README file for the replication package for Ahlfeldt, Albers, Behrens (2025): 'Prime locations', ***American Economic Review: Insights***

Distributed via Havard dataverse doi: 10.7910/DVN/0CYVYN.

Contents

1 Overview 3

2 Data Availability and Provenance Statements 7
- 2.1 Statement about Rights . . . 7
- 2.2 License for Data . . . 8
- 2.3 Summary of Availability . . . 9
- 2.4 Details on each Data Source . . . 10

3 Dataset list 43

4 Computational requirements 46
- 4.1 Software Requirements . . . 46
  - 4.1.1 Clustering (C++) . . . 46
  - 4.1.2 Replication of tables/figures (Stata & Python) . . . 47
  - 4.1.3 Own data collection/transformation of proprietary data/map production (ArcGis & Matlab) . . . 48
- 4.2 Controlled Randomness . . . 48
- 4.3 Memory, Runtime, Storage Requirements . . . 48
  - 4.3.1 Clustering (C++) . . . 48
  - 4.3.2 Replication of tables/figures . . . 49
  - 4.3.3 Own data collection/transformation of proprietary data/map production . . . 49

5 Description of programs/code 49
- 5.1 Overview . . . 49
  - 5.1.1 Clustering . . . 49
  - 5.1.2 Replication of tables and figures . . . 50
- 5.2 License for Code . . . 50

6 Instructions to Replicators 51
- 6.1 Overview . . . 51
  - 6.1.1 C++ clustering . . . 51
  - 6.1.2 Replication of figures/tables . . . 51
  - 6.1.3 A note on the replication of own data collection/ minimal pre-processing/ transformation of proprietary data and large datasets/map production . . . 51
- 6.2 Details . . . 53
- 6.3 List of tables and programs . . . 57

References 63

7 Acknowledgements 64

README template adapted from Vilhuber et al. (2022)

1 Overview

The code in this replication package constructs the analysis file from a large set of data sources. The main analysis can be run using Stata 18, which includes a native Python integration. It reproduces the 3 figures and 2 tables in the paper as well as all appendix figures and tables. The replicator should expect the code to run for about 48-96 hours, depending on the machine. It has been tested on both Windows and Mac Machines (details below). We also provide code to pre-process a large (proprietary) dataset. C++ was used because of speed limitations of other languages. We provide the whole code that was applied to the proprietary data, which can be obtained by a replicator. The corresponding output, which is then processed by the Stata programs, is part of this replication package.

All 27 data sources are detailed in accordance to the AEA guidelines and a full provenance and availability description is provided in this README File. Where we are unable to provide the 'raw data' (= as provided by the originator) because of size restrictions for the replication directory or licensing issues, we provide a detailed access description *and* the relevant data matched to the grid. For this type of matching operation, we used ArcGis Pro/ArcMap. We provide the corresponding shapefiles for the grid such that the replicators will find it easy to match the 'raw data' once they obtain the data by the provider. We provide extensive source-by-source descriptions. However, ArcGis Pro/ArcMap are not necessary to reproduce the figures and tables in the paper and the paper's appendix, except for one purely descriptive 3D representation in a sub-figure of Figure 1 and a purely descriptive display of the sample on the world map—corresponding files are provided.[1] We provide, where applicable, corresponding scripts and describe them in this README file along with the provenance and availability of each data source (Section 2.4).

In the rest of this overview section, we describe how to unzip parts of this directory (as Harvard Dataverse has a 1,000 files limit) and the general structure of this replication directory, both in terms of the analysis and folders.

[1] In addition, for the creation of our own dataset on industry leaders created for this paper, we employed Matlab to georeference manually collected addresses through a script that queries the google API. We provide an example script, but want to make replicators aware that there are now easy-to-use freeware alternatives (e.g. the plugin 'MMQGIS' in QGIS) or that replicators can simply parse the address into `maps.google.com` to retrieve coordinates.

Instructions to download:

- Download from Dataverse ('ReplicationPrimeLocations'). After unzipping the main directory in a location of your choice, you should see the following contents of the root directory:
  - This README file (`README.pdf`)
  - Folder `_Scripts` (code to reproduce tables and figures of paper and appendix)
  - Folder `_Work` (output will be stored here)
  - Folder `_Data` (all raw data is stored here)
  - Folder `LicenseFiles` (including the `LICENSE.txt`)
  - Folder `PythonDependencies` (containing a file `environment.yml` to install the required python dependencies).
  - Folder `Codebook` (containing a short codebook for relevant variables from proprietary datasets)
  - Zipped folder `C++` (used to process proprietary data; further instructions are given below)
- Instructions to unzip:
  - Some folders in the `_Data` folder are zipped. These are unzipped when executing the Stata Master file.
  - Unzip the folder `C++.zip` in the `_Work` folder if you want to replicate the clustering stage of the algorithm.
  - The folder `_Work/TEMP` contains multiple zip archives with all intermediate outputs. It does not need to be extracted to replicate the paper's figures and tables. We provide this folder for those who have problems with the Stata-Python integration. Leaving python-based scripts out can be easily achieved by using setting the global `runpython_scripts` in the Stata Master file to "0". If this setting is chosen, all intermediate outputs in `_Work/TEMP` will be unzipped and replication can be run successfully without python.

General notes on data

The sources and nature of the data used in this replication directory are discussed in the main paper and the appendix. This replication directory contains micro-geographic grid data, from which all exhibits in the paper and the web appendix can be replicated. These data are stored in various subfolders within the `_Data` folder.

Please note that some of the underlying point pattern data aggregated to the grid level are proprietary (for example, the National Economic Time Series Establishment data). We provide a discussion of the origin of these data and how they can be accessed and then aggregated to grid cells in both this `README` and individual `README`s the respective subfolders from which the grid data are loaded.

General structure of this replication directory

Lines 2–17 of the prime location identification algorithm (Algorithm 1, in Appendix Section B.1), although conceptually straightforward, are computationally demanding due to the dimensions of our point pattern datasets. Therefore, we implement it in C++ (see the dedicated Section 3a below for additional details). We refer to this stage of the analysis as 'clustering.' The remaining parts of Algorithm 2, as well as all subsequent empirical analyses, are carried out using standard software (Stata and Stata's Python integration).

To make the replication directory more accessible, we provide all outputs of the clustering stage as data inputs for the substantive analysis. Therefore, all exhibits in the main paper and the appendix can be generated using scripts called from the following master script:

Table 1: Location and description of master file to replicate figures and tables

Name	Content
`_Work/Scripts/0_Master_PrimeLocations.do`	A Stata do-file that calls various other Stata and Python scripts generates the results reported in the main paper and the appendix.

To successfully execute the master do-file (or any of the scripts called by the master do-file), please set the global root to your local path where you have saved the replication directory. All other paths used in the scripts are relative and will work without further adjustment as long as the folder structure of the replication directory remains unchanged.

All scripts must be executed in the exact order in which they are listed in the master do-file. This will generate all exhibits in the main paper and the appendix, with the exception of the left column of Figure 1, which is purely descriptive and was created in ESRI ArcScene (we provide all inputs and project files in a separate folder).

The meta do-file covers the analysis for the US metropolitan areas (in Section 2) and Global Cities (in Section 3). We indicate in the do-file where outputs from the C++ clustering algorithm are read in as inputs. Additionally, we specify in the meta do-file where weights are generated as inputs for the C++ clustering algorithm when creating clusters for Global Cities.

Folder structure of this replication directory

For the replicators orientation and convenience, each folder contains a `_README.txt` file describing the contents of its subfolders. In folders containing data, we also provide a discussion of the origin of the underlying data and how they can be accessed. Below, we highlight the main folders and their contents.

Table 2: Folder structure of `_Data` directory

Name	Content
125 GLOBAL CITIES	Folder containing the input data and outputs of the clustering algorithm.
US METRO DATA	Folder containing the input data for the analysis of US MSAs presented in Section 2.

Table 3: Folder structure of `_Work` directory

Name	Content
C++	Folder containing C++ clustering algorithm codes (lines 2-17 in Algorithm 1) and additional scripts for computationally demanding spatial analyses.
GIS2D	Folder containing the data and project files used to generate the 2D map in Appendix Figure A.2.1.
GIS3D	Folder containing the data and project files used to generate the 3D maps in Figure 1 (left column).
log	Folder containing the log file for the entire analysis covered by the master do file. The folder will be generated when running the master do file.
Output	Folder containing all outputs presented in the paper and shared via the GitHub toolkit, except for those in the left column of Figure 1. The folder will be generated when running the master do file.
Scripts	Folder containing Stata do files (some of which include Python code) that produce all outputs saved in the Output folder.
TEMP	Temporary folder that will be generated when running the master do file. It contains intermediate outputs only.

2 Data Availability and Provenance Statements

2.1 Statement about Rights

- ☒ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
- ☒ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package. Appropriate permission are documented in the LICENSE.txt file.

2.2 License for Data

The data that we create are licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) License. See LICENSE.txt for details. The file is located in `LicenseFiles/Documentationoflicensesforrepublishedrawdata`.

Regarding the licenses for the redistribution of raw data, consult Table 7 and Table 8 for the respective licenses. Please ensure the correct attribution when using the data. The licenses are also documented source-by-source in:
`LicenseFiles/Documentationoflicensesforrepublishedrawdata`.

2.3 Summary of Availability

☐ All data **are** publicly available.
☒ Some **propriatary raw** data **cannot be made** publicly available. **However, all data are made available at a slightly aggregated grid-cell level.**
☐ **No data can be made** publicly available.

☒ Confidential data used in this paper and not provided as part of the public replication package will be preserved for **5** years after publication, in accordance with journal policies.
☒ We commit to provide reasonable assistance to requests for clarification and replication.

2.4 Details on each Data Source

On the following pages, we list all 26 datasets either created or used for this paper. We use numerate them with IDs with a #. This allows the replicator to easily navigate in the case of cross-references and to find the source in the summary table in Section 3.

To ease the readability for the replicator, we created a standard 'questionnaire' for each data source:

1. Dataset full name:
2. Originator (dataset citation):
3. Full URL/DOI:
4. Raw data provided in public deposit other than ours: **Yes/No**
5. Data provided in our public deposit: **Yes/No**
6. Raw data provided in our public deposit: **YES/No**
    - If applicable: Reason for not providing raw data (e.g. data are proprietary), how it can be accessed and processed; transformation provided in the replication package.
7. Transformation of data are provided: **Yes/No**
8. Format:
9. List availability in the package:
10. Variable dictionary:
11. Additional licensing information (if applicable):

[#1] National Establishment Time-Series (NETS)

1. Dataset full name: National Establishment Time-Series (NETS)
2. Originator (dataset citation): Walls & Associates (2024)
3. Full DOI: not available, data are proprietary.
4. Raw data provided in public deposit other than ours: **No**
5. Data provided in our public deposit: **Yes, in transformed version**
6. Raw data provided: **No**
   - Reason for not providing raw data: Data are proprietary
   - It can be accessed and processed with the following procedure and cost:
     - Cost: USD 30,000
     - Detailed description of access modalities and data processing:
       (a) Send email to Don Wallls (dwalls2@earthlink.net).
       (b) Ask for full dataset and pay the required sum.
       (c) Original data by Walls & Associates (2024) are provided in ASCII CSV format.
       (d) The data can then be processed with the C++ program provided and documented in in `\PrimeLocationsReplication\_Work\C++\extract_NETS`. Please see the associated README file for additional details. The code will: (i) extract the specified year from the NETS data; (ii) reconstruct the geographic coordinates based on the NETS move history file; and (iii) split the data by CBSA bounding box. The final data that will be used in the clustering algorithm is then saved as ASCII CSV file under the form plants_[cbsa_id].txt
       (e) Data are aggregated to clusters with the Mac OS executable program clustering_cmd (command-line version) or clustering_gui (with graphical user interface; requires a recent Metal system to be installed). Those executables are provided in pre-compiled form or as C++ source code. We describe the programs and their clustering output below. Everything (programs and clustering output) is made available in our replication package.
7. Transformation of data are provided: **Yes**
   We have obtained permission from Walls & Associates to share the aggregate employment data at the grid level, which we make available in the aforementioned files. However, we cannot provide access to the underlying establishment-level data. We also provide access to the output of the cluster algorithm which makes use of the establishment-level data.
8. Format: `.txt`
9. List availability in the package:
   - `PrimeLocationsReplication\_Data\USMETRODATA\CLUSTEROUTPUT`
     It contains 6 subfolders
     - `cells_p98_5`,
     - `cells_p99_0`,
     - `cells_p99_5`,
     - `cells_p99_9`,
     - `cells_validation_first`,
     - `cells_validation_last`)

In each of these, the cell-level employment data by industries are provided, but the clustering output differs. File names in all subfolders have the following structure:`gridcells_XXXXX_YY_output.txt`, where XXXXX is the CBSA identifier and YY indicates the establishment employment input into the algorithm (3: tradable services; 5: search terms; 99: total employment; 0: Employment predicted for the second batch of cities using weights from the first batch; 1: Employment predicted for the first batch of cities using weights from the second batch).

10. Variable dictionary:

Table 4: Dictionary for NETs data files

v3	cell row identifier
v4	cell col identifier
v6	cell UL longitude (in radians)
v7	cell UL latitude (in radians)
v8	cell LR longitude (in radians)
v9	cell LR latitude (in radians)
v11	employment (manufacturing and wholesale)
v12	employment (non-traded services)
v13	employment (tradable services)
v14	employment (public services)
v15	employment (other)
v17	total employment
v19	ClusterID (0 = no cluster)
v21	Unique cell ID (string)
v23	Undevelopable (1= not developable)

11. Additional licensing information (if applicable): Permission was obtained to publish cell-level aggregates in this replication directory.

[#2] TIGER/Line® Shapefiles

We employ TIGER/Line® Shapefiles for both Metropolitan areas (part of the Core Based Statistical Areas) and ZIP Code Tabulation Areas (ZCTAs).

2a: Official delineation of Core Based Statistical Areas (CBSAs)

1. Dataset full name: 2014 TIGER/Line® Shapefiles: Core Based Statistical Areas
2. Originator (dataset citation): U.S. Census Bureau (2014)
3. Full DOI/URL: https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2014&layergroup=Core+Based+Statistical+Areas (last accessed February 2, 2026). Given the unsecure future of the data's preservation, a copy of it has been stored in this Dataverse (https://doi.org/10.7910/DVN/OCYVYN) in the path specified below (9.).
4. Raw data provided in public deposit/website other than ours: **Yes**
   (a) Click on the above link.
   (b) Under the header "Metropolitan/Micropolitan Statistical Area (current)", click "download".
   (c) Unzip the file `tl\_2014\_us\_cbsa.zip` to access the shapefile.
5. Data provided in our public deposit: **Yes**
6. Raw data provided in our archive: **Yes**
7. Transformation of data are provided: **Not applicable**
8. Format: `.shp`
9. List availability in the package:
   - Shapefile and documentation: `PrimeLocationsReplication/_Data/USMETRODATA/GISData/USMETROGRIDS/GRIDCREATIONDATAANDSCRIPTS/OFFICIALDELINIATION/tl_2014_us_cbsa`
10. Variable dictionary: see variable names in shapefile
11. Additional licensing information (if applicable): The documentation (provided in `Licenses` folder states: 'Copyright protection is not available for any work of the United States Government (Title 17 U.S.C., Section 105). Thus, you are free to reproduce census materials as you see fit.'

2b: Official delineation of ZIP-Code-Tabulation-Areas (ZCTAs)

1. Dataset full name: 2020 TIGER/Line® Shapefiles: ZIP Code Tabulation Areas
2. Originator (dataset citation): U.S. Census Bureau (2014)
3. Full DOI/URL: https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2020&layergroup=ZIP+Code+Tabulation+Areas (last accessed February 20, 2026). Given the unsecure future of the data's preservation, a copy of it has been stored in this Dataverse (https://doi.org/10.7910/DVN/OCYVYN) in the path specified below (9.).
4. Raw data provided in public deposit/website other than ours: **Yes**
   (a) Click on the above link.
   (b) Under the header "ZIP Code Tabulation Areas (2020)", click "download".
   (c) Unzip the file `tl_2020_us_zcta520.zip` to access the shapefile.
5. Data provided in our public deposit: **Yes**

6. Raw data provided in our archive: **Yes**
7. Transformation of data are provided: **Not applicable**
8. Format: `.shp`
9. List availability in the package:
    - Shapefile and documentation: `PrimeLocationsReplication/_Data/USMETRODATA/GISData/CountyBusinessPatternBoundaries/tl_2020_us_zcta520`[2]
10. Variable dictionary: see variable names in shapefile
11. Additional licensing information (if applicable): The documentation (provided in `Licenses` folder states: 'Copyright protection is not available for any work of the United States Government (Title 17 U.S.C., Section 105). Thus, you are free to reproduce census materials as you see fit.'

[2]We have also archived the 2010 version of the Tiger/Line file shall replicators want to access it (not necessary for replication of results as the 2020 version has been used); the download process is analogous as is the location in the directory.

[#3] ESA WorldCover 10 m 2020 v100

1. Dataset full name: ESA WorldCover 10 m 2020 v100
2. Originator (dataset citation): European Space Agency (2022), Zanaga et al. (2022)
3. Full DOI: https://doi.org/10.5281/zenodo.7254221
4. Raw data provided in public domain: **Yes**, please click link above to download the data.
5. Data provided in our public deposit: **Yes, in transformed version**
6. Raw data provided in our archive: **No**
    - Reason for not providing raw data: Data are too large
        - Raw data are securely, and publicly stored under the above DOI (Zenodo archive).
        - Detailed description of access modalities and data processing:
            (a) These data are used for classifying whether a given cell is developable or not.
            (b) Download the data the above DOI.
            (c) Open the file in a GIS application (ArcGis or QGIS). Extract raster value for grid cell.
            (d) If the centroid of our 250m x 250m cell lies in a water body, it is classified as undevelopable.
            (e) In data item [#5], we provide a description of the location for a script to automate this process for cells of all city grids.
7. Transformation of data are provided: **Yes**
    - Together, with data item [#4] (slope), this information is used to identify developable/undevelopable cells.
8. Format: `.csv`
9. List availability in the package:
    - `PrimeLocationsReplication/_Data/USMETRODATA/GISData/USMETROGRIDS`. In this folder, there is .csv file for each US metro (containing all grid cells).
10. Variable dictionary: Variable name 'developable' in .csv file.
11. Additional licensing information (if applicable): Data are distributed under a Creative Commons Attribution 4.0 International license by the originator (https://doi.org/10.5281/zenodo.7254221).

[#4] Global Multi-resolution Terrain Elevation Data 2010

1. Dataset full name: Global Multi-resolution Terrain Elevation Data 2010 (7.5 arc second resolution, statistic: mean)
2. Originator (dataset citation): Earth Resources Observation And Science Center (2017) , Danielson and Gesch (2011)
3. Full DOI/URL: https://doi.org/10.5066/F7J38R2N (data description) https://edcintl.cr.usgs.gov/downloads/sciweb1/shared/topo/downloads/GMTED/Grid_ZipFiles/mn75_grd.zip (data)
4. Raw data provided in public deposit other than ours: **Yes**, please click link above to download the data
5. Data provided in our public deposit: **Yes**, in a transformed version.
6. Raw data provided: **No**
    - Reason for not providing raw data: Data are too large
    - Data are stored by Earth Resources Observation And Science Center (2017), part of the US Geological Survey. We have also keep a copy of the data on a local server for the next 5 years.
    - Detailed access modality:
        - Click on above link
        - Data can be opened with GIS application.
7. Transformation of data are provided: **Yes**
    - Together, with data item [#3] (land use), this information is used to identify developable/undevelopable cells.
    - Details on transformations:
        - Open data with ARCGIS Pro.
        - Use 'slope' function to generate a slope raster.
        - Extract values to the grids (data item [#5]). In data item [#5], we provide a description of the location for a script to automate this process for cells of all city grids.
8. Format: `.csv`
9. List availability in the package:
    - `PrimeLocationsReplication/_Data/USMETRODATA/GISData/USMETROGRIDS`. In this folder, there is .csv file for each US metro (containing all grid cells).
10. Variable dictionary: Variable name 'developable' in .csv file.
11. Additional licensing information (if applicable): According to the University of Manitoba, there are no access and use restrictions (https://canwinmap.ad.umanitoba.ca/documents/270#more). We provide the full text in the `Licenses` folder.

[#5] Definition of MSA grids

1. Dataset full name: Prime locations dataset (**Data generated for this paper**)
2. Originator (dataset citation): Ahlfeldt, Albers and Behrens (2025*b*)
3. Full DOI: https://doi.org/10.7910/DVN/OCYVYN
4. Raw data provided in public deposit/website other than ours: **No**
5. Data provided in our public deposit: **Yes**
6. Raw data provided in our public deposit: **Yes**, we created these data. We compiled the grids in the following way, mainly using ArcGIS and its python console:
   (a) We identify the zip code with the greatest employment density within five kilometres around its centroid within each MSA according to the County Business Pattern Data of 2015 (United States Census Bureau, 2017), see also data item [#12]
   (b) We then put a 70 x 70km grid (with 280 x280 cells of 250 x 250 meters) around this centroid.
   (c) We intersect this with the official MSA delineation of 2014 (data item [#2])
   (d) Deviations from strict rule and manual adjustment: Producing sensible grids requires manual checking in all cases and adjustments in some cases - a process that cannot be fully automated:
      i. In a very few cases, official MSA delineations cut through continuous metro areas because of the presence of a state border. In this case, we ignore the official metro delineation.
      ii. Our criterion sometimes leads to the exclusion of cities/counties that appear in the MSA name. In response,
         - we move the centroid manually to cover all localities in a metro's name
         - where necessary, because the Metro is too large for a 70kmx70km grid (e.g. Dallas-Fort Worth), we enlarge the grid.
      iii. For each cell in each grid, we determine if a cell is developable according to the following rules:
         - Cells that are in permanent water bodies are not developable (land cover data: see data item [#3])
         - Cells that are in glaciers are not developable (land cover data: see data item [#3])
         - Cells that have a slope steeper than 20% are not developable (for data on slopes, see [#4])
         - If a cell fulfils one of the above criteria but is classified as built in the landcover data, we classify it as built (land cover data: see data item [#3])
   (e) We save the results in .csv files for each city, with a definition of whether cells are developable or not. These data provided in the replication package.
   (f) To ease the replication of creating the grids, we provide the following material:
      - List of centroids for the cities.
      - The shapefile with the official delineation (U.S. Census Bureau, 2014).
      - A python script that can be executed in ArcGis. The paths have to be adapted to the user's needs.
7. Transformation of data are provided: **Not applicable**
8. Format: `.csv`
9. List availability in the package:

- Grid cell data: `PrimeLocationsReplication/_Data/USMETRODATA/GISData/USMETROGRIDS`. In this folder, there is .csv file for each US metro (containing all grid cells).
- Shapefiles with grid central points: `PrimeLocationsReplication/_Data/USMETRODATA/GISData/USMETROGRIDS/GRIDCREATIONDATAANDSCRIPTS/SHAPEFILEWITHCENTRALPOINTS` (Special/large grid and normal cases)
- Script: `PrimeLocationsReplication/_Data/USMETRODATA/GISData/USMETROGRIDS/GRIDCREATIONDATAANDSCRIPTS/SCRIPTS` (Special/large grid and normal cases)

10. Variable dictionary: see variable names in .csv file.

Table 5: Dictionary for NETs data files

v3	cell row identifier
v4	cell col identifier
v6	cell UL longitude (in radians)
v7	cell UL latitude (in radians)
v8	cell LR longitude (in radians)
v9	cell LR latitude (in radians)
v11	employment (manufacturing and wholesale)
v12	employment (non-traded services)
v13	employment (tradable services)
v14	employment (public services)
v15	employment (other)
v17	total employment
v19	ClusterID (0 = no cluster)
v21	Unique cell ID (string)
v23	Undevelopable (1= not developable)

11. Additional licensing information (if applicable): Not applicable - we created these data.

[#6] Adjusted height dataset

1. Dataset full name: Adjusted height dataset
2. Originator (dataset citation): Barr and Jedwab (2023)
3. Full DOI: As Barr and Jedwab (2023) do not provide a replication directory, permission was obtained from Barr and Jedwab (2023) to add the raw data by to this repository: https://doi.org/10.7910/DVN/OCYVYN
4. Raw data provided in public deposit/website other than ours: **No**
5. Data provided in our public deposit: **Yes, in raw and transformed version**
6. Raw data provided: **Yes**
7. Transformation of data are provided: **Yes**

   A spatial join was required. It was performed manually using ArcGIS, see `PrimeLocationsReplication/_Data/USMETRODATA/HeightGap/RAW` for a description.
8. Format: `.dta`
9. List availability in the package:
   - `PrimeLocationsReplication/_Data/USMETRODATA/HeightGap/RAW/cbsa-AdjHeightGIS`
10. Variable dictionary: see labels in stata .dta file.
11. Additional licensing information (if applicable): We obtained written permission by Remi Jedwab to redistribute these data via our replication dataset.

[#7] Zillow CBSA house price index

1. Dataset full name: Zillow ZHVI Singe-Family Homes Time Series
2. Originator (dataset citation): Zillow Research (2024)
3. Full URL: https://www.zillow.com/research/data/ (data accessed October 15, 2024)
4. Raw data provided in public deposit/website other than ours: **Yes** (Zillow website above)
    - Click on URL above.
    - Choose under Home values:
        - Data value: 'ZHVI Singe-Family Homes Time Series'
        - Geography: 'Metro & US'
        - Click download.
        - Choose all month of year 2015 and take by metro and average.
        - save as csv file
5. Data provided in our public deposit: **Yes)**
6. Raw data provided: **Yes, in slightly transformed version (annual average)**
7. Transformation of data are provided: **Yes**, data are merged to the CBSA identifiers, the procedure is described in detail in `PrimeLocationsReplication/_Data/USMETRODATA/Housing/RAW`.
8. Format: `.dta` (matched data), .csv (raw data)
9. List availability in the package:
    - Matched data: `PrimeLocationsReplication/_Data/USMETRODATA/Housing/cbsa_hpi`
    - Raw data: `PrimeLocationsReplication/_Data/USMETRODATA/Housing/RAW/Metro_zhvi_uc_sfr_tier_0.33_0.67_sm_sa_month.csv`
10. Variable dictionary: see labels in stata .dta file.
11. Additional licensing information (if applicable): We have obtained the written permission by Edward Berchick and Zillow legal to republish the raw data as an annual average.

[#8] Wharton Residential Land Use Regulatory Index (WRLURI)

1. Dataset full name: Wharton Residential Land Use Regulatory Index (WRLURI)
2. Originator (dataset citation): Gyourko, Hartley and Krimmel (2021), Wharton School (2020)
3. Full URL: https://real-faculty.wharton.upenn.edu/wp-content/uploads/~gyourko/WRLURI/WHARTONLANDREGULATIONDATA_1_15_2020.zip (last accessed Feb 18, 2025)
4. Raw data provided in public deposit/website other than ours: **Yes** (Wharton website)
    - Click on URL above.
    - Download starts.
5. Data provided in our public deposit: **Yes, in raw and transformed version**
6. Raw data provided: **Yes**
7. Transformation of data are provided: **Yes**, data are merged to the CBSA identifiers/CBSA identifier variable is renamed. The procedure is described in detail in `PrimeLocationsReplication/_Data/USMETRODATA/LandUseRegulation/RAW`.
8. Format: `.dta`
9. List availability in the package:
    - `PrimeLocationsReplication/_Data/USMETRODATA/LandUseRegulation/WRLURI2019_CBSA`
10. Variable dictionary: see labels in stata .dta file.
11. Additional licensing information (if applicable): Permission to republish the raw data was obtained via email from the lead author (Joe Gyourko).

[#9] Historical manufacturing data for 1940

1. Dataset full name: 1940 Census - Population, Housing, Agriculture & Economic Data; NHGIS code: 1940_cPHAE
2. Originator (dataset citation): Manson et al. (2024*a*)
3. Full DOI: http://doi.org/10.18128/D050.V19.0 (last accessed Feb 18, 2025)
4. Raw data provided in public deposit/website other than ours: **Yes** (IPUMS National Historical Geographic Information System (NHGIS)); access in the following way:
    - Click on DOI given above.
    - Click 'get data'. (NGHIS only provides the most recent versions via their website. In the case of future revisions of the data and if researchers wish to compare our provided series with the vintage series, this can be obtained by contacting IPUMS via ipums@umn.edu).
    - Click datasets.
    - Choose '1940 Census: Population, Housing, Agriculture & Economic Data [US, States & Counties]'
    - Click 'submit'.
    - Put the following variables in the 'data cart':
        - Manufacturing Establishments, 1939 (Source code: NT18)
        - Value Added by Manufacture, 1939 (Source code: NT23)
        - Retail, Wholesale and Service Establishments (Source code: NT32)
    - Press continue on data cart.
    - Choose geography: 'County (by State)'
    - Choose file format '.csv' and choose
    - Press submit (Login/registration is required, but free).
    - Data can then be downloaded.
5. Data provided in our public deposit: **Yes, in raw and transformed version**
6. Raw data provided: **Yes**
7. Transformation of data are provided: **Yes**, data are merged to the CBSA identifiers/CBSA identifier variable is renamed. The procedure is described in detail in `PrimeLocationsReplication/_Data/USMETRODATA/HistoricMfg/RAW` and the corresponding crosswalk file are provided.[3]
8. Format: `.dta` (transformed data) & .csv (raw)
9. List availability in the package:
    - Transformed data: `PrimeLocationsReplication/_Data/USMETRODATA/HistoricMfg/cbsa_mfgshare1940`
    - Raw data: `PrimeLocationsReplication/_Data/USMETRODATA/HistoricMfg/RAW`
10. Variable dictionary: see labels in stata .dta file.
11. Additional licensing information (if applicable): NGHIS websites states 'You may publish a subset of the data to meet journal requirements for accessing data related to a particular publication.' (https://www.nhgis.org/citation-and-use-nhgis-data, accessed August 22, 2025).

[3]The crosswalks can be accessed at https://github.com/vbehnam/ICPSR_FIPS/blob/master/README.md (ICP-FIPS) and https://www.nber.org/research/data/census-core-based-statistical-area-cbsa-federal-information-processing-series-fips-county-crosswalk (FIPS-CBSA)

[#10] 2013 American Community Survey: 5-Year Data; NHGIS code: 2009_2013_ACS5a

1. Dataset full name: 2013 American Community Survey: 5-Year Data [2009-2013, Block Groups & Larger Areas]; NHGIS code: 2009_2013_ACS5a
2. Originator (dataset citation): Manson et al. (2024*b*)
3. Full DOI: http://doi.org/10.18128/D050.V19.0 (last accessed Feb 25, 2025)
4. Raw data provided in public deposit/website other than ours: **Yes** (IPUMS National Historical Geographic Information System (NHGIS)), access can be obtained by the following steps:
    - Click on DOI given above.
    - Click 'get data'. (NGHIS only provides the most recent versions via their website. In the case of future revisions of the data and if researchers wish to compare our provided series with the vintage series, this can be obtained by contacting IPUMS via ipums@umn.edu).
    - Click datasets.
    - Choose '2013 American Community Survey: 5-Year Data [2009-2013, Block Groups & Larger Areas]'
    - Click 'submit'.
    - Put the following variables in the 'data cart':
        - Total population (Source code: B01003)
        - Median Household Income in the Past 12 Months (in 2013 Inflation-Adjusted Dollars) (Source code: B19013)
        - Aggregate Household Income in the Past 12 Months (in 2013 Inflation-Adjusted Dollars) (Source code: B19025)
        - Median Family Income in the Past 12 Months (in 2013 Inflation-Adjusted Dollars) (Source code: B19113)
        - Per Capita Income in the Past 12 Months (in 2013 Inflation-Adjusted Dollars) (Source code: B19301)
        - Housing Units (Source code: B25001)
        - Median Contract Rent (Source code: B25058)
        - Median Gross Rent (Source code: B25064)
        - Owner-occupied housing units (Source code: B25075)
    - For the geographic coordinates of the block data, choose GIS files and 'Block Group' (Extent: United States).
    - Press continue on data cart.
    - Choose geography: 'Block Group (by State–County–Census Tract)'
    - Choose file format '.csv' and choose
    - Press submit (Login/registration is required, but free).
    - Data can then be downloaded.
    - Export the coordinates and block data from the .shp GIS file as .csv.
5. Data provided in our public deposit: **Yes**
6. Raw data provided: **Yes**
7. Transformation of data are provided: **Yes**, data saved as .dta file. Coordinates saved in .csv file.

8. Format: `.dta` (for data); .csv for coordinate file.
9. List availability in the package:
    - Data: `PrimeLocationsReplication/_Data/USMETRODATA/BlockData/block_group_data.dta`
    - Coordinates: `PrimeLocationsReplication/_Data/USMETRODATA/BlockData/block_group_coordinates.csv`
10. Variable dictionary: see labels in stata .dta file.
11. Additional licensing information (if applicable): NGHIS websites states 'You may publish a subset of the data to meet journal requirements for accessing data related to a particular publication.' (https://www.nhgis.org/citation-and-use-nhgis-data, accessed August 22, 2025).

[#11] Commuting times 2010 American Community Survey; NHGIS code: 2010_ACS1

1. Dataset full name: 2010 American Community Survey; NHGIS code: 2010_ACS1
2. Originator (dataset citation): Manson et al. (2024c)
3. Full DOI: doi: /10.18128/D050.V19.0 (last accessed Feb 25, 2025)
4. Raw data provided in public deposit/website other than ours: **Yes** (IPUMS National Historical Geographic Information System (NHGIS)); access can be obtained by the following steps:
    - Click on DOI given above.
    - Click 'get data'. (NGHIS only provides the most recent versions via their website. In the case of future revisions of the data and if researchers wish to compare our provided series with the vintage series, this can be obtained by contacting IPUMS via ipums@umn.edu).
    - Click datasets.
    - Choose '2010 American Community Survey: 1-Year Data'.
    - Click 'submit'.
    - Put the following variables in the 'data cart':
        - Travel Time to Work (Source code: B08303)
    - Press continue on data cart.
    - Choose geography: 'Metropolitan Statistical Area/Micropolitan Statistical Area'
    - Choose file format '.csv' and choose
    - Press submit (Login/registration is required, but free).
    - Data can then be downloaded.
5. Data provided in our public deposit: **Yes**
6. Raw data provided: **Yes**
7. Transformation of data are provided: **Yes**, data has been saved as .dta file.
8. Format: `.dta` (for data)
9. List availability in the package:
    - `PrimeLocationsReplication/_Data/USMETRODATA/BlockData/block_group_data.dta`
10. Variable dictionary: see labels in stata .dta file.
11. Additional licensing information (if applicable):NGHIS websites states 'You may publish a subset of the data to meet journal requirements for accessing data related to a particular publication.' (https://www.nhgis.org/citation-and-use-nhgis-data, accessed August 22, 2025).

[#12] County Business Patterns: 2014 & 2015

1. Dataset full name: County Business Patterns: 2014 & 2015
2. Originator (dataset citation): United States Census Bureau (2016), United States Census Bureau (2017)
3. Full URL: https://www.census.gov/data/datasets/2014/econ/cbp/2014-cbp.html (2014 version; last accessed Feb 25, 2025), https://www.census.gov/data/datasets/2015/econ/cbp/2015-cbp.html (2015 version; last accessed Feb 25, 2025)
4. Raw data provided in public deposit/website other than ours: **Yes** (Census Bureau),
   - Click on the URL given above.
   - Click on button 'ZIP Code Totals File'.
   - Download zip file (name:'zbp14totals.txt').
   - Process for the 2015 edition is analogous.
5. Data provided in our public deposit: **Yes**
6. Raw data provided: **Yes**
7. Transformation of data are provided: **Yes**, we provide a cross-walk file and merge the ZIP codes to ZCTAs (ZIP Code Tabulation Areas). The cross-walk was originally by John Snow, Inc. and is distributed by via github (https://github.com/jjchern/zipzcta)
8. Format: `.dta` (for data)
9. List availability in the package:
   - Matched data: `PrimeLocationsReplication/_Data/USMETRODATA/RawNumericData/CBP_ZCTAFILES/ZCTA_EMP.dta`
   - Raw data and cross-walks: `PrimeLocationsReplication/_Data/USMETRODATA/RawNumericData/CBP_ZCTAFILES/RAW`
10. Variable dictionary: see labels in stata .dta file.
11. Additional licensing information (if applicable): Licensing information is not directly provided on the sources website. However, the data are also provided by https://catalog.data.gov/dataset/2014-county-business-patterns with a Creative Commons CCZero license. A screenshot of this licensing information is provided in the `Licenses` folder.

[#13] Location of local tradable service establishments

1. Dataset full name: Prime locations dataset (**Data generated for this paper**)
2. Originator (dataset citation): Ahlfeldt, Albers and Behrens (2025*b*)
3. Full DOI: https://doi.org/10.7910/DVN/0CYVYN (dataverse deposit for this paper)
4. Raw data provided in public deposit/website other than ours: **No**
5. Data provided in our public deposit: **Yes**
6. Raw data provided: **Yes**, data was created the following way:
    - We use Google's Nearby Places API to scrape the coordinates of all Places of Interest (POI); The cost was approximately USD 1,000.
    - Places of interest are tradable service firms. We associated specific keywords with these (Table A.2.1 in the paper's appendix). We then search with the API in the global cities in our sample.
    - While Google's API identifies all POI within a circle with a user-specified centroid and radius (with a maximum of 50 kilometres), it returns at most 60 POI per search query. To collect the universe of POI despite this query restriction, we therefore apply an iterative search strategy for each city-keyword pair:
        - We perform an initial scrape on each city's centroid, using a radius of 50 kilometres. If the query returns less than 60 POI, we stop.
        - If the query returns 60 POI, we perform additional scrapes within 4 circles with 25km radius each, shifting their centroids by 25km in each inter-cardinal direction from the original circle's centroid.
        - We continue to divide these circles into 4 overlapping subcircles in the same manner until the respective query returns less than 60 POI. As each step of our iterative search strategy generates sub-circles that cover the whole area of their parent, we are guaranteed to obtain the universe of all POI within our area of interest matching the specified keywords.
        - Finally, we delete all duplicates.
    - We save the results in a .csv file
    - We provide the corresponding python script to query the google API below. As described in the README in the corresponding folder, this has to be manually configured using the replicator's API key. Please contact us if you have any questions. Since business die and new ones emerge, the output from the script will change over time (data was collected in January 2020), the API will not return the exact same results we use in this paper.
    - We hence archived the data from our query and provide it in this replication directory (see below).
7. Transformation of data are provided: **Yes**, we provide a merged .dta file with the global office locations described in data item **[#13]** (just below).
8. Format: `.dta`
9. List availability in the package:
    - Data: `PrimeLocationsReplication/_Data/125GLOBALCITIES/CLUSTERING/RAW_Data/RAW/primepoints_includingHQs.dta`
    - Script to query google API: `/_Data/125GLOBALCITIES/RAW/TooltoqueryGoogleAPI`.
10. Variable dictionary: see labels in stata .dta file.
11. Additional licensing information (if applicable): Not applicable - we created these data.

[#14] Location of industry leaders, headquarters, central banks, stock exchanges

1. Dataset full name: Prime locations dataset (**Data generated for this paper**)
2. Originator (dataset citation): Ahlfeldt, Albers and Behrens (2025*b*)
3. Full DOI: https://doi.org/10.7910/DVN/0CYVYN (dataverse deposit for this paper)
4. Raw data provided in public deposit/website other than ours: **No**
5. Data provided in our public deposit: **Yes**
6. Raw data provided: **Yes**, data was created the following way:
    - Manual collection: Selection of firms based on manually consulting websites, in particular three types of establishments were chosen:
        (a) Global industry leaders in accounting, consultancy, law, investment banks, and insurances (company names are provided in Appendix Table A.2.3)
        (b) Tradable service companies in global stock markets. The simple criterion is that the company is part of the country's main stock market index. For example, the insurer Munich RE is a tradable services company, is listed in the DAX 40 (the German leading stock market index), and has its HQ in Munich. In total, we geocode the location of 562 listed tradable services companies. To identify all potential stock exchanges, we rely on a list compiled by Meri Paterson (http://www.meripaterson.com).
        (c) Addresses of all central banks and stock markets of the countries in our sample as well as their domestic regional and international representations.
    - We manually collect the coordinates of all the offices/representations of the respective companies/central banks via their websites. Where coordinates are not available, we retrieve addresses instead.
    - When we have only addresses, we convert these with a self-written matlab program. An example of this program can be accessed in the replication package under the path `/_Data/125GLOBALCITIES/CLUSTERING/RAW/ToolToConvertAddresses/Work`. We provide this as an additional service. However, other free programs can be used to transform an address into a coordinate (e.g. QGIS) and their use is recommended.
    - We save all results in a dta file.
7. Transformation of data are provided: **Not applicable**
8. Format: `.dta`
9. List availability in the package:
    - Data: `PrimeLocationsReplication/_Data/125GLOBALCITIES/CLUSTERING/RAW/primepoints_includingHQs.dta`
    - Data including company names: `PrimeLocationsReplication/_Data/125GLOBALCITIES/CLUSTERING/RAW/primepoints_global_CompanyNames`
10. Variable dictionary: see labels in stata .dta file.
11. Additional licensing information (if applicable): Not applicable - we created these data.

[#15] Landcover (for land developability classification)

1. Dataset Full Name: Global land cover classification
2. Originator (dataset citation): Hansen et al. (2000*a*,*b*)
3. Full DOI: Since the original distributor (Global Land Cover Facility (GLCF) at University of Maryland) no longer hosts the file, we provide a copy as part of this replication package (`https://doi.org/10.7910/DVN/0CYVYN`)
4. Raw data provided in public deposit/website other than ours: **Yes**
5. Data provided in our public deposit: **Yes**
6. Raw data provided: **Yes**
7. Transformation of data are provided: **Yes,** land cover information was retrieved manually using ArcGis and added to grid file. Results can be reproduced in the following manner:
   (a) Access the raw land cover data (path given below)
   (b) Open with ArcGis.
   (c) Open the Grid shape files described in data item [#18]
   (d) Extract values to point
   (e) Classify as 'developable' if landcover has the value 13 (built)
8. Format: `.shp` (shapefiles) & .txt (as part of intermediate input) `.tif` (raw data)
9. List availability in the package:
   - Raw data (tif): `PrimeLocationsReplication/_Data/125GLOBALCITIES/GEODATA/LANDCOVERAVHHR/AVHRR_1km_LANDCOVER_1981_1994.GLOBAL.tif`
   - Data (Shapefiles): `PrimeLocationsReplication/_Data/125GLOBALCITIES/GIS/CELLS` (contains one shapefile file per metro ; filename: `CELL_XXX` where `XXX` is the metro id)
   - Data (.txt): Undevelopable classification is part of intermediate output in `PrimeLocationsReplicatio_Data/125GLOBALCITIES/CLUSTERING/cells` (contains one .txt file per metro ; filename: `CELL_XXX` where `XXX` is the metro id)
10. Variable dictionary:
    - Variables in shapefile are named such that they are self-explanatory. Developable cells receive the value '1' for the variable 'devle'.
    - Variables in .txt are abbreviated, see data item [#18] for dictionary.
11. Additional licensing information (if applicable): Data are in the public domain (see license information in `Licenses` folder).

[#16] Water bodies (for land developability classification)

1. Dataset Full Name: Global, high-resolution (30-m) inland water body dataset
2. Originator (dataset citation): Originator:Hansen et al. (2013), Original distributor United States Geological Survey (n.d.).
3. Full DOI: Since the original distributor (Global Land Cover Facility (GLCF) at University of Maryland) no longer hosts the file (the link for the 'datamask' at `https://earthenginepartners.appspot.com/science-2013-global-forest/download_v1.0.html` are not functional) and nor the United States Geological Survey (n.d.) is easily accessible, we provide a copy of the relevant tiles for this replication package (doi: /10.7910/DVN/0CYVYN).
4. Raw data provided in public deposit/website other than ours: **Yes**
5. Data provided in our public deposit: **Yes**
6. Raw data provided: **Yes**
7. Transformation of data are provided: **Yes,** land cover information was retrieved manually using ArcGis and added to grid file. Results can be reproduced in the following manner:
   (a) Access the raw data (path given below)
   (b) Unzip
   (c) Open the tiles with ArcGis.
   (d) Open the Grid shape files described in data item [#18]
   (e) Extract values to point (0 NA; 1 land; 2 water)
   (f) Classify as 'developable' if cell has the value 1 (land)
   (g) Additionally, we calculate the share of water cells by metro grid and save the data as .dta file
8. Format: `.shp` (shapefiles) & .txt (as part of intermediate input) & .dta (city-level share of water cells)
9. List availability in the package:
   - Raw data (tif): `PrimeLocationsReplication/_Data/125GLOBALCITIES/GEODATA/WATERBODIES/water_30m_clip_to_grid.zip` is a zip file with the all relevant tiles in .tif format.
   - Raw data (Shapefiles): `PrimeLocationsReplication/_Data/125GLOBALCITIES/GIS/CELLS` (contains one shapefile file per metro ; filename: `CELL_XXX` where `XXX` is the metro id)
   - Data (.txt): Undevelopable classification is part of intermediate output in `PrimeLocationsReplicatio_Data/125GLOBALCITIES/CLUSTERING/cells` (contains one .txt file per metro ; filename: `CELL_XXX` where `XXX` is the metro id)
   - City-level share of water cells (.dta): `PrimeLocationsReplication/_Data/125GLOBALCITIES/METRO_LEVEL_COVARIATES/metro_covariates.dta`
10. Variable dictionary:
    - Variables in shapefile are named such that they are self-explanatory. Developable cells receive the value '1' for the variable 'devle'.
    - Variables in the .dta file are labeled.
    - Variables in .txt are abbreviated, see data item [#18] for dictionary.
11. Additional licensing information (if applicable): Creative Commons Attribution 4.0 International License.

[#17] Global Multi-resolution Terrain Elevation Data 2010 for global grids

1. Dataset full name: Global Multi-resolution Terrain Elevation Data 2010 (7.5 arc second resolution, statistic: mean)
2. Originator (dataset citation): Earth Resources Observation And Science Center (2017) , Danielson and Gesch (2011)
3. Full DOI/URL: https://doi.org/10.5066/F7J38R2N (data description) https://edcintl.cr.usgs.gov/downloads/sciweb1/shared/topo/downloads/GMTED/Grid_ZipFiles/mn75_grd.zip (data)
4. Raw data provided in public deposit other than ours: **Yes**, please click link above to download the data
5. Data provided in our public deposit: **Yes**, in a transformed version.
6. Raw data provided: **No**
    - Reason for not providing raw data: Data are too large
    - Data are stored by Earth Resources Observation And Science Center (2017), part of the US Geological Survey. We have also keep a copy of the data on a local server for the next 5 years.
    - Detailed access modality:
        - Click on above link.
        - Data can be opened with GIS application.
7. Transformation of data are provided: **Yes**
    - Together, with data items [#15] (land use) and [#16] (water bodies), this information is used to identify developable/undevelopable cells.
    - Details on transformations:
        - Open data with ARCGIS Pro.
        - Use 'slope' function to generate a slope raster.
        - Extract values to the grids (data item [#18]).
        - Additionally, we calculate manually using ArcGis the share of steep slope cells (>8.5 degree) by metro grid and save the data as .dta file.
8. Format: `.shp` (shapefiles) & .txt (as part of intermediate input) & .dta (city-level share of water cells)
9. List availability in the package:
    - Data (Shapefiles): `PrimeLocationsReplication/_Data/125GLOBALCITIES/GIS/CELLS` (contains one shapefile file per metro ; filename: `CELL_XXX` where `XXX` is the metro id)
    - Data (.txt): Undevelopable classification is part of intermediate output in `PrimeLocationsReplicatio_Data/125GLOBALCITIES/CLUSTERING/cells` (contains one .txt file per metro ; filename: `CELL_XXX` where `XXX` is the metro id)
    - City-level share of steep slope cells (.dta): `PrimeLocationsReplication/_Data/125GLOBALCITIES/METRO_LEVEL_COVARIATES/metro_covariates.dta`
10. Variable dictionary: not applicable.
    - Variables in shapefile are named such that they are self-explanatory. Developable cells receive the value '1' for the variable 'devle'.
    - Variables in .txt are abbreviated, see data item [#18] for dictionary.
11. Additional licensing information (if applicable): According to the University of Manitoba, there are no access and use restrictions (https://canwinmap.ad.umanitoba.ca/documents/270#more). We provide the full text in the `Licenses` folder.

[#18] Grid definitions - Global dataset

1. Dataset Full Name: Prime locations dataset - grid definitions (**Data generated for this paper**)
2. Originator (dataset citation): Ahlfeldt, Albers and Behrens (2025*b*)
3. Full DOI: https://doi.org/10.7910/DVN/OCYVYN (dataverse deposit for this paper)
4. Raw data provided in public deposit/website other than ours: **No**
5. Data provided in our public deposit: **Yes**
6. Raw data provided: **Yes**, 'raw grids' are shared. However, to determine the center and extent of them, we require an input that proprietary. Sharing raw data: SNL data (SNL Real Estate Research, 2014, see data item **#23** for detailed description) are proprietary and cannot not be shared in their raw format. Grids were created using the following procedure in ArcGis:
   (a) We start from the SNL database and drop any office buildings that are more than 30 kilometers away from the city-median (based on the other office locations) in either x or y coordinates.
   (b) We then define the extent of the grid with the widest x and y coordinates still found in the SNL data + an additional 5 kilometers buffer around these.
   (c) We create bounding box around these.
   (d) For each city, we then create midpoints for 250 meters grids for the entire city. In ArcGis, the function for creating the grid is called 'Fishnet'.
   (e) For each cell in each grid, we determine if a cell is developable according to the following rules:
      - Cells whose centroid is in permanent water bodies are not developable (See data item [#16] on water bodies)
      - Cells that have a slope steeper than 20% are not developable (for data on slopes, see data item [#17])
      - If a cell fulfils one of the above criteria but is classified as 'built' in the landcover data, we classify it as built (land cover data: see data item [#15])
   (f) We provide the corresponding shapefiles in our replication package.
7. Transformation of data are provided: **Not applicable**
8. Format: `.shp` (shapefiles) & .txt (as part of intermediate input)
9. List availability in the package:
   - Data (Shapefiles): `PrimeLocationsReplication/_Data/125GLOBALCITIES/GIS/CELLS` (contains one shapefile file per metro ; filename: `CELL_XXX` where `XXX` is the metro id)
   - Data (.txt): Undevelopable classification is part of intermediate output in `PrimeLocationsReplicatio_Data/125GLOBALCITIES/CLUSTERING/cells` (contains one .txt file per metro ; filename: `CELL_XXX` where `XXX` is the metro id)
10. Variable dictionary:
    - Variables in shapefile are named such that they are self-explanatory. Developable cells receive the value '1' for the variable 'devle'.
    - Variables in .txt are abbreviated, dictionary:
11. Additional licensing information (if applicable): Not applicable - we created these data.

Table 6: Dictionary for global grid files

v3	cell row identifier
v4	cell col identifier
v6	cell UL longitude (in radians)
v7	cell UL latitude (in radians)
v8	cell LR longitude (in radians)
v9	cell LR latitude (in radians)
v11	employment (manufacturing and wholesale)
v12	employment (non-traded services)
v13	employment (tradable services)
v14	employment (public services)
v15	employment (other)
v17	total employment
v19	ClusterID (0 = no cluster)
v21	Unique cell ID (string)
v23	Undevelopable (1= not developable)

[#19] Tweet density - Business-related tweets

1. Dataset full name: Tweet density - Business-related tweets
2. Originator (dataset citation): We generated these data for this paper based on data we acquired from the social media analytics firm GNIP (2014) in 2014. In particular, we acquired business-related tweets posted in 2014. The firm no longer exists (acquired by twitter in 2014 which is now X).
3. Full DOI/URL: We provide a 'gridded version' of these proprietary data as part of this replication package (`https://doi.org/10.7910/DVN/0CYVYN`). It represents simple counts of tweets matched to our grid cells (no contents or handles can be provided).
4. Raw data provided in public deposit other than ours: **No**
5. Data provided in our public deposit: **Yes**, in a transformed version.
6. Raw data provided: **No**
    - Reason for not providing raw data: Data are/were proprietary in their raw form.
    - Cost for acquiring the raw data: Data was acquired for 1,500 USD in 2014.
    - It is potentially possible to obtain a similar dataset from $X.com$, but we cannot guarantee this.
7. Transformation of data are provided: **Yes**,
    - We match the data to the global city grids (data item [#18]) using ArcGis on the raw data.
    - To anonymize the data completely we aggregate to simple counts.
    - Since our data refers to 2014 it is unlikely that the exact same data can still be obtained from X.
    - We provide a matched `.dta` file that relates the tweet counts to the cells in the global city grids (for around half of the metro areas).
8. Format: `.shp` (shapefiles) & .txt (as part of intermediate input)
9. List availability in the package:
    - Data (dta file): `PrimeLocationsReplication/_Data/125GLOBALCITIES/TWITTER/DATA.dta`
10. Variable dictionary: not applicable (see labels in dta file)
11. Additional licensing information (if applicable): We cannot redistribute the raw data - data has been aggregated to grid cells.

[#20] The Geotaggers' World Atlas

1. Dataset full name: The Geotaggers' World Atlas
2. Originator (dataset citation): Fischer (2016) is the originator of this dataset, and she shared the data with us. We are allowed to use these data for our research. In order to preserve her artwork, we provide a transformed version in this replication directory.
3. Full DOI/URL: We provide a 'gridded version' of these data as part of this replication package (https://doi.org/10.7910/DVN/OCYVYN). A visualization of Erica Fisher's data (Geotagger's world atlas) is available in jpeg format via flickr[4], but those are not the raw data shared with us.
4. Raw data provided in public deposit other than ours: **No**
5. Data provided in our public deposit: **Yes**, in a transformed version.
6. Raw data provided: **No**
    - Reason for not providing raw data: Data underlie the Geotagger's World Atlas by Erica Fisher (enf1234567890@gmail.com), who can be contacted for sharing the raw data.
7. Transformation of data are provided: **Yes**,
    - We match the data to the global city grids (data item [#18]) using ArcGis on the raw data.
    - We provide simple counts by cell using the corresponding ArcGIS function.
    - We provide a matched `.dta` file that relates the photo counts to the cells in the global city grids (for around half of the metro areas).
8. Format: `.shp` (shapefiles) & .txt (as part of intermediate input)
9. List availability in the package:
    - Data (dta file): `PrimeLocationsReplication/_Data/125GLOBALCITIES/PHOTOS/PHOTOCOUNT_250M.dta`
10. Variable dictionary: not applicable (see labels in dta file)
11. Additional licensing information (if applicable): We cannot redistribute the raw data - data has been aggregated to grid cells.

[4] https://www.flickr.com/photos/walkingsf/albums/72157623971287575/

[#21] Building heights data/tall buildings data

1. Dataset full name: Building heights data
2. Originator (dataset citation): Data was bought from the company Emporis and permission was obtained by Ahlfeldt and McMillen (2017) to publish a certain anonymized extract of it. We use the data extract published by Ahlfeldt and McMillen (2017).
3. Full DOI/URL: https://doi.org/10.7910/DVN/FFPAXW
4. Raw data provided in public deposit other than ours: **No**
5. Data provided in our public deposit: **Yes**
6. Raw data provided: **Yes**
7. Transformation of data are provided: **Yes**,
   - Open raw data (point data) by either
     (a) Download replication directory of Ahlfeldt and McMillen (2017) at https://doi.org/10.7910/DVN/FFPAXW and open `DATA/CONSTR_WORLD.dta`
     (b) Open file `CONSTR_WORLD.dta` in this replication directory in `PrimeLocationsReplication/_Data/125GLOBALCITIES/EMPORIS/RAW`
   - Match the data to the global city grids (data item [#18]) using ArcGis on the raw data.
   - Make counts of by cell using the corresponding ArcGIS function for variables `USE_COMMERCIAL` and `USE_RESIDENTIAL` as well as maximum height for buildungs (both all years and conditional on being built after 2000).
   - We provide a matched `.dta` file that provides the corresponding result.
8. Format: `.shp` (shapefiles) & .txt (as part of intermediate input)
9. List availability in the package:
   - Data (dta file):
10. Variable dictionary: not applicable (see labels in dta file)
11. Additional licensing information (if applicable): The license of the Ahlfeldt and McMillen (2017) used in this paper is CC0 1.0.

[#22] Starbucks establishments

1. Dataset full name: Starbucks establishments - All Starbucks Locations In The World
2. Originator (dataset citation): The data were originally scraped by Meller (2021) and stored with an open deposit called Socrata. However, the data have since been removed and are now deposited/distributed by the San Diego Regional Data Library. Since the version history is not entirely clear, we provide the raw data that we had downloaded from Socrata.
3. Full DOI/URL: https://data.sandiegodata.org/dataset/chrismeller-github-com-starbucks/
4. Raw data provided in public deposit other than ours: **Yes**
5. Data provided in our public deposit: **Yes**
6. Raw data provided: **Yes**
7. Transformation of data are provided: **Yes**,
   - We match the data to the global city grids (data item [#18]) using ArcGis on the raw data.
   - We checked the data that are matched to our city grids manually and correct (few) erroneous data points (e.g. missing coordinates) where applicable.
   - We provide simple counts by cell using the corresponding ArcGIS function.
   - We provide a matched `.dta` file that relates the count of starbucks establishments to the cells in the global city grids.
8. Format: `.dta` ('data matched to grid) & `.csv` (raw data) .
9. List availability in the package:
   - Data (dta file): `PrimeLocationsReplication/_Data/125GLOBALCITIES/STARBUCKS_SNL/Pgrid_StarSNL.dta`
   - Data (raw file, current public archive): `PrimeLocationsReplication/_Data/125GLOBALCITIES/STARBUCKS_SNL/RAW/all_starbucks(SanDiegoRegionalDataLibrary).csv`
   - Data (raw file, orginal data downloaded from Socrata): `PrimeLocationsReplication/_Data/125GLOBALCITIES/STARBUCKS_SNL/RAW/All_Starbucks_Locations_in_the_World(SocrataA.csv`
10. Variable dictionary: not applicable (see labels in .dta and column headers in .csv files)
11. Additional licensing information (if applicable): The originator Chris Meller marked this dataset as 'open data' on github (see https://github.com/chrismeller/StarbucksLocations, accessed last August 26,2025). This tag has the following definition: 'Open data is data that can be freely used, re-used, and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike.' Screenshots are used to document this licensing.

[#23] Coworking spaces

1. Dataset full name: Prime locations dataset (**Data generated for this paper**)
2. Originator (dataset citation): Ahlfeldt, Albers and Behrens (2025*b*)
3. Full DOI: https://doi.org/10.7910/DVN/0CYVYN (dataverse deposit for this paper)
4. Raw data provided in public deposit/website other than ours: **No**
5. Data provided in our public deposit: **Yes**
6. Raw data provided: **Yes**, data was created the following way:
    - Manual collection from the websites of two firms for the cities in our global city sample:
        - Regus (https://www.regus.com/):
            * Choose coworking search
            * Choose city in sample
            * Press search
            * Extract address
            * Parse into maps.google.com to get latitude/longitude.
        - WeWork https://www.wework.com/:
            * Choose city in sample
            * Press search
            * Extract adress
            * Parse into maps.google.com to get latitude/longitude.
        - Total establishments collected this way in January 2020:
            * Regus: 2,253
            * WeWork: 742
        - We save the data in a `.dta file` file called `pp_CoWorking`
        - We match these to our global city grids using ArcGIS and provide counts per cell.
7. Transformation of data are provided: **Yes**, matched to global city grids.
8. Format: `.dta`
9. List availability in the package:
    - Data raw: `PrimeLocationsReplication/_Data/125GLOBALCITIES/COWORKING/RAW/pp_CoWorking.dta`
    - Data grid: `PrimeLocationsReplication/_Data/125GLOBALCITIES/COWORKING/RAW/GRID_COWORING.dta`
10. Variable dictionary: see labels in stata .dta file.
11. Additional licensing information (if applicable): Not applicable - we created these data.

[#24] Commercial office stock - Real estate investment trust (REITS) investments of 2014

1. Dataset full name: Real estate investment trust (REITS) investments of 2014
2. Originator (dataset citation): The data were originally bought from SNL Real Estate Research (2014), which was acquired in 2015 by McGraw Hill (now called S&P Global Inc). In principal, the data can still be bought as the SNL database is now maintained by S&P Global Inc. The dataset's name now is 'SNL Real Estate Property dataset'. For price, license, and acquisition details, send an email to `support.datafeed.mi@spglobal.com`
3. Full DOI/URL: `https://www.marketplace.spglobal.com/en/datasets/snl-real-estate-property-(42)`
4. Raw data provided in public deposit other than ours: **No**
5. Data provided in our public deposit: **Yes**
6. Raw data provided: **No**
    - Reason for not providing raw data: Data are proprietary.
    - Cost for acquiring raw data: The data was obtained under an institutional license at LSE.
7. Transformation of data are provided: **Yes**,
    - We match the data to the global city grids (data item [#18]) using ArcGis on the raw data.
    - We perform simple counts and means by cell using the corresponding ArcGIS functions.
    - We provide a `.dta` file with the cell-level counts and means based on the SNL data.
8. Format: `.dta` ('data matched to grid)
9. List availability in the package:
    - Data (dta file): `PrimeLocationsReplication/_Data/125GLOBALCITIES/STARBUCKS_SNL/Pgrid_StarSNL.dta`
10. Variable dictionary: not applicable (see labels in .dta file)
11. Additional licensing information (if applicable): We cannot redistribute the raw data - data has been aggregated to grid cells.

[#25] Population grid - Gridded Population of the World Dataset Version 4.10

1. Dataset full name: Gridded Population of the World Dataset Version 4.10
2. Originator (dataset citation): NASA Socioeconomic Data and Applications Center (SEDAC)
3. Full URL: Since the acceptance of our paper, the NASA Socioeconomic Data and Applications Center (SEDAC) has ended the hosting with NASA. Because of this and since the Gridded Population of the World Dataset Version 4.10 is licensed under the terms of a Creative Commons Attribution 4.0 International License, we deposit the raw data in the deposit for this paper https://doi.org/10.7910/DVN/OCYVY (dataverse deposit for this paper).
4. Raw data provided in public deposit/website other than ours: **No**
5. Data provided in our public deposit: **Yes**
6. Raw data provided: **Yes**
7. Transformation of data are provided: **Yes**, data were manually matched to our grid using ArcGis
    - Unzip zip file
    - Open tif in ArcGIS
    - Import Global city grids into ArcGIS (data item [#18])
    - Assign the population density of a population grid cell to the centroid of a Global City grid cell (based the population grid cell a Global City grid cell centroid falls in).
    - Export and save as .dta and label data accordingly
8. Format: `.dta` (for matched data) ; `.zip`/`.tif` for raw data
9. List availability in the package:
    - Matched data: `PrimeLocationsReplication/_Data/125GLOBALCITIES/GEODATA/POP/GRID_POP.dta`
    - Raw data: `PrimeLocationsReplication/_Data/125GLOBALCITIES/GEODATA/POP/RAW/gpw-v4-population-count-rev10_2010_30_sec_tif.zip`
10. Variable dictionary: see labels in stata .dta file.
11. Additional licensing information (if applicable): data are licensed with the Creative Commons Attribution 4.0 International License

[#26] Metro population

1. Dataset full name: World Urbanization Prospects: The 2014 Revision (except for 4 Swedish cities)
2. Originator (dataset citation): United Nations Population Division (2014) (except for Swedish cities), Statistics Sweden (2016) for 4 Swedish cities
3. Full URL: `https://population.un.org/wup/assets/Download/Archive/WUP2014-Excel-files.zip` (last accessed August 20, 2025)
4. Raw data provided in public deposit/website other than ours: **Yes**, see URL above
5. Data provided in our public deposit: **Yes**
6. Raw data provided: **Yes**
    - The license agreement explicitly prohibits uploading and redistributing the entire raw dataset. We will keep a copy of the raw data. However, we add to our replication package the data on the cities that are both in the United Nations dataset and our global city dataset. In this sense, the raw data are provided. We also add data for four Swedish cities.
7. Transformation of data are provided: **Yes**, data were manually matched to our metros
    - Download zip archive from above URL and unzip
    - `WUP2014_XLS_CD_FILES/WUP2014-F12-Cities_Over_300K.xls`
    - Open list of metropolitan areas analysed in the 'Prime Locations' Paper: `PrimeLocationsReplicatio _Data/125GLOBALCITIES/METRO_LEVEL_COVARIATES/metro_data`
    - Match manually with the corresponding UN metropolitan areas and choose 2015 data
    - We add data for 4 Swedish cities (Helsingborg (province Skane), Linkoeping(province Östergötland), Malmoe (province Skane), Örebroe (province Örebro) from Statistics Sweden (2016, data refer to 2014 instead of 2015)
    - Save data.
8. Format: `.dta` (for matched data) ; `.xls` for raw data
9. List availability in the package:
    - Matched data: `PrimeLocationsReplication/_Data/125GLOBALCITIES/METRO_LEVEL_COVARIATES/metro_covariates.dta`
10. Variable dictionary: see labels in stata .dta file.
11. Additional licensing information (if applicable): Creative Commons Attribution 4.0 International License

[#27] Subway dataset - Subway census (v1)

1. Dataset full name: Subway census (v1)
2. Originator (dataset citation): Gonzalez-Navarro and Turner (2019)
3. Full URL: https://doi.org/10.7910/DVN/VHPRTA (last accessed August 20, 2025).
4. Raw data provided in public deposit/website other than ours: **Yes**
5. Data provided in our public deposit: **Yes**
6. Raw data provided: **Yes**
7. Transformation of data are provided: **Yes**, data were manually matched to our metros
    - Download zip archive from above URL and unzip or accesss it directly in our replication director (`PrimeLocationsReplication/_Data/125GLOBALCITIES/METRO_LEVEL_COVARIATES/SUBWAYCENSUS/RAW/GT_JUE2018_replication.7z`)[5]
    - Open file `replication_file/data/analyze_me2.dta` with Stata
    - Type 'keep if year==2010'
    - Record stations
    - Open and make copy of list of metropolitan areas analysed in the 'Prime Locations' Paper: `PrimeLocationsReplication/_Data/125GLOBALCITIES/METRO_LEVEL_COVARIATES/metro_data`
    - Match our cities manually with the corresponding cities in the Gonzalez-Navarro and Turner (2019). The crosswalk of our and the ids of Gonzalez-Navarro and Turner (2019) is contained in the file `PrimeLocationsReplication/_Data/125GLOBALCITIES/METRO_LEVEL_COVARIATES /SUBWAYCENSUS/CORRECTIONSANDCROSSWALK/PL_METRO_OPENINGS.xlsx`
    - We noticed some mistakes for the opening years and stations as of 2010. These are corrected and listed in the following excel file: `PrimeLocationsReplication/_Data/125GLOBALCITIES/METRO_LEVEL_COVARIATES /SUBWAYCENSUS/CORRECTIONSANDCROSSWALK/PL_METRO_OPENINGS.xlsx`.
    - Save data maintaining the original (`stations_2010_GNT`) (`stations_2010_TA_edit`) data and our corrected as separate variables.
8. Format: `.dta` (for matched data) and `.zip/.dta` (for raw data)
9. List availability in the package:
    - Matched data: `PrimeLocationsReplication/_Data/125GLOBALCITIES/METRO_LEVEL_COVARIATES/metro_covariates.dta`
    - Raw data: `PrimeLocationsReplication/_Data/125GLOBALCITIES/METRO_LEVEL_COVARIATES/SUBWAYCENSUS/RAW/GT_JUE2018_replication.7z`
10. Variable dictionary: see labels in stata .dta file.
11. Additional licensing information (if applicable): data are licensed with CC0 1.0 Universal Deed

[5]Indeed, Gonzalez-Navarro and Turner shared these data privately with us as of 2016 but there seems no difference with the data in the replication package in the dataverse regarding our variables of interest.

3 Dataset list

The following to tables (Table 7 and 8) provide a detailed overview about all datasets used in this paper. In particular, they summarize the following information:

1. Data item number - this is the number to identify the source-specific description from the previous section.
2. Name of the dataset
3. Data Type:
    - Own: dataset create by authors for this paper
    - External (dataset created by others/for other research paper)
4. Data provided:
    - raw
    - transformed
    - Short note on if transformed, why so.
5. License information if raw data are provided
6. Source/dataset citation. The folder `Licenses` in the root folder documents the corresponding licenses/permissions for redistribution whenever raw data is included that was not created for this paper.
7. Location/path in replication package

Table 7: Source Overview US MSA Data

Item	Data: Name	Type: Own	Type: External	Data provided: Raw	Data provided: Transformed	Data provided: Why transformed (matched to grid)	License (if raw data provided)	Source/dataset citation	Location/path
[#1]	National Establishment Time-Series (NETS)		✓		✓	Proprietary data: Republishing of raw (point) data not permitted	NA	Walls & Associates (2024); Ahlfeldt, Albers and Behrens (2025*b*)	PrimeLocationsReplication_Data\USMETRODATA\CLUSTEROUTPUT
[#2]	Tiger/Line Shapefiles (MSA and ZCTA)		✓	✓			Public domain	U.S. Census Bureau (2014)	Metro: PrimeLocationsReplication/_Data/USMETRODATA/GISData/USMETROGRIDS/GRIDCREATIONDATAANDSCRIPTS/OFFICIALDELINIATION/tl_2014_us_cbsa ZCTA: PrimeLocationsReplication/_Data/USMETRODATA/GISData/CountyBusinessPatternBoundaries/tl_2020_us_zcta520
[#3]	ESA WorldCover 10 m 2020 v100		✓		✓	Data are too large	CC BY 4.0	European Space Agency (2022); Zanaga et al. (2022)	PrimeLocationsReplication/_Data/USMETRODATA/GISData/USMETROGRIDS
[#4]	Global Multi-resolution Terrain Elevation Data 2010		✓		✓	Data are too large	Public domain	Earth Resources Observation And Science Center (2017); Danielson and Gesch (2011)	PrimeLocationsReplication/_Data/USMETRODATA/GISData/USMETROGRIDS
[#5]	Definition of MSA grids	✓		✓			CC BY 4.0 (own data)	Ahlfeldt, Albers and Behrens (2025*b*)	PrimeLocationsReplication/_Data/USMETRODATA/GISData/USMETROGRIDS
[#6]	Adjusted height dataset		✓	✓			Permission to redistribute by authors	Originator Barr and Jedwab (2023), distributor: Ahlfeldt, Albers and Behrens (2025*b*)	PrimeLocationsReplication/_Data/USMETRODATA/HeightGap/RAW/cbsa-AdjHeightGIS
[#7]	Zillow ZHVI Singe-Family Homes Time Series		✓	✓				Zillow Research (2024)	PrimeLocationsReplicatio/_Data/USMETRODATA/Housing/cbsa_hpi
[#8]	"Wharton Residential Land Use Regulatory Index (WRLURI) "		✓	✓			Permission to redistribute by authors	Gyourko, Hartley and Krimmel (2021); Wharton School (2020)	PrimeLocationsReplication/_Data/USMETRODATA/LandUseRegulation/WRLURI2019_CBSA
[#9]	Historical manufacturing data for 1940		✓	✓			Permission obtained (see file folder)	Manson et al. (2024*a*)	PrimeLocationsReplication/_Data/USMETRODATA/HistoricMfg/RAW
[#10]	2013 American Community Survey: 5-Year Data; NHGIS code: 2009_2013_ACS5a		✓	✓			Permission obtained (see file folder)	Manson et al. (2024*b*)	PrimeLocationsReplication/_Data/USMETRODATA/BlockData/block_group_data.dta
[#11]	Commuting times 2010 American Community Survey; NHGIS code:2010_ACS		✓	✓			Permission obtained (see file folder)	Manson et al. (2024*c*)	PrimeLocationsReplication/_Data/USMETRODATA/BlockData/block_group_data.dta
[#12]	County Business Patterns: 2014 & 2015		✓	✓			CCZero	United States Census Bureau (2016, 2017)	PrimeLocationsReplication/_Data/USMETRODATA/RawNumericData/CBP_ZCTAFILES/RAW

Table 8: Source Overview Global Cities Data

Item	Data: Name	Type: Own	Type: External	Data provided: Raw	Data provided: Trans-formed	Data provided: Why transformed (matched to grid)	License (if raw data provided)	Source/dataset citation	Location/path
[#13]	Location of local tradable service establishments	✓		✓			CC BY 4.0 (own data)	Ahlfeldt, Albers and Behrens (2025*b*)	PrimeLocationsReplication/_Data/125GLOBALCITIES/CLUSTERING/RAW/primepoints_includingHQs.dta
[#14]	Location of industry leaders, headquarters, central banks, stock exchanges	✓		✓			CC BY 4.0 (own data)	Ahlfeldt, Albers and Behrens (2025*b*)	PrimeLocationsReplication/_Data/125GLOBALCITIES/CLUSTERING/RAW/primepoints_includingHQs.dta
[#15]	Landcover (for land developability classification)		✓	✓			Public domain	Hansen et al. (2000*b*)	PrimeLocationsReplication/_Data/125GLOBALCITIES/GEODATA/LANDCOVERAVHHR/AVHRR_1km_LANDCOVER_1981_1994.GLOBAL.tif
[#16]	Water bodies (for land developability classification)		✓	✓			CC BY 4.0	Hansen et al. (2013),United States Geological Survey (n.d.)	PrimeLocationsReplication/_Data/125GLOBALCITIES/GEODATA/WATERBODIES/water_30m_clip_to_grid.zip
[#17]	Elevation dataset - global cities (for land developability classification)		✓		✓	Data are too large	Public domain	Earth Resources Observation And Science Center (2017) ,Danielson and Gesch (2011)	
[#18]	Grid definitions - Global dataset	✓		✓			CC BY 4.0 (own data)	Ahlfeldt, Albers and Behrens (2025*b*)	PrimeLocationsReplication/_Data/125GLOBALCITIES/METRO_LEVEL_COVARIATES/metro_covariates.dta
[#19]	"Tweet density - Business-related tweets"		✓		✓	Proprietary data: Republishing of raw (point) data not permitted	NA	GNIP (2014); Ahlfeldt, Albers and Behrens (2025*b*)	PrimeLocationsReplication/_Data/125GLOBALCITIES/TWITTER/DATA.dta
[#20]	Geotaggers World Atlas		✓		✓	Permission not given to provide raw (point) data	NA	Fischer (2016); Ahlfeldt, Albers and Behrens (2025*b*)	PrimeLocationsReplication/_Data/125GLOBALCITIES/PHOTOS/PHOTOCOUNT_250M.dta
[#21]	Building heights/tall buildings data		✓	✓			CC 1.0	Ahlfeldt and McMillen (2017)	
[#22]	Starbucks		✓	✓			Permission for redistribution documented	Meller (2021)	PrimeLocationsReplication/_Data/125GLOBALCITIES/STARBUCKS_SNL/RAW/All_Starbucks_Locations_in_the_World(SocrataArchive).csv
[#23]	Coworking spaces	✓		✓			CC BY 4.0 (own data)	Ahlfeldt, Albers and Behrens (2025*b*)	PrimeLocationsReplication/_Data/125GLOBALCITIES/COWORKING/RAW/pp_CoWorking.dta
[#24]	REITS investments of 2014		✓		✓	Proprietary data: Republishing of raw (point) data not permitted	NA	SNL Real Estate Research (2014); Ahlfeldt, Albers and Behrens (2025*b*)	PrimeLocationsReplication/_Data/125GLOBALCITIES/STARBUCKS_SNL/Pgrid_StarSNL.dta
[#25]	Gridded Population of the World Dataset Version 4.10		✓	✓			CC BY 4.0	NASA Socioeconomic Data and Applications Center (SEDAC)	PrimeLocationsReplication/_Data/125GLOBALCITIES/GEODATA/POP/RAW/gpw-v4-population-count-rev10_2010_30_sec_tif.zip
[#26]	Metro population		✓	✓		Copyright on original data, but we combine relevant raw data with other (see description)	NA, CC BY 4.0	United Nations Population Division (2014); Statistics Sweden (2016)	PrimeLocationsReplication/_Data/125GLOBALCITIES/METRO_LEVEL_COVARIATES/metro_covariates.dta
[#27]	Subway census (v1)		✓	✓			CCo 1.0 Universal Deed	Gonzalez-Navarro and Turner (2019)	PrimeLocationsReplication/_Data/125GLOBALCITIES/METRO_LEVEL_COVARIATES/SUBWAYCENSUS/RAW/GT_JUE2018_replication.7z

4 Computational requirements

We structure the README on software, randomness, and computational requirements along three types of tasks:

1. **Clustering: The application of C++ to the raw (proprietary) point pattern NETS data by Walls & Associates (2024, see data item [#1] above).**
2. **Replication of all tables and figures of the paper and main appendix except for one 3D figure.**
3. **Own data collection/transformation of proprietary data/map production (Figure 1; Appendix Figure A.2.1)**

We use sub-subsections corresponding to these tasks.

4.1 Software Requirements

4.1.1 Clustering (C++)

☒ The replication package contains one or more programs. Sources codes for all dependencies are provided in the replication directory.

The complete archive for all C++ codes is in the replication directory subfolder C++. Due to its size and number of files, it is zipped.

The first task that we do in C++ is to extract the data from the huge NETS archives and preprocess it. With more than 87 million establishments split across 33 years, these files are huge (the raw NETS data folder in ASCII format is more than 156 GB in size). The Mac OS executable `extract_NETS` extracts and pre-processes the data. The most complicated part is to reconstruct the geography in the past using the latest geographic coordinates (in one file) and the establishments' move history in another.

The second task we do in C++ is cluster algorithm. Applied to the exhaustive US data it is very demanding in terms of computations. This is especially true for large cities where we need to cluster more than 1 million establishments (New York CBSA has, for example: 1,176,488 establishments; see Table A.1.2), map them to 250 x 250 meters grid cells, and cluster the grid cells. This procedure has to be done repeatedly (for all employment, tradable services, and 'search terms'; and for the different p-values that we use). To operationalize this procedure efficiently, we have developed a C++ package that allows us to run the clustering in an automated way across all 381 US CBSAs and 125 Global Cities. There are two versions of the executables: one command-line version (`clustering_cmd`) to be run in the shell; and one GUI version (`clustering_gui`) that has a graphical interface. The latter requires Mac OS with a recent version of Apple's Metal graphics API.

We provide detailed README files for each of the C++ code archives. While we have taken care to write our code in an ISO standard compatible format (C++20), we do not guarantee that the code runs correctly on any architecture and machine. We compiled and ran the GUI version on a 2021 MacBook Pro M1 with 16GB of ram (with Mac OS 12.6.2 and Xcode 14.2); we did the same for the command line version. We also compiled the command line version on a late 2013 MacPro with 64GB of ram (with Mac OS 12.7.6 and Xcode 14.2). The GUI version does not run on this architecture because of the Metal version.

We have successfully compiled the code on Windows using the mingw32 cross-platform compiler, so this is possible to do. However, cross-platform compilation is always tricky and we cannot guarantee that our code compiles on other cross-platform compilers or any architecture. However, the code should always compile on XCode and Mac OS using the settings included in the `.xcodeproj` files provided with the source codes.

Our code uses a number of dependencies that are clearly indicated in the README files:

- `GNUscientificlibrary(gsl2.8)`

- DearImGUI(v1.89.7)
- Glfw(libglfw3.4)
- stb_image(v2.27)
- tinyfiledialogues(v3.8.8)
- The Apple frameworks: AppKit,QuartzCore,Metal,MetalKit

These dependencies (except for the Apple frameworks) are also provided as parts of the archives. Users familiar with C++ should be able to build the executables (cmd version) on their architecture. For Mac users, the Xcode project files .xcodeproj are provided.

Note that all C++ code, the data, and parametrization files are in a folder called "C++". In the work-flow table in Section 6, all steps that indicate "Software C++" in the last column will be relative to that directory. Detailed README files are provided in each of the executables' subfolder.

4.1.2 Replication of tables/figures (Stata & Python)

☒ The replication package contains one or more programs to install all dependencies and set up the necessary directory structure.

Stata-ado files

The master do file (_Work/Scripts/0_Master_PrimeLocations.do) can be executed in Stata (tested on versions 17 and 18). Our scripts use the following ado files: grc1leg,distinct, texsave,spmap,shp2dta,filelist,geonear,estout,ineqdeco,ftools,reghdfe,rangejoin,rangestat, modern,nnls,texsave. These ado files will be automatically installed when executing the master do file.

Stata-Python integration

Some of the Stata do files called by the master do file contain Python code. These can be executed in Stata (the integration is native from Stata 16 onwards.). However, a Python environment must be set up on the local machine with several packages. We recommend the Anaconda distribution, available at: https://www.anaconda.com/products/distribution. We have used the following version: 3.11.7 (64-bit architecture). There is an excellent introduction on setting up Python in Stata on the Stata blog: https://blog.stata.com/2020/08/18/stata-python-integration-part-1-setting-up-stata-to-use-python/.

We have installed the following dependencies:

- Shapely:2.0.6
- Pandas:2.1.4
- GeoPandas:1.0.1
- NumPy:1.26.4
- tqdm:4.65.0
- SciPy:1.11.4

These dependencies are installed via calling the shell from the Stata Master script (_Work/Scripts/0_Master_PrimeLocations.do). Shall any problems occur, we recommend making use of the environment.yml file we provide in PythonDependencies folder in the root directory. A corresponding tutorial on how to install dependencies via the shell can be found here: https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html#sharing-an-environment

4.1.3 Own data collection/transformation of proprietary data/map production (ArcGis & Matlab)

- We used ArcGis (ArcMap and ArcPro) to match proprietary data to the grids and to create the 3D figures in the left column of Figure 1 as well to display the cities in our sample on a world map (Figure A.2.1).
- Matlab 2018 was used in conjunction with the Google API to reference addresses.

4.2 Controlled Randomness

☒ Random seed 1251368 is set at line 10 of the parametrization file `params_cluster_engine.txt`. It is accessed via the parameter cluster.random_seed in the C++ code.
☐ No Pseudo random generator is used in the analysis described here.

Remark on concurrent execution: We use multiple threads to process the data. To ensure replicability using a single PRNG, the main thread draws a unique series of random numbers which are then dispatched to the worker threads using a threadsafe queue. This ensures that irrespective of the execution order the same random points are processed (in a non-deterministic order) during each program run.

4.3 Memory, Runtime, Storage Requirements

4.3.1 Clustering (C++)

Summary Approximate time needed to reproduce the analyses on a standard 2025 desktop machine:

☐ <10 minutes
☐ 10-60 minutes
☒ 1-2 hours
☐ 2-8 hours
☐ 8-24 hours
☐ 1-3 days
☐ 3-14 days
☐ > 14 days

Approximate storage space needed:

☐ < 25 MBytes
☐ 25 MB - 250 MB
☐ 250 MB - 2 GB
☒ 2 GB - 25 GB
☐ 25 GB - 250 GB
☐ > 250 GB
☐ Not feasible to run on a desktop machine, as described below.

Details The code was run on a MacBook Pro (2021 model) with an M1 Pro processor, 16GB of ram, and a 1TB SSD harddisk. The operating system is Sequoia 15.6.1 (24G90). The code was compiled using XCode 16.2 and the settings provided in the `.xcodeproj` file included with the code in the replication directory. The clustering code runs for about a little more than 1 hour.

4.3.2 Replication of tables/figures

Summary Approximate time needed to reproduce the analyses on a standard 2025 desktop machine:

- ☐ <10 minutes
- ☐ 10-60 minutes
- ☐ 1-2 hours
- ☐ 2-8 hours
- ☐ 8-24 hours
- ☐ 1-3 days
- ☒ 3-14 days
- ☐ > 14 days

Approximate storage space needed:

- ☐ < 25 MBytes
- ☐ 25 MB - 250 MB
- ☐ 250 MB - 2 GB
- ☐ 2 GB - 25 GB
- ☒ 25 GB - 250 GB
- ☐ > 250 GB
- ☐ Not feasible to run on a desktop machine, as described below.

Details The code was last run on a **8-core M3-based laptop with 16GB Memory, running on MacOS version 14.6 with 800GB of free space. The whole replication directory once run takes up approximately 227GB.** We have also successfully run this replication directory on 2 Windows machines.

4.3.3 Own data collection/transformation of proprietary data/map production

We ran ArcGis on a Windows Server (via remote desktop). However, these operations are not very computationally intensive. As these are manual operations that do not require much computing power, we do not provide the estimated time as this will vary by user.

5 Description of programs/code

5.1 Overview

Here, we provides a high-level overview about the code:

5.1.1 Clustering

- First, we provide a program called `extract_NETS` that extracts a cross section (year) from the raw NETS data archives provided by Walls & Associates. We have written this program since when extracting specific years of the NETS data we need to reconstruct the locations of the plants in that year from the last year and the move history of the plants. While this, in princple, can be done in other software (e.g., Stata), the reshaping of the huge data to do this is prohibitively complicated. Our code reads from the raw ASCII files and uses mappings to assemble the data into the correct format in an efficient and fast way. It also splits plants by CBSAs using the CBSAs bounding boxes, thereby generating the raw input files that are used in the clustering algorithm.

- Second, we provide a program called `clustering_cmd` that detects clusters of plants in each CBSA and aggregates those to the grid cell level. A detailed description of how to use the code is provided in the README file `./C++/clustering_cmd/readme.txt`. The code: (i) reads in the parameters from a parametrization file (e.g., input- and output file locations, p-values for clustering, initial value of the random seed etc.); (ii) then loops through the specified list of CBSAs to cluster all employment, tradable services employment, and tradable services employment based on search terms (see documentation for more information). In a first step, the code generates a random baseline distribution of 'excepted' employment distributions around 100K randomly drawn points. Then, in a second step the algorithm determines the observed employment distribution in the neighborhood of each observed establishment. In a third step, the algorithm identifies all plants that have an 'abnormally high' concentration of employment around them and flags them as potentially clustered plants. Last, the algorithm maps these potentially clustered plants to cells and aggregates those cells to contiguous clusters using a recursive function. The output is then a file that contains, for each grid cell, information on: (i) whether the cell belongs to a cluster (and the cluster id); and (ii) what the total employment is in each cell.

5.1.2 Replication of tables and figures

- All programs are stored in `_Scripts`. The master file `0_Master_PrimeLocations` will execute them all.
  - Users can choose to run Python code or not by setting global `runpython_scripts` to either 0 (no Python scripts) or 1 (Python will be called from Stata).
  - If you don't want to run the Pyhton scripts, make sure that you extract the `TEMP.zip` in the `_Work` folder such that the intermediate outputs can be retrieved.
- The Code has three major parts
  - `1_DefinePrograms` defines programs that are used for both the US MSA sample and the global city sample
  - Programs starting with `3_` generally refer to the analysis of the US MSA sample. In section 6.3, we detail which do file produces which table/figure.
  - Programs starting with `4_` provide the link between the US-MSA and global city dataset, for example with regards to estimating weights. In section 6.3, we detail which do file produces which table/figure.
- Programs starting with `5_` provide the preparation and analysis of the global city dataset. In section 6.3, we detail which do file produces which table/figure.
- Programs starting with `6_` provide the results on sub-additivity and overidentification tests for the big data establishments in the global city sample. In section 6.3, we detail which do file produces which table/figure.

5.2 License for Code

The code is licensed under a BSD-3-Clause license (see https://opensource.org/license/BSD-3-Clause for an extended description of this license).

6 Instructions to Replicators

6.1 Overview

As in the previous sections, we split this overview into the description of the Clustering algorithm that requires C++, the 'Replication of figures/tables' that only requires Stata with a Python aggregation, and 'Own data collection/transformation of proprietary data/map production'.

6.1.1 C++ clustering

Below we provide a high-level step-by-step description on how to run the C++ clustering code. To keep the replication directory and this README accessible, we provide more details (including example) in specific README files in the respective folders detailed below

1. Go to `./C++/extract_nets/exec` and run the pre-compiled `extract_NETS` command-line tool (see the file `./C++/extract_nets/readme_extract_NETS` for the syntax. This will create the plant files for each CBSA that are used in the clustering algorithm.
2. Go to `./C++/clustering_cmd/exec` and run the pre-compiled `clustering_cmd` command-line tool (see the file `./C++/clustering_cmd/readme_clustering_cmd` for the syntax. Alternatively, run the `clustering_gui` tool, which does the same thing but has a graphical user interface (GUI). These programs will cluster the plants in each CBSA using the provided parametrization files (see the README files for a complete description).
3. The output of the clustering algorithm is written to `./C++/clustering_cmd/output/cbsa`. The `cells` and `weights` subfolder contain the output for the clusters and the data needed to estimate the employment weights. The subfolders `ecs` and `logs` contain some transformed data and log files that should be of no specific interest to the user.

6.1.2 Replication of figures/tables

- Navigate to `root/_Scripts`
- Open the master file `0_Master_PrimeLocations`. It will reproduce all tables/figures of the main paper except for a 3D and a sample overview (for which ArcGIS is needed, see Section).
- Choose a user in line 22, default is "3".
- Set your root directory, i.e., the path in which you stored the replication directory, for the user you chose (line 31 for user 3)
- Set parameter `runpython_scripts` (line 52); default is 1
  - Users can choose to run Python code or not by setting global `runpython_scripts` to either 0 (no Python scripts) or 1 (Python will be called from Stata).
  - If you don't want to run the Python scripts, make sure that you extract the `TEMP.zip` in the `_Work` folder such that the intermediate outputs can be retrieved.
- Execute `0_Master_PrimeLocations`

6.1.3 A note on the replication of own data collection/ minimal pre-processing/ transformation of proprietary data and large datasets/map production

For the replicators' and other researchers' convenience, we publish all data necessary to replicate all tables and figures. Two figures are manual map exports that could, at the time of the writing of the article, not be produced at the same quality using scripts.

This pertains to the 3D-Maps in **Figure 1** in the main text and **Appendix Figure A.2.1**. **Figure 1** can be generated using ArcScene, the corresponding project is saved in `_Work/GIS3D/...` where '...' refers to the respective MSA identifiers and 'Chicago','LA', and 'NY'. **Appendix**

Figure A.2.1 shows the global sample of cities on a World Map. This map was created with ArcGIS pro. The respective project file is sotred in `_Work/GIS2D/GlobalCitiesMap.mxd`

For proprietary data, we obtained written permission to republish the raw data where the licensing was unclear or prohibited a republishing without permission and modification (e.g. *Zillow* allowed us to publish an annual average of the raw data). However, in two cases, it was not possible to obtain permission since the companies do not exist anymore (data items [#24]; [#19]) or it would violate the terms we agreed to with at the point of the initial data provision (data item [#20]). In these cases, we provide data aggregated at the cell level. For these aggregations, we have used ArcGis Pro. In particular, we match points to the cells provided by our shape files (see data items for the description of the location of the grid files [#5] and [#18]).

Beyond these licensing cases, we also do not provide the raw data for items [#3], [#4], [#17]. The data are simply too large to be included in our replication archive. In these cases, we simply extracted the raster values to the cell centroids using the corresponding ArcGis function. We have also provided the respective details on the data source and transformation in the description of the data sources. If the replicator requires help/an introduction for value to point extraction in ArcGis, please do not hesitate to contact us, especially Thilo Albers (who performed this task).

Finally, we provide data on industry leaders, which we manually collected from websites (data item [#14]). We wrote a little Matlab script to geocode addresses. The tool queries the Google API and provides latitude and longitude. We consider this minimal pre-processing and provide an example script. Integrating this script in the replication workflow did not strike us as reasonable. First, it would require the replicator to register for the google API. Second, because APIs often have request limits, it might lead to crashing the replication code. Hence, in order to avoid unnecessary cost and because there is no added benefit in our view, we decided against including the georeferencing part in the 'master script' in order to simplify the replication workflow.

6.2 Details

We provide a complete walk-through for our replication directory in the table below. As indicated in the last column, we use different programming languages at various stages of the analysis. All exhibits – except for the descriptive 3D maps in Figure 1 (generated using ArcScene) – were generated in Stata.

If you do not have a working python environment or problems to set it up, you can unzip the TEMP folder into the `_Work`. Since the extracted TEMP folder will contain all intermediate outputs, this will allow you to run the script without a functioning Python or C++ environment on a local machine. You can simply proceed to the next script if one script is not executable due to missing software.

However, if you start from scratch – with the content of the TEMP folder *and the intermediate output from clustering in C++*[6] deleted – you must execute all scripts in the table below in the exact order provided for the results to replicate.

The table below also documents additional data outputs not in the paper (such as maps for each metro) that we share via our Prime Locations Toolkit (Ahlfeldt, Albers and Behrens, 2025*a*) via GitHub (`https://github.com/Ahlfeldt/AABPL-toolkit`) as an additional service to other researchers. We created this toolkit upon the request of the editor.

[6]The latter are not deleted by default to facilitate an easier replication. To delete this output, remove the files that are stored in `_Data/USMETRODATA/CLUSTEROUTPUT` and `_Data/125GLOBALCITIES/CLUSTERING/cells`.

Table 9: Code Overview

File	Description	Output	Software
_Work/C++/extract_NETS/exec/extract_NETS(MacOSexecutable); [the source code is in the _Work/C++/extract_NETS/ folder]; the documentation is in the readme_extract_NETS.pdf file in the same folder	Extracts the NETS data in a format that can be used in the clustering algorithm.	[yourspecifiedoutputfolder]/plants_XXXX.txt files, where XXXXX is the CBSA (or global cities) identifier; these are used as inputs in the following stages.	C++
_Work/C++/dofiles/TableA12.do(alternatively,_Scripts/2a_TabNETSdata.do)	"Tabulation of NETS establishments by industry in selected US cities (Chicago, LA, NY). Not called by the master do file since it requires the files plants_16980.txt, plants_31080.txt,plants_35620.txt (obtained after running the extract_NETS tool above), which contain proprietary establishment-level data"	Figures from Table A.1.2 (mfg_wholesale=0,nontradable_services=1,public_services=2,tradable_services=3,others=4)	Stata
_Work/C++/clustering_cmd/exec/clustering_cmd (MAC OS command-line executable), alternative clustering_gui (MAC OS Metal GUI executable); [the source codes are in the _Work/C++/clustering_cmd or _Work/C++/clustering_gui folders]; the documentation is in the readme_clustering_cmd.pdf file in the first folder (the GUI version works in the same way). Note that we provide both a GUI and a command line version. The former requires a recent Mac with a recent Metal API Toolkit. The latter should run on older machines. It needs to be recompiled from source for non-Mac OS platforms. Compatibility is not guaranteed.	Reads in MSA grids, determines whether a grid cell is developable, aggregates establishment-level NETS data to grid cells, and performs lines 2-17 of Clustering Algorithm 1 on the establishment-level data. Saves the output to MSA-specific txt files. Further generates also the total NETS employment and counts of big data establishments within randomly drawn disks for estimating the employment weights.	"_Data/USMETRODATA\CLUSTEROUTPUT/gridcells_XXXXX_YY_output.txt, where XXXXX is the CBSA identifier and YY indicates the establishment employment input into the algorithm (99 =total employment, 5 =base on search terms, 3 =tradable services). Further output: pp_sampling_XXXXX_output.txt in _Data\USMETRODATA\WEIGHTS\weights.	C++
_Work/GIS2D/GlobalCitiesMap.mxd	Project file that generates a 2D world map illustrating the Global Cities sample	Appendix Figure A.2.1	ESRI ArcMap
_Work/GIS3D/16980_Chicago/16980_Chicago.sxd	Project file that generates a 3D profile of employment densities in Chicago	Upper-left 3D map in Figure 1	ESRI ArcScene
_Work/GIS3D/31080_LA/31080_LA.sxd	Project file that generates a 3D profile of employment densities in Los Angeles	Middle-left 3D map in Figure 1	ESRI ArcScene
_Work/GIS3D/35620_NY/35620_NY	Project file for generating a 3D profile of employment densities in New York.	Bottom-left 3D map in Figure 1	ESRI ArcScene
_Scripts/0_Master_PrimeLocations.do	Meta do file that defines the working directory for Stata/Python analysis, including relative paths; generates subdirectories; loads external ado files and internal programs; sets various global parameters and the seed; and calls all other Stata scripts discussed below.	Outputs are generated by scripts called by this meta script.	Stata with Python integration
_Scripts/1_DefinePrograms.do	Defines various Stata programs that are called by the scripts below.	Outputs are generated by scripts that call programs defined in this script.	Stata
_Scripts/3a_0_GenerateLegend.do	Processes MSA grid files to ensure all MSA files contain basic information such as identifiers and coordinates. Appends the data into one large file containing all MSAs.	Intermediate outputs (data files) used by other scripts.	Stata
_Scripts/3a_2_CompileEmploymentGrid_CBSAs_pval.do	Processes clustering datasets generated by C++ code for US MSAs, labels variables, merges undeveloped cells, and saves Stata files for use in subsequent scripts.	Intermediate outputs (data files) used by other scripts.	Stata
_Scripts/3a_3_GeneratePLs.do	Uses the processed output files generated by the previous script (containing the clustering output from Algorithm 1, lines 2-17) to complete the delineation of prime locations in US MSAs (Algorithm 1, lines 18-34).	"Intermediate outputs (data files) used by other scripts. Simple maps for quick inspection (not the final maps in paper and toolkit)"	Stata
_Scripts/3b_1_SummaryStats_PLs_US_A.do	Generates outputs summarizing various attributes of prime locations.	"The following outputs will be generated for all four p-values. The outputs for the preferred p-value will be chosen later. Table 1, except last two columns Appendix Figure B.2.1 Appendix Table B.2.4 Toolkit data set PL-data.csv containing various attributes of all prime locations in all US MSAs"	Stata
_Scripts/3b_1_SummaryStats_PLs_US_B.do	Generates outputs that summarize the geography of prime locations across US MSAs.	"The following outputs will be generated for all four p-values. The outputs for the preferred p-value will be chosen later. Figure 2, panel a Appendix Figure B.2.2 Appendix Table B.2.1 Appendix Table B.2.2 Appendix Table B.2.3 Appendix Table B.2.7 Toolkit data set metro-data.csvcontaining various measures of prime location geography of US MSAs"	Stata
_Scripts/3b_3_CalculateDistancesPLs.do	Generates internal and external distances from prime location boundaries.	Intermediate outputs (data files) used by other scripts.	Stata

Code Overview (continued)

File	Description	Output	Software
_Scripts/3c_3_ValidateThreshold.do	Generates BDD plots showing the change in employment density at the prime location boundary by p-value.	Figure B.1.1	Stata
_Scripts/3c_BaselineResults.do	Selects outputs from the runs with the preferred p-value and copies them to the base folder.	"From the loop over p-values, all outputs with the preferred p-value of 99.5 are copied from temporary folders to final output folders (Figure 2, panel a, Table 1, except last two columns, Appendix Figure B.2.1, Appendix Figure B.2.2, Appendix Table B.2.1 Appendix Table B.2.2, Appendix Table B.2.3, Appendix Table B.2.4, Appendix Table B.2.7, Toolkit data sets metro-data.csv, PL-data.csv)"	Stata
_Scripts/3d_GenShapes4Mapping.do	Generates MSA grid shapefiles with identifiers for mapping purposes.	Intermediate outputs (data files) used by other scripts.	Stata with Python integration
_Scripts/3d_MAPPING.do	Generates maps of prime locations for each US MSA.	"2D maps showing US prime locations in Figure 1, right column Similar 2D maps for all US MSAs published via the toolkit"	Stata
_Scripts/3d_GenGrids4toolkit.do	Generates grid shapefiles with employment data by sector.	Employment grid shapefiles for all US MSAs for distribution via the GitHub toolkit	Stata with Python integration
_Scripts/3e_GenPLshapes.do	Generates shapefiles of prime location boundaries for all US MSAs	Prime location shapefiles for all US MSAs for distribution via the GitHub toolkit	Stata with Python integration
_Scripts/3f_GenBlockGroup2PLdist.do	Generates shapefiles of prime location boundaries for all US MSAs.	Intermediate outputs (data files) used by other scripts.	Stata with Python integration
_Scripts/3g_BlockGroupAnalysis.do	Uses block group data to provide gradient estimates and generate measures of prime location accessibility.	"Intermediate outputs (data files) used by other scripts. Appendix Table B.2.6"	Stata
_Scripts/3h_PLaccessibility.do	Analyzes the effect of polycentric city structure on accessibility.	Appendix Table B.2.5	Stata
_Work/C++/geog_tools/exec/geog_tools (Mac OS executable); [the source code is in the _Work/C++/geog_tools/ folder]; the documentation is in the readme_geog_tools.pdf file in the same folder	Generates employment data within prime location catchment areas using the geographic centroids of the identified prime locations and census block data.	_Data/USMETRODATA/PL_emp20km/PL_emp20km.dta	C++
_Scripts/3h_PLsummaryTable.do	Finalizes descriptive statistics of prime locations using the output from the previous code.	"Final version of Table 1 (adding last two columns) Final version of PL-data.csv for toolkit (adding the last two columns)"	Stata
_Scripts/3i_Validation_CBP_datagen.do	Overlays the ZCTA shapefile with the prime location shapefile to compute the share of area covered by prime locations.	Intermediate outputs (data files) used by other scripts.	Stata with Python integration
_Scripts/3j_Validation_CBP.do	Computes employment densities by ZCTAs based on their degree of overlap with prime locations.	Appendix Figure B.2.3	Stata
_Scripts/4a_0_POP_MSAs.do	Merges metropolitan area population data with US MSAs that overlap with the Global Cities sample.	Intermediate outputs (data files) used by other scripts.	Stata
_Scripts/4a_1_MATCH_PrimePointsToGrid.do	Matches grid identifiers to big data establishments for US MSAs that overlap with the Global Cities sample.	Intermediate outputs (data files) used by other scripts.	Stata
_Scripts/4a_2_CompilingWeights.do	Reads data from pp_sampling_XXXXX_output.txt in _Data\USMETRODATA\WEIGHTS\weights obtained in step 3 from the clustering_cmd C++ code into Stata and prepares it for further analysis.	Intermediate outputs (data files) used by other scripts.	Stata
_Scripts/4a_3_Weights_CBSAs.do	Estimates employment weights to assign to big-data establishments.	"Intermediate outputs (data files) used by other scripts. Appendix Figure B.3.1 Appendix Table B.3.1"	Stata
_Work/C++/dofiles/prepare_prime_points_weighted_US.do	Prepare the employment weighted prime points for the overid on US MSAs	Requires the weights file created before.	Stata
_Work/C++/clustering_cmd/exec/clustering_cmd (MAC OS command-line executable), alternative clustering_gui (MAC OS Metal GUI executable); [the source codes are in the _Work/C++/clustering_cmd or _Work/C++/clustering_gui folders]; the documentation is in the readme_clustering_cmd.pdf file in the first folder (the GUI version works in the same way).	Applies the clustering algorithm to employment-weighted big data establishments to generate clusters (lines 2-17 of Algorithm 1) for US MSAs that overlap with the Global Cities sample, including two sub-samples based on the first and second halves of the sample for validation.	"_Data\USMETRODATA\CLUSTEROUTPUT/cells_validation_first _Data\USMETRODATA\CLUSTEROUTPUT/cells_validation_last"	C++
_Scripts/4b_1_CompileEmploymentGrid_CBSA_validations.do	Processes the clustering dataset generated by the first C++ code, labels variables, merges undeveloped cells, and saves Stata files for use in subsequent scripts.	Intermediate outputs (data files) used by other scripts.	Stata

Code Overview (continued)

File	Description	Output	Software
_Scripts/4b_2_GeneratePLs_val This script is executed three times in a loop over different employment types	Uses the processed output files generated by the previous script (containing the clustering output from Algorithm 1, lines 2-17) to complete the delineation of prime locations based on tradable services and search-term employment (Algorithm 1, lines 18-34).	"Intermediate outputs (data files) used by other scripts. Simple maps for quick inspection (not the final maps in paper and toolkit)"	Stata
_Scripts/4b_3_CompileEWPPfirst.do	Processes the clustering dataset generated by the C++ code using employment-weighted big data establishments (weights from the first half of metros).	Intermediate outputs (data files) used by other scripts.	Stata
_Scripts/4b_4_CompileEWPPlast.do	Processes the clustering dataset generated by the C++ code using employment-weighted big data establishments (weights from the second half of metros).	Intermediate outputs (data files) used by other scripts.	Stata
_Scripts/4b_5_GeneratePLs_val_first.do	Uses the processed output files to complete the delineation of prime locations based on big data establishments and employment weights estimated from the first half of cities (Algorithm 1, lines 18-34).	Intermediate outputs (data files) used by other scripts.	Stata
_Scripts/4b_6_GeneratePLs_val_last.do	Uses the processed output files to complete the delineation of prime locations based on big data establishments and employment weights estimated from the second half of cities (Algorithm 1, lines 18-34).	Intermediate outputs (data files) used by other scripts.	Stata
_Scripts/4c_1_Validation.do	Correlates prime locations from alternative employment measures with prime locations from the preferred employment measure to validate the big data approach.	Appendix Table B.3.2	Stata
_Scripts/5a_0_Establishments.do	Tabulates big data establishments collected for Global Cities sample by establishment type.	"Appendix Table A.2.1 Appendix Table A.2.2 Appendix Table A.2.3"	Stata
_Work/C++/dofiles/prepare_prime_points_weighted_global.do	Prepare the employment weighted prime points for the global cities sample	Requires the weights file created before.	Stata
_Work/C++/clustering_cmd/exec/clustering_cmd (MAC OS command-line executable), alternative clustering_gui (MAC OS Metal GUI executable); [the source codes are in the _Work/C++/clustering_cmd or _Work/C++/clustering_gui folders]; the documentation is in the readme_clustering_cmd.pdf file in the first folder (the GUI version works in the same way).	Applies the clustering algorithm to employment-weighted big data establishments to generate clusters (lines 2-17 of Algorithm 1) for the Global Cities sample	" _Data\USMETRODATA\CLUSTEROUTPUT/cells_validation_{last,first}	C++
_Scripts/5a_1_Compile125cityDataset.do	Processes clustering datasets generated by C++ code for Global Cities, labels variables, merges undeveloped cells, and saves Stata files for use in subsequent scripts.	Intermediate outputs (data files) used by other scripts.	Stata
_Scripts/5a_1_5b_1_DelineatePLs_125cityDataset.do	Uses the processed output files generated by the previous script (containing the clustering output from Algorithm 1, lines 2-17) to complete the delineation of prime locations for Global Cities (Algorithm 1, lines 18-34).	"Intermediate outputs (data files) used by other scripts. Simple maps for quick inspection (not the final maps in the toolkit)"	Stata
_Scripts/5b_2_SummaryStats_125CitiesPLs.do	Summarizes the geography of prime locations in Global Cities.	"Appendix Figure B.3.2 Data sets PL-data.csv and metro-data.csv for toolkit. Text for Table 2 notes."	Stata
_Scripts/5b_3_GenPLShapes.do	Generates shapefiles of prime location boundaries for all Global Cities.	Prime location shapefiles for all Global Cities for distribution via the GitHub toolkit	Stata with Python integration
_Scripts/5b_3_MAPPING.do	Generates maps of prime locations for all Global Cities.	2D maps showing prime locations in for all Global Cities published via the toolkit	Stata
_Scripts/5b_4_GengGridDataShapes.do	Generates grid shapefiles with predicted employment and additional data.	Grid shapefiles for Global Cities for distribution via the GitHub toolkit	Stata with Python integration
_Scripts/5c_1_CalculateDistanceToPLBorder.do	Generates internal and external distances from prime location boundaries.	Intermediate outputs (data files) used by other scripts.	Stata
_Scripts/5c_2_Gradients.do	Generates gradients illustrating how various outcomes change with distance from prime locations within Global Cities.	"Figure 3 Appendix Figure B.4.1"	Stata
_Scripts/6d_1_CorrelatesAndSubadditivity.do	Explores the determinants of urban spatial structure across Global Cities.	Figure 2, panel b Table 2 Appendix Table B.4.1	Stata
_Work/C++/overid_row/exec/overid_row (Mac OS executable); [the source code is in the _Work/C++/overid_row folder]; the documentation is in the readme_overid_row.pdf file in the same folder	Generates counts of coworking establishments, real estate investments, Starbucks franchises, and big data establishments within 100k randomly drawn disks per Global City.	_Data/125GLOBALCITIES/OVERIDdata_overid.dta	C++
_Scripts/6e_1_125CitiesValidation.do	Correlates the counts of establishments within randomly drawn disks generated in the previous step to validate big data establishments.	Appendix Table B.3.3	Stata

6.3 List of tables and programs

The provided code reproduces:

- ☒ All numbers provided in text in the paper (refer to the tables)
- ☒ All tables and figures in the paper
- ☐ Selected tables and figures in the paper, as explained and justified below.

Table 10: Code-Figure/Table Mapping (main paper)

Figure/Table	Program/shapefile	Line number	Output file	Note
Figure 1a	_Work/GIS3D/16980_Chicago/16980_Chicago.sxd	NA	_Work/GIS3D/16980_Chicago/16980_Chicago.png	Open ArcGis project file to generate figure
Figure 1c	_Work/GIS3D/31080_LA/31080_LA.sxd	NA	_Work/GIS3D/31080_LA/31080_LA.png	Open ArcGis project file to generate figure
Figure 1e	_Work/GIS3D/35620_NY/35620_NY	NA	_Work/GIS3D/35620_NY/35620_NY.png	Open ArcGis project file to generate figure
Figure 1b	Work/Scripts/3d_MAPPING.do	126	_Work/Output/Figures/US-CBSAs/Figure_1/MAPS_PL_TotalEmp_BaseP_16980.png	
Figure 1d	Work/Scripts/3d_MAPPING.do	128	_Work/Output/Figures/US-CBSAs/Figure_1/MAPS_PL_TotalEmp_BaseP_35620.png	
Figure 1f	Work/Scripts/3d_MAPPING.do	130	_Work/Output/Figures/US-CBSAs/Figure_1/MAPS_PL_TotalEmp_BaseP_31080.png	
Figure 2a	_Scripts/3b_1_SummaryStats_PLs_US_B.do	136	_Work/Output/Figures/p99_5/FIG_2a_US-CBSA-correlations.pdf	
Figure 2b	_Scripts/Scripts/6d_1_CorrelatesAndSubadditivity.do	68	_Work/Output/Figures/GlobalCities/FIG_2b_GlobalCities-correlations.pdf	
Figure 3	_Scripts/5c_2_Gradients.do	743	_Work/Output/Figures/GlobalCities/FIG_3_1_PLGRADIENTS_NA_ROW.png	
Table 1	_Scripts/3b_1_SummaryStats_PLs_US_B.do	216	_Work/Output/Tables/US-CBSAs/TAB_1_PL_SUMMARY_A.tex	
Table 2 (Panel a & b)	_Scripts/Scripts/6d_1_CorrelatesAndSubadditivity.do	165, 185	_Work/Output/Tables/GlobalCities/TAB_2b_PanelB.tex	
Table 2 (table notes)	_Scripts/5b_2_SummaryStats_125CitiesPLs.do	230	_Work/Output/Data/Footnote.csv	

The provided code reproduces:

☐ All tables and figures in the paper
☒ Selected tables and figures in the paper, as explained and justified below.

- 3 tables describe sources or industry codes.
- 1 table provides summary statistics of the proprietary NETS dataset.

Code-Figure Mapping (appendix)

Figure/Table	Program/shapefile	Line number	Output file	Note
Figure A.2.1	_Work\GIS2D/GlobalCitiesMap.mxd	NA	_Work/GIS2D/WORLD/Fig125.png	Open ArcGis project file to generate figure (Export map as .png)
Figure B.1.1a	_Scripts/3c_3_ValidateThreshold.do	47	_Work/Output/Figures_Appendix/US-CBSAs/FIG_B1_1b_PctCutoff-log-top10.pdf	
Figure B.1.1b	_Scripts/3c_3_ValidateThreshold.do	74	_Work/Output/Figures_Appendix/US-CBSAs/FIG_B1_1b_PctCutoff-log-top10.pdf	
Figure B.2.1	_Scripts/3b_1_SummaryStats_PLs_US_A.do	115	_Work/Output/Figures_Appendix/US-CBSAs/FIG_B2_1_US-PL-histograms.pdf	
Figure B.2.2	_Scripts/3b_1_SummaryStats_PLs_US_B.do	91	_Work/Output/Figures_Appendix/US-CBSAs/FIG_B2_2_US-CBSA-histograms.pdf	
Figure B.2.3	_Scripts/3j_Validation_CBP.do	37	_Work/Output/Figures_Appendix/US-CBSAs/FIG_B2_3_NETSdataValidation.pdf	
Figure B.3.1	_Scripts/4a_3_Weights_CBSAs.do	168	_Work/Output/Figures_Appendix/US-CBSAs/weightsFIG_B3_1_Weights.pdf	
Figure B.3.2	_Scripts/5b_2_SummaryStats_125CitiesPLs.do	151	_Work/Output/Figures_Appendix/US-CBSAs/GlobalCities/FIG_B3_2_distribution.pdf	
Figure B.4.1	_Scripts/5c_2_Gradients.do	389	_Work/Output/Figures_Appendix/US-CBSAs/GlobalCities/FIG_B4_1_PLGRADIENTS_CITYSTRUCTURE.png	
Table A1.1	Not applicable	NA		Table about included industry codes.
Table A1.2	_Scripts/2a_TabNETSdata.do	75,93	NA	Figures created by the do file, but propriatary data input required.

(Table continues on next page)

Table 11: Code-Figure Mapping (appendix, continued)

Figure/Table	Program/shapefile	Line number	Output file	Note
Table A.2.1	_Scripts/5a_0_Establishments.do	24	_Work/Output/Tables_Appendix/validation/GlobalCities/TAB_A2_1_Establishments.tex	
Table A.2.2	_Scripts/5a_0_Establishments.do	43	_Work/Output/Tables_Appendix/validation/GlobalCities/TAB_A2_2_Establishments.tex	
Table A.2.3	_Scripts/5a_0_Establishments.do	71	_Work/Output/Tables_Appendix/validation/GlobalCities/TAB_A2_3_EstablishmentsCompanies.tex	
Table A.2.4	Not applicable	NA (no data)		Data source summary table
Table A.2.5	Not applicable	NA (no data)		Data source summary table
Table B.2.1	_Scripts/3b_1_SummaryStats_PLs_US_B.do	172	_Work/Output/Tables_Appendix/US-CBSAs/TAB_B2_1_HistoricMFGshare.tex	
Table B.2.2	_Scripts/3b_1_SummaryStats_PLs_US_B.do	187	_Work/Output/Tables_Appendix/US-CBSAs/TAB_B2_2_LUR.tex	
Table B.2.3	_Scripts/3b_1_SummaryStats_PLs_US_B.do	202	_Work/Output/Tables_Appendix/US-CBSAs/TAB_B2_3_AdjustedHeight.tex	
Table B.2.4	_Scripts/3b_1_SummaryStats_PLs_US_A.do	155	_Work/Output/Tables_Appendix/US-CBSAs/TAB_B2_4_Specialization.tex	
Table B.2.5	_Scripts/3h_PLaccessibility.do	40	_Work/Output/Tables_Appendix/US-CBSAs/TAB_B2_5_PolycentricityEffects.tex	
Table B.2.6	_Scripts/3g_BlockGroupAnalysis.do	61, 70, 79, 84	_Work/Output/Tables_Appendix/US-CBSAs/TAB_B2_6a_Gradients_A_all.tex	
Table B.2.7	_Scripts/3b_1_SummaryStats_PLs_US_B.do	281	_Work/Output/Tables_Appendix/US-CBSAs/TAB_B2_7_PL_SUMMARY_B.tex	
Table B.3.1	_Scripts/4a_3_Weights_CBSAs.do	178	_Work/Output/Tables_Appendix/weights/TAB_B3_1_Weights-Continuous.tex	
Table B.3.2	_Scripts/4c_1_Validation.do	55	_Work/Output/Tables_Appendix/validation/TAB_B3_2_Validation.tex	
Table B.3.3	_Scripts/6e_1_125CitiesValidation.do	95	_Work/Output/Tables_Appendix/validation/GlobalCities/Table_B3_3_125Cities ValidationCoworkingStarSNL.tex	
Table B.4.1	_Scripts/6d_1_CorrelatesAndSubadditivity.do	92	_Work/Output/Tables_Appendix/validation/GlobalCities/TAB_B4_1_GlobalSummaryStats.tex	

References

Ahlfeldt, Gabriel M., and Daniel McMillen. 2017. "Data and Code for: 'Tall Buildings and Land Values: Height and Construction Cost Elasticities in Chicago, 1870–2010'." Harvard Dataverse [distributor]:https://doi.org/10.7910/DVN/FFPAXW, Version 1.3.

Ahlfeldt, Gabriel M, Thilo Albers, and Kristian Behrens. 2025*a*. "Prime locations toolkit." GitHub[distributor]: https://github.com/Ahlfeldt/AABPL-toolkit (last accessed: Feb 25, 2025), Version 1.0.

Ahlfeldt, Gabriel, Thilo Albers, and Kristian Behrens. 2025*b*. "Data and code for 'Prime Locations'." Harvard Dataverse [distributor]: https://doi.org/10.7910/DVN/0CYVYN, Version 1.0.

Barr, Jason, and Remi Jedwab. 2023. "Exciting, boring, and nonexistent skylines: Vertical building gaps in global perspective." *Real Estate Economics*, 51(6): 1512–1546.

Danielson, Jeffrey J, and Dean B Gesch. 2011. "Global multi-resolution terrain elevation data 2010 (GMTED2010)." United States Geological Service [distributor]: https://edcintl.cr.usgs.gov/downloads/sciweb1/shared/topo/downloads/GMTED/Grid_ZipFiles/mn75_grd.zip (last accessed: August 20, 2025).

Earth Resources Observation And Science Center. 2017. "Global Land Cover Characterization (GLCC)." United States Geological Service [distributor]: https://edcintl.cr.usgs.gov/downloads/sciweb1/shared/topo/downloads/GMTED/Grid_ZipFiles/mn75_grd.zip (last accessed: August 20, 2025).

European Space Agency. 2022. "ESA WorldCover 10 m 2021 v100." Zenodo [distributor]: https://doi.org/10.5281/zenodo.5571936, Version 100.

Fischer, Erica. 2016. "The Geotaggers' World Atlas." Cell level aggregates generated by Gabriel Ahlfeldt and provided via Harvard Dataverse[distributor]: https://doi.org/10.7910/DVN/0CYVYN.

GNIP. 2014. "Business-related Tweets." Data acquired as of 2014 from GNIP (firm no longer exists). Cell level aggregates generated by Gabriel Ahlfeldt and provided via Harvard Dataverse [distributor]: https://doi.org/10.7910/DVN/0CYVYN.

Gonzalez-Navarro, Marco, and Matthew A. Turner. 2018. "Subways and urban growth: Evidence from earth." *Journal of Urban Economics*, 108: 85–106.

Gonzalez-Navarro, Marco, and Matthew A. Turner. 2019. "Code and Data for 'Subways and Urban Growth: Evidence From Earth'." Harvard Dataverse [distributor]: https://doi.org/10.7910/DVN/VHPRTA.

Gyourko, Joseph, Jonathan S. Hartley, and Jacob Krimmel. 2021. "The local residential land use regulatory environment across U.S. housing markets: Evidence from a new Wharton index." *Journal of Urban Economics*, 124: 103337.

Hansen, Matthew C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and John RG Townshend. 2013. "High-Resolution

Global Maps of 21st-Century Forest Cover Change." *Science*, 342(6160): 850–853. Given access difficulties at the original distributor (USGS), copy of relevant data archived by Ahlfeldt, Albers, and Behrens via Harvard Dataverse [distributor]: https://doi.org/10.7910/DVN/0CYVYN.

Hansen, Matthew C, Ruth S DeFries, John RG Townshend, and Rob Sohlberg. 2000*a*. "Data for 'Global land cover classification at 1 km spatial resolution using a classification tree approach'. Global Land Cover Facility (GLCF) at University of Maryland." Copy of data resulting from article and previously distributed by University of Maryland archived by Ahlfeldt, Albers, and Behrens via Harvard Dataverse [distributor]: : https://doi.org/10.7910/DVN/0CYVYN.

Hansen, Matthew C, Ruth S DeFries, John RG Townshend, and Rob Sohlberg. 2000*b*. "Global land cover classification at 1 km spatial resolution using a classification tree approach." *International Journal of Remote Sensing*, 21(6-7): 1331–1364. Copy of data resulting from article and previously distributed by University of Maryland archived by Ahlfeldt, Albers, and Behrens via Harvard Dataverse [distributor]: : https://doi.org/10.7910/DVN/0CYVYN.

Manson, Steven, Jonathan Schroeder, David Van Riper, Katherine Knowles, Tracy Kugler, Finn Roberts, and Steven Ruggles. 2024*a*. "IPUMS National Historical Geographic Information System: Version 19.0. [1940_cPHAE]." Minneapolis, MN: IPUMS.http://doi.org/10.18128/D050.V19.0.

Manson, Steven, Jonathan Schroeder, David Van Riper, Katherine Knowles, Tracy Kugler, Finn Roberts, and Steven Ruggles. 2024*b*. "IPUMS National Historical Geographic Information System: Version 19.0. [2009_2013_ACS5a]." Minneapolis, MN: IPUMS.http://doi.org/10.18128/D050.V19.0.

Manson, Steven, Jonathan Schroeder, David Van Riper, Katherine Knowles, Tracy Kugler, Finn Roberts, and Steven Ruggles. 2024*c*. "IPUMS National Historical Geographic Information System: Version 19.0. [2010_ACS1]." Minneapolis, MN: IPUMS.http://doi.org/10.18128/D050.V19.0.

Meller, Chris. 2021. "Starbucks Locations: A list of Starbucks locations, scraped from the web in 2017." Data provided by San Diego Regional Data Library at https://doi.org/10.5066/F7J38R2N. We provide a copy of the raw data that Chris Meller previously shared via Socrata, which is no longer available: Copy of data archived by Ahlfeldt, Albers, and Behrens via Harvard Dataverse [distributor]: https://doi.org/10.7910/DVN/0CYVYN.

NASA Socioeconomic Data and Applications Center (SEDAC). 2017. "Gridded Population of the World, Version 4 (GPWv4): Basic Demographic Characteristics, Revision 10. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC)." Data was originally distributed via NASA Socioeconomic Data and Applications Center (SEDAC) at https://doi.org/10.7927/H45H7D7F, but is no longer available there. A copy of the data is archived by Ahlfeldt, Albers, and Behrens via Harvard Dataverse [distributor]: https://doi.org/10.7910/DVN/0CYVYN.

SNL Real Estate Research. 2014. "Real estate investment trust (REITS) investments." The data were originally acquired from SNL. The company does no longer exist. It

was acquired by by McGraw Hill (now called S&P). Raw data cannot be shared but cell level aggregates with simple counts have been created by Ahlfeldt, Albers and Behrens and can be accessed via Dataverse[distributor]: https://doi.org/10.7910/DVN/0CYVYN.

Statistics Sweden. 2016. "Folkmängd i riket, län och kommuner 30 september 2014 och befolkningsförändringar 1 juli - 30 september 2014. Totalt, Table 17_kv3_14." Data was originally provided by Statistics Sweden at https://www.scb.se/en/finding-statistics/statistics-by-subject-area/population-and-living-conditions/population-composition-and-development/population-statistics/ (last accessed in Agust 2016). An extract of the relevant data (see description in this README) is merged into to the population data provided by Ahlfeldt, Albers, and Behrens via Harvard Dataverse [distributor]: https://doi.org/10.7910/DVN/0CYVYN.

United Nations Population Division. 2014. "World Urbanization Prospects: The 2014 Revision, CD-ROM Edition." Data was originally distributed via London Data Store at https://londondatastore-upload.s3.amazonaws.com/dataset/global-city-population-estimates/global-city-population-estimates.xls (last accessed Feb 25, 20025). However, it is no longer maintained there. An extract of the relevant data (see description in this README) is merged into to the population data provided by Ahlfeldt, Albers, and Behrens via Harvard Dataverse [distributor]: https://doi.org/10.7910/DVN/0CYVYN.

United States Census Bureau. 2016. "County Business Patterns: 2014." Distributed by the United States Census Bureau at https://www.census.gov/data/datasets/2014/econ/cbp/2014-cbp.html last accessed: Feb 25, 2025); copy of data archived by Ahlfeldt, Albers, and Behrens via Harvard Dataverse [distributor]: https://doi.org/10.7910/DVN/0CYVYN.

United States Census Bureau. 2017. "County Business Patterns: 2015." Distributed by United States Census Bureau at https://www.census.gov/data/datasets/2015/econ/cbp/2015-cbp.html last accessed: Feb 25, 2025); copy of data archived by Ahlfeldt, Albers, and Behrens via Harvard Dataverse [distributor]: https://doi.org/10.7910/DVN/0CYVYN.

United States Geological Survey. n.d.. distributed by the United States Geological Survey; https://lpdaac.usgs.gov/products/gfcc30wcv001/ (last accessed: Feb 25, 2025); copy of data archived by Ahlfeldt, Albers, and Behrens via Harvard Dataverse [distributor]: https://doi.org/10.7910/DVN/0CYVYN.

U.S. Census Bureau. 2014. "2014 TIGER/Line Shapefiles (machine- readable data files)." Provided by U.S. Census Bureau: https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2014&layergroup=Core+Based+Statistical+Areas; Copy of data archived by Ahlfeldt, Albers, and Behrens via Harvard Dataverse [distributor]: https://doi.org/10.7910/DVN/0CYVYN.

Vilhuber, Lars, Marie Connolly, Miklós Koren, Joan Llull, and Peter Morrow. 2022. "A template README for social science replication packages (v1.1)." Social Science Data Editors. https://doi.org/10.5281/zenodo.7293838.

Walls & Associates. 2024. "National Establishment Time-Series Database©." Proprietary data provided by Walls & Associates (see replication package for access modalities and cost); contact: dwalls2@earthlink.net. Cell-level aggregates of data provided by Ahlfeldt, Albers, and Behrens via Harvard Dataverse [distributor]: https://doi.org/10.7910/DVN/0CYVYN.

Wharton School. 2020. "Wharton Residential Land Use Regulatory Index." Wharton School, University of Pennsylvania [distributor:] https://real-faculty.wharton.upenn.edu/wp-content/uploads/~gyourko/WRLURI/WHARTONLANDREGULATIONDATA_1_15_2020.zip (last accessed: Feb 25, 2025); data are also available at: Harvard Dataverse [distributor]: https://doi.org/10.7910/DVN/0CYVYN.

Zanaga, D., R. Van De Kerchove, D. Daems, W. De Keersmaecker, C. Brockmann, G. Kirches, J. Wevers, O. Cartus, M. Santoro, S. Fritz, M. Lesiv, M. Herold, N.E. Tsendbazar, P. Xu, F. Ramoino, and O. Arino. 2022. "European Space Agency WorldCover 10m 2021." Zenodo[distributor]: https://doi.org/10.5281/zenodo.5571936, Version 100.

Zillow Research. 2024. "ZHVI Singe-Family Homes Time Series, January-December 2015." Data distributed via Zillow's official website; https://www.zillow.com/research/data/ (data retrieved Oct 15, 2024); data are also archived at: Harvard Dataverse [distributor]: https://doi.org/10.7910/DVN/0CYVYN.

7 Acknowledgements

We thank the AEA data editor team for their guidance in creating this replication directory.

We thank Remi Jedwab and Jason Barr for sharing the dataset underlying Barr and Jedwab (2023) and allowing us to redistribute it via our directory.

In terms of the data collection, we thank our research assistant team (Katia Del-Frari, Peter Jeffrey, Stephen Law, Sascha Moebius, Alice Neht, Tiantong Song, Timo Stieglitz, and Sevrin Waights) as well as Kalle Kappner for help in creating our novel datasets.

Behrens gratefully acknowledges financial support from SSHRC (Insight Grant 435-2021-1017). Ahlfeldt & Albers gratefully acknowledge the support of the Deutsche Forschungsgemeinschaft (CRC TRR 190–project 280092119).