Isolating personal knowledge spillovers: co-inventor deaths and spatial citation differentials

Benjamin Balsmeier, Lee Fleming, and Sonja Lück

DATA AVAILABILITY STATEMENT

The data for this project come from both public use data and commercial data. This is a list of all input datasets used in the paper:

Public use data:

1. Source: Balsmeier et al., 2018. "Machine Learning and Natural Language Processing on the Patent Corpus: Data, Tools, and New Measures." Journal of Economics and Management Strategy.
   Available for download at:
   https://doi.org/10.7910/DVN/KPMMPV
   The following dataset was used:
   - inventor.geo.assignee.combo.disambig.zip (downloaded on November 12, 2019)
     Saved as: ...\data\raw\inventor.geo.assignee.combo.disambig1.dta
     ...\data\raw\inventor.geo.assignee.combo.disambig2.dta
     ...\data\raw\inventor.geo.assignee.combo.disambig3.dta
     ...\data\raw\inventor.geo.assignee.combo.disambig4.dta
     ...\data\raw\inventor.geo.assignee.combo.disambig5.dta
     ...\data\raw\inventor.geo.assignee.combo.disambig6.dta
     ...\data\raw\inventor.geo.assignee.combo.disambig7.dta

2. Source: USPTO PatentsView (United States Patent and Trademark Office PatentsView)
   Available for download at: https://patentsview.org/download/data-download-tables under a Creative Commons Attribution 4.0 International License.
   The following datasets were used:
   - application.zip (downloaded on April 7, 2021)
     Saved as: ...\data\raw\application.dta
   - inventor.zip (downloaded on March 13, 2021)
     Saved as: ...\data\raw\inventor.dta
   - location.zip (downloaded on March 13, 2021)
     Saved as: ...\data\raw\location.dta
   - patent.zip (downloaded on April 7, 2021)
     Saved as: ...\data\raw\patent1.dta
     ...\data\raw\patent2.dta
     ...\data\raw\patent3.dta
     ...\data\raw\patent4.dta
     ...\data\raw\patent5.dta
   - patent_inventor.zip (downloaded on March 13, 2021)
     Saved as: ...\data\raw\patent_inventor.dta
   - persistent_inventor_disambig.zip (downloaded on June 26, 2021)
     Saved as: ..\data\raw\persistent_inventor_disambig.zip

- rawinventor.zip (downloaded on March 13, 2021)
  Saved as: ...\data\raw\rawinventor1.dta
  ...\data\raw\rawinventor2.dta
- uspatentcitation.zip (downloaded on April 6, 2021)
  Saved as: ...\data\raw\uspatentcitation1.dta
  ...\data\raw\uspatentcitation2.dta
  ...\data\raw\uspatentcitation3.dta
  ...\data\raw\uspatentcitation4.dta
  ...\data\raw\uspatentcitation5.dta
  ...\data\raw\uspatentcitation6.dta
  ...\data\raw\uspatentcitation7.dta
  ...\data\raw\uspatentcitation8.dta

3. Source: Kaltenberg, Mary, Adam Jaffe, and Margie E. Lachman, 2021, "Matched inventor ages from patents, based on web scraped sources"
   Available for download at: https://doi.org/10.7910/DVN/YRLSKU
   The following dataset was used:
   - inventor_age_score_gender.csv (downloaded on May 6, 2021)
     Saved as: ...\data\raw\inventor_age_score_gender.csv

4. Source: National Weather Service
   Shapefile for U.S. States and Territories available for download at:
   https://www.weather.gov/gis/USStates
   The following dataset was used:
   - s_11au16.zip (downloaded on July 21, 2021)
     Saved as: ...\data\raw\s_11au16.dbf
     ...\data\raw\s_11au16.prj
     ...\data\raw\s_11au16.shp
     ...\data\raw\s_11au16.shx

5. Source: Own manual city spelling correction (created on March 22, 2021)
   Available for download at: https://doi.org/10.7910/DVN/S5BBFY
   The following dataset was used: ..\data\raw\city_spelling_correction.dta

Commercial data:

6. Source: Simple maps (https://simplemaps.com)
   Researchers can buy the data for $99 or $199 (see https://simplemaps.com/data/us-cities for further information). The basic version is for free but less complete. Replicators will need to buy the Pro or Comprehensive version to get access to the full data. We used the Comprehensive dataset. Filename: uscities.csv, v1.72 (downloaded on November 26, 2020)

SOFTWARE REQUIREMENTS

Stata (codes run with version Stata / MP 15.1)
The following packages were used:

- ftools
- reghdfe
- ppmlhdfe (version 2.2.0)
- estout
- egenmore
- shp2dta

The program setup_stata.do installs all dependencies locally, and should be run once.

MEMORY AND RUNTIME REQUIREMENTS

A computer with the following specifications took several hours to run all Stata codes:
Stata MP 15.1, Intel® Core™ i7-8750H (6 Core) CPU @2.2GHz with 32GB RAM

Do file *...**code**setup_stata.do* took a few minutes.
Do file *...**code**create_sample.do* took 4.5 hours.
Do file *...**code**analyses.do* took 1 hour.
Do file *...**code**FigureA10.do* took a few minutes.
Do file *...**code**replicate_all.do* took 5.5 hours.

INSTRUCTIONS

1. Folder structure

In the directory, there are two folders: code and data.

- Folder *…**code* contains all Stata do files.
- In Folder *...**data*, you find two folders *...**data**raw* and *…**data**work*. You also find the analysis sample dataset and a dataset which identifies the observations defined by coarsened exact matching.
  - Folder *…**data**raw* provides the original used raw data.
  - Folder *...**data**work* is empty and will store the datasets created by the do files.

2. Stata

- Do file *...**code**setup_stata.do* installs all used Stata packages, and should be run once previous to the other do files.
- Do file *...**code**setup_datasets.do* unzips all datasets, and should be run once previous to the other do files.
- Do file *...**code**create_sample.do* merges all relevant raw datasets and generates the analysis sample.
- Do file *...**code**analyses.do* replicates all results (except Figure A10) based on our analysis sample dataset.
- Do file *...**code**FigureA10.do* replicates the map in Figure A10.
- Do file *...**replicate_all.do* serves as a master do file and will run the Stata do files referenced above.

Create a folder *C:\replication* on your computer and copy the folders *code* and *data* to this folder. (If you bought the SimpleMaps data, please save the file uscities.csv in *C:\replication\data\raw*).

All results (except Figure A10) can be replicated with the do file analyses.do and the dataset analyses_sample.dta. Figure A10 can only be reproduced if the commercial data has been purchased from SimpleMaps. The do files create_sample.do and replicate_all.do also require the commercial data from SimpleMaps.

Remarks:
Replicators should expect the number of observations in the Stata output to deviate from the numbers reported in the following tables of manuscript:
Table 2 (Panel A, B, C), Appendix tables A11, A12, A13, A17, A18, A21, A22, A23, A24, A25. The reason is the ppmlhdfe command (version 2.2.0) that is used to estimate Poisson models. It detects so-called 'separated' observations and excludes them from the count of observations that are in the dataset and reported in the manuscript. For further explanation see:

- Correia, Sergio, Paulo Guimarães, and Thomas Zylkin. 2019. "ppmlhdfe: Fast Poisson Estimation with High-Dimensional Fixed Effects". arXiv:1903.01690.
- Correia, Sergio, and Paulo Guimarães, Thomas Zylkin. 2019. "Verifying the existence of maximum likelihood estimates for generalized linear models". arXiv:1903.01633.