model\results_model\future_scenarios_results.csv**") Running time: about 2min
- "**historical_trends.m**": This code generates Figures 1 and 2, and Appendix Figures A.1, A.5, and A.6. Running time: about 1min

3. The Stata code "**model\predictions_quantitative_model.do**" must be run after the codes in Point 2 above have been run. The current directory must be set to the one of the deposit. Generates Table 3, Appendix Table A.4. Running time: about 1min

4. The Stata code "**model\income_validation.do**" produces the datasets used for the validation described in Section 4.3 (the resulting figures will be produced in point 5 below). The current directory must be set to the one of the deposit. Running time: about 1min

5. The matlab code "**model\create_graphs.m**": The working directory should be set to "**model**" (that is, the same directory where the file is located). This code generates Figure 3 and Appendix Figures A.2, A.3, A.4, A.8, A.9, A.10. Running time: about 1min

Part 3 - Codes that generate the robustness table (Appendix F)

To generate the results in Appendix Table F.7, the following codes must be run *in the following order*:

1. The matlab code "**model\run_robustness.m**": The working directory should be set to "**model**" (that is, the same directory where the file is located) Running time: about 2h53min

2. The Stata code "**model\robustness_quantitative_model.do**" produces columns 1, 3, 4, 6, 7 of Appendix Table F.7 (Panels A and B). The other columns are taken from Table 3. The current directory must be set to the one of the deposit. Running time: about 1min