

Data and replication files for 'Urban growth and its aggregate implications'

by Gilles Duranton and Diego Puga

This page distributes and documents computer programs and data to replicate the results obtained by Gilles Duranton and Diego Puga in their article '**Urban growth and its aggregate implications**', to be published in *Econometrica*.

The replication files

The full replication package is available for download from this site as a zip file: [urbangrowth_replication.zip](#) (18.12 Gb).

This replication package contains all the required code and data except for the restricted-access files with the residence (county) of respondents in the us National Longitudinal Survey of the Youth 1979 (nlsy79) and the restricted-access file with the residence (block group) of households in the 2009 us National Household Travel Survey (nhts).

For researchers not intending to replicate the construction of the geographic variables from raw sources, a smaller replication package is also available for download as a zip file: [urbangrowth_replication_nogis.zip](#) (0.35 Gb). This still replicates all the results, but relies on intermediate data files from our own run of the relevant Python scripts. The only difference with respect to the full replication package is that the very large geographic data sources contained under `data/src/gis/` in the full package are not included (with the exception of block-group and metropolitan area boundaries).

Obtaining the restricted-access location data for the NLSY79 and 2009 NHTS

Fully replicating the results of the published article requires, in addition to the code and data files provided here, access to the restricted-access geocode files with the residence (county) of respondents in the nlsy79 and the restricted-access file with the residence (block group) of households in the 2009 nhts.

See *Using the code without the restricted-access location data for the NLSY79 and 2009 NHTS* below for information on how to run a partial replication without these restricted-access data.

Regarding the restricted-access nlsy79 data, only employees and students of us universities, employees of us federally-funded research centers, and employees of eligible us government institutions and non-profits can request access to the nlsy79 geocode data. The us Bureau of Labor Statistics (bls) has no provisions for accessing the nlsy79 geocode data from outside the United States.

At the time of writing, one can find the application for obtaining access to the geocode nlsy79 data and information about the process at <https://www.bls.gov/nls/geocodeapp.htm>. In the application, the researcher must describe the project's research objectives in a few paragraphs. If the application is approved, the bls will send the researcher a Letter of Agreement to be signed by an official institution signatory. The researcher must sign additional agreements, and in the case of students, their research advisor must be the signatory. Data access agreements are between the bls and the recipient institution, not between bls and individual researchers. All geocode data access must occur on the recipient institution's physical premises.

When we requested access to the data for this research project, the us Bureau of Labor Statistics (bls) sent authorised users of the geocode nlsy79 data a CD-ROM with a series of additional files containing the restricted-access location data for nlsy79 respondents. Shortly before the article's publication, the bls transitioned its mode of provision of nlsy79 geocode data to a virtual data enclave (vde). In this managed environment, researchers can analyze the geocode data. Statistical software available for use in the VDE includes Stata. Researchers can bring external files (such as the replication files for this project) and extract analysis results from the vde, following a bls approval process.

Regarding the restricted-access 2009 nhts data, the Census block group of households surveyed for the 2009 nhts is only provided to researchers by the Federal Highway Administration after explaining why their project involves legitimate research that could not otherwise be accomplished and signing a confidentiality agreement. To initiate the process, researchers should contact the Data User Support person for the nhts (at the time of writing, the contact details are available at <https://nhts.ornl.gov/contactUs.shtml>).

Instructions and overview of the replication files

These are the steps to construct the data and replicate the results of the *Econometrica* article:

- Download and place the uncompressed replication package under some directory that will be the root directory of the replication files.
- Edit `code/_hcgrowth_run.do` to specify in the line beginning with `global PathProjectRoot` for your operating system the path to the root directory of the replication files (note there are three such lines, one for Windows, one for Unix, and one for macOS). The subdirectories `code`, `data`, and `results` will be under this directory on your computer.

- Users without the restricted-access data can skip the next two steps and follow the additional instructions under *Using the code without the restricted-access location data for the NLSY79 and 2009 NHTS* below. Users with these restricted-access data should:
 - Add to the `data/src/nlsy/geocode` directory the four extra files that the BLS provides to authorised users of the restricted-access geocode NLSY79 data: `location.dct`, `location-value-labels.do`, `survey_and_created_variables.dct`, and `survey_and_created_variables-value-labels.do`.
 - Set the flag `global NLSYGeocodeUnavailable = 0` in `code/_hcgrowth_run.do`.
 - Add to the `data/src/nhts/geocode` directory the extra file that the DOT provides to authorised users of the restricted-access 2009 NHTS data: `HCTBG.CSV`.
 - Set the flag `global NHTSBGUnavailable = 0` in `code/_hcgrowth_run.do`.
- If you want to use our pre-processed spatial data, leave the flag `global DisableGIS = 1` in `code/_hcgrowth_run.do` as provided and proceed to the next step. If instead you wish to reprocess these spatial data from the raw sources, you will need to change this flag to `global DisableGIS = 0` and use Python in addition to Stata (see *Software and hardware notes* below for details).
- Run `code/_hcgrowth_run.do` in Stata.

The Stata script `code/_hcgrowth_run.do` first runs `code/1_hcgrowth_builddata.do` to perform the data construction, creating the processed data files used for the analysis (described under *Processed data* below) and placing them in the `data/processed/` directory. Next, the Stata script `code/_hcgrowth_run.do` automatically runs `code/2_hcgrowth_analysis.do` to perform the analysis of the processed data and stores all the results (described under *Results* below) in the `results/` directory.

Figures 1 and 2, which illustrate the theoretical model rather than empirical results, are produced by running Mathematica notebooks `code/analysis/hcgrowth_fig1.nb` and `code/analysis/hcgrowth_fig2.nb` (in the second notebook, specifying first `Year = 1980`; and then `Year = 2010`; in the first line to produce the two panels).

Using the code without the restricted-access location data for the NLSY79 and 2009 NHTS

While it is not possible to fully replicate the results of the *Econometrica* article without the restricted-access location data for the NLSY79 and the 2009 NHTS, researchers can run the code without these additional data under two scenarios.

Researchers who wish to perform a partial replication can produce without the restricted access data all of the figures in the paper, as well as table 1 (except for column 1), table 3, and table C.2. To do this, simply leave the flags `global NLSYGeocodeUnavailable = 1` and `global NHTSBGUnavailable = 1` in `code/_hcgrowth_run.do`, as provided. Replication of table 2 and table C.1 will be skipped.

Researchers without the restricted-access location data for the NLSY79 and the 2009 NHTS that wish to check that the replication code runs smoothly can edit `code/_hcgrowth_run.do` and set the flags `global NLSYGenerateFakeLocations = 1` and `global NHTSGenerateFakeLocations = 1`, while leaving the flags `global NLSYGeocodeUnavailable = 1` and `global NHTSBGUnavailable = 1`. This adjustment will randomly generate a fake location history for each NLSY79 and NHTS respondent, allowing the code to run but generating meaningless results in column (1) of table 1, table 2, and table C.1 with the same format but different values than the actual results in the article.

Software and hardware notes

The results and figures in the *Econometrica* article have been produced in Stata version 18, Python version 3.9.15, and Mathematica version 13.2 using the code and data provided.

The code is highly portable; nevertheless, one should keep in mind the following considerations:

- **Stata:** The following required Stata packages are included in the replication package under the `code/ado/plus/` directory:
 - `estout`: module to make regression tables from stored estimates, by Ben Jann.
 - `grstyle`: module to customize the overall look of graphs, by Ben Jann.
 - `ivreg2`: Stata module for extended instrumental variables/2SLS, GMM and AC/HAC, LIML and k-class regression, by Christopher F. Baum, Mark E. Schaffer, and Steven Stillman.
 - `palettes`: module to provide color palettes, symbol palettes, and line pattern palettes, by Ben Jann.
 - `ranktest`: module to test the rank of a matrix, by Frank Kleibergen, Mark E. Schaffer, and Frank Windmeijer.
 - `regsave`: module to save regression results to a Stata-formatted dataset, by Julian Reif.
 - `shp2dta`: module to converts shape boundary files to Stata datasets, by Kevin Crow.
 - `ftools`: module to provide alternatives to common Stata commands optimized for large datasets, by Sergio Correia.
 - `reghdfe`: module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects, by Sergio Correia.
 - `ivreghdfe`: module for extended instrumental variable regressions with multiple levels of fixed effects, by Sergio Correia.
 - `SJ6-3 sto109`: module for linear partial regression, by Michael Lokshin.

By default, the code temporarily sets Stata's `PLUS` directory to `code/ado/plus/` to ensure that the same versions of these packages used to produce the results in the article are available. Users who prefer to use their standard `PLUS` directory can set the flag `global InstallPackages = 1` in `code/_hcgrowth_run.do`. With this option, the code will instead look for the

required packages in the user's standard **PLUS** directory, installing from the internet any packages that are missing. Note versions installed from the internet may differ from those we used, so using the provided versions by leaving the flag `global InstallPackages = 0` unchanged is recommended.

The code has been tested with Stata versions 18 and 17. The code will impose `version 18` for more robust replicability when run on newer Stata versions.

- **Python:** This is only needed if you wish to reprocess the spatial data from the raw sources instead of using our pre-processed spatial data. The code assumes you have used conda to set up Python. Edit `code/_hcgrowth_run.do` to set the flag `global DisableGIS = 0` and to specify in the line beginning with `global PathConda` for your operating system the path to the conda executable in your computer (note there are three such lines, one for Windows, one for Unix, and one for macOS). The code has been tested with Python version 3.9.15. The following additional Python libraries are required: dask, dask-geopandas, distributed, gdal, geopandas, libspatialindex, matplotlib, numba, rasterio, rasterstats, requests, rioxarray, scikit-learn-intelex, typing, and xarray-spatial. You can install the required Python libraries with the command `conda install -c conda-forge dask dask-geopandas distributed gdal geopandas libspatialindex matplotlib numba rasterio rasterstats requests rioxarray scikit-learn-intelex typing xarray-spatial`. If you wish to use the same exact versions as we did, run `conda install -c conda-forge dask=2023.1.1 dask-geopandas=0.3.0 distributed=2023.1.1 gdal=3.4.3 geopandas=0.12.2 libspatialindex=1.9.3 matplotlib=3.7.1 numba=0.56.4 rasterio=1.2.10 rasterstats=0.18.0 requests=2.28.2 rioxarray=0.14.0 scikit-learn-intelex=2023.0.2 typing=3.10.0.0 xarray-spatial=0.3.5`.
- **Mathematica:** Figures 1 and 2, which illustrate the theoretical model rather than empirical results, are produced with Mathematica notebooks. We used Mathematica version 13.2.
- **Sed:** The sed command is used by `code/analysis/hcgrowth_counterfact_morepermits.do` to manipulate the table it produces. The `sed` command is available as part of any UNIX-based operating system, including Linux and macOS. Windows users can obtain a native port of `sed` from <https://unxutils.sourceforge.net>.
- **Operating system:** None of the Stata or Python code is operating-system-specific and we have run it successfully under Linux, macOS, and Windows.
- **Hardware:** The run of the code producing the results reported in the published version of the article was performed on a DELL PowerEdge R940 Server with two 2.4 GHz 24-Core Intel Xeon Platinum 8260 processors and 1.5 Tb of DDR4-2933MHz RAM running Linux. The run was started on 16 June 2023 and took 16 hours, 33 minutes, and 51 seconds. The data construction took 16 hours, 32 minutes, and 56 seconds, all but 2 minutes and 46 seconds of which was spent running the Python scripts that generate the geographic data. The analysis took 54 seconds. The log of this run is provided with the replication files in `code/logs/log_2023.06.16_19.03.32.txt`.

However, the code will run on more modest hardware. The most demanding part is running the Python scripts that generate the geographic data, and we have run these successfully on a r6i.4xlarge instance from Amazon Elastic Compute Cloud using 16 virtual cpus from a 2.9 GHz Intel Xeon Platinum 8375C processor and 128 Gb of DDR4-3200MHz RAM running Windows, as well as on an Apple iMac Pro with a single 3 GHz 10-Core Intel Xeon W processor and 128 Gb of DDR4-2666MHz RAM running macOS. As already noted, the intermediate data files produced by running these scripts are included with the replication package. Everything else does not have any demanding hardware needs.

Data sources and treatments

City definitions: Our empirical and quantitative analysis focuses on the conterminous United States during the period 1950–2010. To define cities, we use Metropolitan Statistical Area and Consolidated Metropolitan Statistical Area (MSA) definitions outside of New England and New England County Metropolitan Area (NECMA) definitions in New England, as set by the Office of Management and Budget on 30 June 1999. This defines 275 metropolitan areas.

Population: We use county-level population data from the US decennial censuses for 1850, 1920, 1950, 1980, and 2010, that we aggregate to the 1999 MSA/NECMA level. The sources are Schroeder (2016) for 1850 and 1920, Forstall (1996) for 1950 and 1980, and Manson, Schroeder, Riper, Kugler, and Ruggles (2021) for 2010.

City centre and city periphery: We define the city centre as the location indicated by Google Maps for the core city of the metropolitan area.

In addition to defining centres, we need a measure for the spatial extent of the city, corresponding to \bar{x}_{it} in the model. Since, in practice, cities cover two dimensions, there will be different distances between the city periphery and the city centre depending on the direction we follow. When cities have irregular shapes, using the maximum distance to the centre or a very high percentile of the distribution of distances to the centre can be problematic. Also, since metropolitan area definitions are county-based and some urban counties, particularly in the West of the country, extend well into rural areas, a few scattered dwellings very far away from the centre in a county that is part of a city can increase the measured distance between the city periphery and the city centre artificially. To address all of these difficulties, we implement a consistent definition of the city periphery. We take the city periphery to be the longest distance from the

city centre within the metropolitan area boundaries that is within the 95th percentile of dwelling distances and has at least 500 dwelling units per square mile in the 2012 American Community Survey data described below.

Urban fringe: We use the term urban fringe for the area where the city would likely expand next. This is the area where we measure agricultural land prices when calculating replacement costs for housing at the city periphery and where we measure geographical constraints to urban expansion when relating these to the strictness of current planning regulations. The urban fringe is defined as the area within 20km of land developed at medium or high intensity in 2011 that is undeveloped or developed at low intensity, based on land cover data from the 2011 slice of the 2019 version of the National Land Cover Database. The NLCD2019 (Dewitz and us Geological Survey, 2021) offers land cover for years 2001, 2003, 2006, 2008, 2011, 2013, 2016, 2019. We use the 2011 slice of the NLCD2019, since our estimations are all centred around 2010.

Geographical constraints to urban expansion: To obtain an empirical counterpart to our model's z_i , we calculate the share of the area within 30 kilometres of the centre of each city that is geographically unconstrained. This corresponds to $1 / z_i$. We characterise this area with a 30-metre resolution. Each 30-metre cell is classified as geographically unconstrained if it is not covered by slopes steeper than 15%, water, wetlands, or land permanently protected from land cover conversion with a mandate to conserve its natural state, and it does not belong to a foreign country. Slope is calculated on the basis of 1 arc-second Digital Elevation Models from the 3D Elevation Program of the us Geological Survey (2018). Water and wetlands cover is based on the 2011 slice of the NLCD2019. Protected land is identified based on the Protected Areas Database of the United States (us Geological Survey, 2020). This database maps protected areas and assigns them a GAP status code as a measure of intent to permanently protect its natural state. We use GAP status codes 1 and 2 used to isolate land permanently protected from land cover conversion with a mandate to conserve its natural state. As to foreign land, this is identified from the official boundary files of Statistics Canada and Mexico's Instituto Nacional de Estadística y Geografía.

When calculating an empirical counterpart to z_i , we are interested in geographical constraints on the long-term expansion of a city over the course of its history. However, when thinking about geographical constraints to urban expansion as a determinant of the strictness of current planning regulations in panel c of figure 5, it is more appropriate to focus on the area where the city would likely expand next. This is the urban fringe defined above, so we also calculate the share of the urban fringe that is geographically unconstrained for the same set of geographical constraints.

Illustrating the equilibrium with the urban system of the United States: Panels A and B of figure 2 depict the allocation of population across us cities and rural areas in 1980 and 2010 as an equilibrium of the model. To draw this figure, we use parameter values estimated or calibrated in section 5 ($\gamma = 0.07$, $\theta = 0.04$, $\sigma = 0.04$, $\beta = 0.04$, and $\lambda = 0.18$), the actual population in each us metropolitan area and outside metropolitan areas in each year to assign values to N_{it} and N_{rt} , and the share of the area within 30 kilometres of the centre of each city that is geographically unconstrained as the empirical counterpart to $1 / z_i$. We normalise $\tau_t = 1$ in 1980 (which amounts to a choice of numéraire). We set τ_t in 2010 so that population-weighted average growth in y_{it} in the model matches the actual growth in average Gross Domestic Product per person in the United States 1980–2010. For this purpose, we use equation (21) to obtain $\rho^\sigma A_{it}(h_i)^{1+\sigma}$ for each city as a function of τ_t from its values of N_{it} , z_i , and parameters. Substituting this into equation (8) then yields y_{it} for each city as a function of τ_t . We obtain y_{rt} by equating income in rural areas with income in the marginal populated city, where the latter is given by equation (22). We find that increasing τ_t from 1 to 1.569 between 1980 and 2010 makes output per person in the model increase by a factor of 1.658, which matches the ratio of 2010 to 1980 Gross Domestic Product per person in the United States. Note that the numerical computation of τ_t is straightforward since every value of y_{it} is proportional to τ_t (equations 8 and 21).

In the figure, horizontal axis length is total us population, N_t . Total urban population $\sum_i N_{it}$ can be read as the horizontal distance to the left-side axes origin and rural population, $N_{rt} = N_t - \sum_i N_{it}$ can be read as the distance to the right-side axes origin. The thick horizontal segments in the figure represent equilibrium consumption for incumbents in each city, c_{it} (segment height), obtained from equation (40), and population N_{it} (segment length). The thin curves tangent to each thick segment plot consumption for incumbents in each city when population differs from its equilibrium level, as given by equation (41). Incumbents set permitting costs at $p_{it} = c_{it} - c_t$ to achieve the consumption at the maximum of the curve for their city while keeping newcomers indifferent. Rural consumption as a function of rural population is given by the smooth long curve, corresponding to equation (40), where we obtain the value of A_{rt} by equating $c_{rt} = c_t$ for the marginal populated city.

Current Population Survey: Figure 3 plots the evolution of the share of population aged 25–64 who hold a college degree in metropolitan areas of different sizes over the period 1986–2016 in the United States. It uses data from the Annual Social and Economic (ASEC) supplement of the Current Population Survey (cps), obtained from the IPUMS-cps project (Flood, King, Rodgers, Ruggles, and Warren, 2018).

We assign individual observations to a specific metropolitan area based on their county of residence, when available, which we then match to the corresponding 1999 MSA/NECMA; when the county of residence is unavailable, the state of residence is outside of New England, and the cps source data contains the 1999 MSA of residence, we use this; alternatively, we use a purposely-built crosswalk (available with the replication code for this paper) between alternative metropolitan area codes contained in the cps source data and 1999 MSA/NECMA codes. We then group metropolitan areas into three population size categories based on their 2010 population (below 1 million, between 1 and 2.5 million, and above 2.5 million), so that each line in the figure corresponds to the same set of metropolitan areas throughout.

About one-third of the individual observations for residents in metropolitan areas cannot be assigned to a specific area. We assign these observations to the same three size population size categories based first on the metropolitan area size variable and next on the core-based statistical area size variable in the cps. The downside of this procedure relative to be able to assign individual-year observations to a specific metropolitan area is that some observations may be assigned to different curves over time despite corresponding to the same metropolitan area if the population of this area crosses the 1 million or the 2.5 million thresholds.

Up until 1991, the cps contains information on the years of college completed but not on whether the individual has obtained a bachelor's degree, so we classify individuals as having a college degree if they have completed at least 4 years of college. From 1992 onwards, we use the information on whether they have a bachelor's degree or higher. We plot the figure using the ASEC person-level weights.

The cps is also the source for the annual growth rate in the average years of schooling between 1950 and 2010 provided in Section 8. We base the calculation on Table A-1 in us Bureau of the Census (2023) (Years of School Completed by People 25 Years and Over, by Age and Sex: Selected Years 1940 to 2022) and assign to each category of years of school completed the number of years suggested for the United States by De la Fuente and Doménech (2014).

National Household Travel Survey and Census: Column (1) of table 1 estimates γ as the elasticity of distance travelled with respect to the distance between her dwelling and the city centre. Data on household travel behaviour come from the 2009 us National Household Travel Survey (NHTS). The survey is sponsored by various agencies at the us Department of Transportation. For a nationally-representative sample of households, the NHTS provides a travel diary kept by every member of each sampled household where we observe the distance, duration, mode, purpose, and start time for each trip taken on a randomly-assigned travel day. It also includes household and individual demographics.

Household miles travelled are measured using the best estimate of household annual miles computed by the survey administrators, which is their preferred measure. We regress the log of household miles travelled on the log of distance between the household's block-group of residence and the city centre, controls for household and block-group characteristics, and metropolitan area fixed effects. We measure the distance to the centre as the haversine distance between the centroid of each block-group and the centre of each metropolitan area. For consistency with the specifications using housing data, we use all block groups from all metropolitan areas except for college towns, defined as the 46 metropolitan areas with under one million inhabitants in 2010 where at least 10% of them are college students, since the high concentrations of students make housing markets in such college towns very distinct.

The controls for household characteristics, all based on the same NHTS data, are the log of the household size, the log of the number of drivers in the household, the share of drivers that are male, and indicators for a single-person household, for the presence of small children, for the household respondent being Hispanic, White, Black and Asian, and for being a renter.

The controls for block-group characteristics are the percentages of Hispanic, Black, and Asian population (based on 2000 Census data, obtained from the IPUMS-NHGIS project, Manson, Schroeder, Riper, Kugler, and Ruggles, 2021, since the 2009 NHTS records block groups using 2000 boundaries), the performance in standardised tests of the closest public school relative to the city average (from De la Roca, Gould Ellen, and O'Regan, 2014, with variation at the tract level), an indicator for waterfront location (constructed by combining the 2000 block-group boundaries provided in the IPUMS-NHGIS 2000 Census data with the coastline shapefiles from the National Hydrography Dataset and the Great Lakes and watersheds shapefiles from the Great Lakes Restoration Initiative of the us Geological Survey), an indicator for riverfront location (constructed by combining the same block-group boundaries with the major rivers within the United States shapefile included with Esri Data & Maps), and terrain ruggedness (measured by the Terrain Ruggedness Index of Riley, DeGloria, and Elliot, 1999, calculated on the basis of 1 arc-second Digital Elevation Models from the 3d Elevation Program of the us Geological Survey, 2018, and then averaged at the block-group level).

The measure of travel speed in each city included as a control in the regression in column (3) of table 1 and as the dependent variable in column (5) of table 1 is based on the same NHTS data. We keep data on trips in a household vehicle, where this vehicle is a car, van, SUV, or pick-up, and is driven by the survey respondent. Following Couture, Duranton, and Turner (2018), we exclude all trips by households where either the respondent does not recall if they were the driver, or they report one or more trips in top or bottom 0.5% of all trips by distance, time or speed. As they note, removing all trips by the affected household and not only the odd ones is important to avoid biasing the calculations. Since speed varies very substantially depending on trip, individual and household characteristics, we need a minimum number of trips to compute a reliable measure of distance. We restrict our sample to the 182 cities where we have at least 100 trips recorded. We first calculate the speed of individual trips dividing trip miles by trip duration. We then regress the log of travel speed for individual trips on metropolitan area fixed effects, controls for trip characteristics the same controls for household and block-group characteristics as in the regression in column (1) of table 1, and the log of distance between the household's block-group of residence and the city centre. The controls for trip characteristics, all based on the same NHTS data, are the log of trip distance and indicators for day of the week, departure time in 30-minute intervals, and trip purpose. We use the estimated regression coefficients to predict, for each city, the speed of a 15km commuting trip on a Tuesday at 8:00AM by a driver with average characteristics.

To validate the self-reported trip duration estimates of NHTS respondents, we turn to data from Akbar, Couture, Duranton, and Storeygard (2023). They query Google Maps over an extended time period about the duration of a trip with the same origin, destination, day of the week, and departure time as each trip reported by NHTS respondents. Using this alternative trip duration, they recompute an alternative measure of speed that we use in column (6) of table 1.

American Community Survey: All our estimations regarding housing rental prices and values use 5-year 2008–2012 data from the 2012 American Community Survey (ACS), obtained from the IPUMS-NHGIS project (Manson, Schroeder, Riper, Kugler, and Ruggles, 2021). The unit of observation is the block group. We use all block groups between the centre and the periphery of every metropolitan area except for college towns, as defined above, given their distinct housing markets.

All block-group housing regressions use the same controls for housing and block-group characteristics. The controls for housing characteristics are the percentage of dwellings in the block group by type of structure, by number of bedrooms, and by construction decade, all based on the same 2008–2012 ACS data. The controls for block-group characteristics are the same as in the travel regressions, but re-computed for 2012 ACS block groups: the percentages of Hispanic, Black, and Asian population, the performance in standardised tests of the closest public school relative to the city average, an indicator for waterfront location, an indicator for riverfront location, and terrain ruggedness.

Column (2) of table 1 estimates γ based on variation in house prices across locations within a city as a function of distance to the city centre. The dependent variable is the log of the difference between the median rent in the most expensive block group in the city and the median rent in block group under consideration, from the 2008–2012 ACS data. We regress this on the log of the distance between the block group and the centre of its metropolitan area, city fixed effects, and the dwelling and neighbourhood characteristics described above.

Column (3) of table 1 estimates γ based on variation in house prices at the centre of cities as a function of the spatial extent of the city. The first component of the dependent variable for this regression is estimated from an auxiliary regression at the block group level of the log of the median monthly contract rent on city indicators, a third-degree polynomial of distance between the block-group centroid and the city centre, and the aforementioned controls for housing and block-group characteristics. We use this regression to predict the rental price of a national-reference house for city-average neighbourhood characteristics at the centre of each city –i.e. when $x_i^j = 0$. This corresponds to \hat{P}_i on the left-hand side of the empirical specification of equation (35). On the right-hand side of that expression, we have the spatial extent of the city, \bar{x}_{it} , and travel speed $\hat{\tau}_i$. We measure \bar{x}_{it} using the distance between the centre and the periphery of each city as defined above, i.e. the longest distance from the city centre within the metropolitan area boundaries that is within the 95th percentile of dwelling distances and has at least 500 dwelling units per square mile in the 2012 ACS data. Our estimate of speed is the predicted speed of a 15km commuting trip on a Tuesday at 8:00AM by a driver with average characteristics in each city, using NHTS data as described above.

The final component of equation (35) that we need to measure is c_t . Equation (24) tells us this should be proportional to the price of housing at the centre of the cheapest city. Unfortunately, the proportionality constant is itself a function of our key parameter of interest, γ . Since γ appears on both sides of equation (35), we estimate this iteratively. Given a starting value of γ , the values of θ , σ and β obtained below, and the estimated city-centre house price in the cheapest city \hat{P} , we obtain a value for $c(\gamma) = \frac{\gamma + \theta - \sigma - \beta}{(\sigma + \beta)(\gamma + 1)} \hat{P}$. This value allows us to compute our dependent variable in regression (35), $\ln[\hat{P}_i + c(\gamma)]$. Estimating this regression by ordinary least squares yields an updated value of γ , which allows recomputing $c(\gamma)$ and thus $\ln[\hat{P}_i + c(\gamma)]$. We then re-estimate regression (35), and so on until convergence is achieved.

Figure 4 plots housing price gradients for five US cities. We predict the monthly rent of a dwelling with average national characteristics in a neighbourhood with average city characteristics as a function of distance to the city centre with a semilinear regression at the block-group level for each city using Yatchew's (1998) difference estimator. The dependent variable is the median contract rent in the block group. The linear component includes the same dwelling and neighbourhood controls as column (2) of table 1 while distance to the city centre is treated nonparametrically.

Panel A of figure 5 plots the city-periphery monthly rent against 2010 city population. City-periphery monthly rent is the monthly rent of a dwelling with average national characteristics in a neighbourhood with average city characteristics located at the city periphery. This is estimated from the same regression used to estimate the city-centre monthly rent used in column (3) of table 1, but valued at a distance from the centre corresponding to the periphery of each city instead of at a distance zero.

Planning regulations: The strictness of planning regulations in each metropolitan area plotted in panels B, C, and D of figure 5 is measured using the Wharton Residential Land Use Regulatory Index (WRLURI). This index is constructed by Gyourko, Saiz, and Summers (2008) applying factor analysis to responses from a 2006 nationwide survey of residential planning regulations in over 2,600 communities across the United States. Gyourko, Hartley, and Krimmel (2021) construct an updated index based on a 2018 survey with some differences with respect to the 2006 survey, both in terms of questions and responding communities. To aggregate the 2006 index to the level of metropolitan areas, we retain data on the 1896 responding communities that are part of a 1999 MSA/NECMA, average their index to the level of primary metropolitan statistical areas weighting by their population with a correction for community response probability provided by Gyourko, Saiz, and Summers (2008), and then average these values to the level of metropolitan areas weighting by population. To aggregate the 2018 index to the level of metropolitan areas, we retain data on the 1877 responding communities that are part of a 1999 MSA/NECMA, and then average their index to the level of metropolitan areas using the weights provided for large metropolitan areas by Gyourko, Hartley, and Krimmel (2021) and population weights for the rest. Finally, we interpolate the 2006 and 2018 values of the index to obtain a value for 2010 to match the timing of our other data.

Housing replacement costs and price-cost wedges: The periphery house price-cost wedge plotted against the strictness of planning regulations in panel d of figure 5 is the difference between the value of a house and its replacement cost in the periphery of the city. The house value corresponds to a four-bedroom single-family detached house built 2000–2009 in a neighbourhood with average city characteristics located at the city periphery. This is estimated based on a regression of the log of the median house value in the block group on a third-degree polynomial of distance to the city centre, and the same dwelling and neighbourhood controls as column (2) of table 1 using 2008–2012 American Community Survey (ACS) data. City periphery is again defined as the longest distance from the city centre that is within the 95th percentile of dwelling distances and has at least 500 dwelling units per square mile in the block group.

The replacement costs are the sum of city-specific construction costs for an economy-quality single-family detached house of 2000 square feet and the price of a quarter-acre vacant plot of land used for agriculture at the urban fringe. Construction costs are based on RSMeans data for 2010 obtained from Glaeser and Gyourko (2018). The urban fringe is defined as the area within 20km of land developed at medium or high intensity in 2011 that is undeveloped or developed at low intensity, based on land cover data from the 2011 slice of the NLCD2019. Within the urban fringe, we isolate land devoted to agricultural use based on the same 2011 slice of the NLCD2019. We then calculate the average price of vacant plots used for agriculture at the urban fringe of each city using gridded land value data for vacant plots from Nolte, 2020 derived from parcel sales 2000–2019. All prices converted to 2012 dollars using the Consumer Price Index for all urban consumers from US Bureau of Labor Statistics (2023).

Building permits: Data about the number of building permits plotted in figure C.1 and used for our counterfactuals are from the US Department of Housing and Urban Development (HUD). The source data is at the county level and we aggregate this up to the 1999 MSA/NECMA level. The variable annual permits relative to housing stock on the vertical axis of figure C.1 divides for each city the total number of residential construction permits during the period 2008–2012 (to match the timing of the ACS housing data) by the total number of housing units in the city for that period as recorded in the ACS data.

National Longitudinal Survey of Youth: Our estimation of the parameters governing agglomeration economies in table 2 uses panel data from the “cross-sectional sample” of the National Longitudinal Survey of Youth 1979 (NLSY79). The survey, conducted by the US Department of Labor’s Bureau of Labor Statistics, follows a nationally representative sample of 6,111 men and women who were 14–22 years old when they were first surveyed in 1979. These individuals were interviewed annually through 1994 and were interviewed on a biennial basis since 1996. We use data for the period 1979–2012. The NLSY79 contains information on a rich set of personal characteristics and tracks individuals’ labour market activities. Our starting panel is the same as in De la Roca, Ottaviano, and Puga (2023) and we refer the reader to that paper for further details. For each respondent, the confidential geocoded portion of the NLSY79 reports the county and state where they were located at birth, at age 14, and at each interview date since 1979. We use that location information both to record the 1999 MSA/NECMA where each worker is currently employed and to split work experience accumulated until then into work experience in cities with populations equal or greater than 5 million, in cities with populations equal or greater than 2 million but below 5 million, and elsewhere. Since we need a reasonable number of observations to estimate city fixed effects, we include indicators for all metropolitan areas with a population above 2 million and additional indicators for groups of similar-size metropolitan areas with a population below 2 million. In particular, we have a common indicator for cities in groups that start at 75,000 people in increments of 25,000 until 600,000, then in increments of 50,000 people until 800,000, and then in increments of 100,000 people until 2 million. This aggregates the 261 metropolitan areas included in the panel into 63 groups.

In the TSLS estimation of column (2) in table 2, we instrument the log of city size with the percentage of the area in a 30-kilometre radius around the city centre that has slopes greater than 15% and the percentage covered by wetlands (both computed as in our geographical constraints to urban expansion), the inverse hyperbolic sine of the city’s population in 1850 and 1920 (from Schroeder, 2016), the inverse hyperbolic sine of the distance to the Eastern Seaboard (computed using coastline shapefiles from the National Hydrography Dataset of the US Geological Survey), and heating degree days (from Burchfield, Overman, Puga, and Turner, 2006).

Processed data

The Stata script `code/_hcgrowth_run.do` first runs `code/1_hcgrowth_builddata.do` to perform the data construction, creating the processed data files used for the analysis and placing them in the `data/processed/` directory. The processed data consist of the following files and variables:

- `data/processed/citypop_msa_necma.dta`. This data file has population data for each US metropolitan area (1999 MSA/NECMA definitions) for 1850 and each decennial census year 1920–2020 and contains the following variables:
 - **msa**. MSA/CMSA/NECMA FIPS code (1999 definitions).
 - **msa_name**. MSA/CMSA/NECMA name.
 - **pop1850**. Population 1850.
 - **pop1920**. Population 1920.
 - **pop1930**. Population 1930.
 - **pop1940**. Population 1940.
 - **pop1950**. Population 1950.
 - **pop1960**. Population 1960.
 - **pop1970**. Population 1970.
 - **pop1980**. Population 1980.
 - **pop1990**. Population 1990.

- **pop2000.** Population 2000.
- **pop2010.** Population 2010.
- **pop2020.** Population 2020.
- data/processed/citypop_msa_necma_wtotal.dta has the same contents as data/processed/citypop_msa_necma.dta with an extra observation for total population in the conterminous United States (msa value missing and msa_name value *Conterminous United States*).
- data/processed/citypop_pmsa.dta. This data file has population data for each Primary Metropolitan Statistical Area (PMSA) along with the msa it is part of (1999 msa definitions) for 1850 and each decennial census year 1920-2020. MSAs not broken down into PMSAs and NECMAS are included as additional observations. The file is used to aggregate the 2006 Wharton Residential Land Use Regulatory Index from the PMSA level to the msa level (see *Planning regulations* above for details) and contains the following variables:
 - **msa.** MSA/CMSA/NECMA FIPS code (1999 definitions).
 - **msa_name.** MSA/CMSA/NECMA name.
 - **pmsa.** PMSA FIPS code (1999 definitions).
 - **pmsa_name.** PMSA name.
 - **pop1850.** Population 1850.
 - **pop1920.** Population 1920.
 - **pop1930.** Population 1930.
 - **pop1940.** Population 1940.
 - **pop1950.** Population 1950.
 - **pop1960.** Population 1960.
 - **pop1970.** Population 1970.
 - **pop1980.** Population 1980.
 - **pop1990.** Population 1990.
 - **pop2000.** Population 2000.
 - **pop2010.** Population 2010.
 - **pop2020.** Population 2020.
- data/processed/county2msa1999.dta. This data file (generated with code from Duranton and Puga, 2020) maps us counties to 1999 county-based metropolitan areas and contains the following variables:
 - **msa.** MSA/CMSA/NECMA FIPS code (1999 definitions).
 - **msa_name.** MSA/CMSA/NECMA name.
 - **pmsa.** PMSA FIPS code (1999 definitions)
 - **pmsa_name.** PMSA name.
 - **fips.** County FIPS code (state and county).
 - **state.** State FIPS code for the county.
 - **county.** County code.
 - **county_name.** County name.
- data/processed/controls_msa.dta. This data file has the following non-housing variables for individual metropolitan areas:
 - **msa.** MSA/CMSA/NECMA FIPS code (1999 definitions).
 - **msa_name.** MSA/CMSA/NECMA name.
 - **msa_maincity.** Main city in MSA/CMSA/NECMA.
 - **msa_lat.** MSA/CMSA/NECMA centre latitude.
 - **msa_lon.** MSA/CMSA/NECMA centre longitude.
 - **msa_area.** MSA/CMSA/NECMA area (hectares).
 - **west_mississippi.** West of Mississippi indicator.
 - **msa_eastseab.** Distance to Eastern Seaboard from city centre (km).
 - **msa_cdd.** Cooling degree days.
 - **msa_hdd.** Heating degree days.
 - **msa_cc30km_constrained.** Percentage that cannot be developed within 30km of city centre.
 - **msa_cc30km_foreign.** Percentage in a foreign country within 30km of city centre.
 - **msa_cc30km_protected.** Percentage undevelopable public land within 30km of city centre.
 - **msa_cc30km_stEEP.** Percentage with slope > 15% within 30km of city centre.
 - **msa_cc30km_water.** Percentage covered by water within 30km of city centre.
 - **msa_cc30km_wetland.** Percentage covered by wetlands within 30km of city centre.
 - **msa_fringe_constrained.** Percentage that cannot be developed in urban fringe.
 - **msa_fringe_foreign.** Percentage in a foreign country in urban fringe.
 - **msa_fringe_protected.** Percentage undevelopable public land in urban fringe.
 - **msa_fringe_stEEP.** Percentage with slope > 15% in urban fringe.
 - **msa_fringe_water.** Percentage covered by water in urban fringe.
 - **msa_fringe_wetland.** Percentage covered by wetlands in urban fringe.
 - **dist2cbd_edge.** Distance between MSA/CMSA/NECMA centre and edge.

- **college_town**. College town indicator (population < 1 million & > 10% college students 2010).
- data/processed/housing_msa.dta. This data file has the following housing variables for individual metropolitan areas:
 - **msa**. MSA/CMSA/NECMA FIPS code (1999 definitions).
 - **msa_name**. MSA/CMSA/NECMA name.
 - **ag_land_value_fringe**. Agricultural land value in urban fringe (2012 \$/ha).
 - **cost_construction**. Construction cost for 2000 sq.-foot property of economy quality (2012 \$).
 - **permits_2008_2012**. Annual housing permits 2008-2012.
 - **permit_rate_1980_2010**. Total permits per initial housing unit 1980-2010.
 - **wrluri2006**. Wharton Residential Land Use Regulatory Index 2006.
 - **wrluri2018**. Wharton Residential Land Use Regulatory Index 2018.
- data/processed/housing_blockg_acs.dta. This data file has 2008–2012 housing data from the 2012 American Community Survey (acs) at the block-group level for all block groups within a metropolitan area (1999 MSA/NECMA definitions) and additional block-group controls. Note that the block group definitions for the 2012 ACS differ from those for the 2009 NHTS data, so data from data/processed/housing_blockg_acs.dta and data/processed/travel_nhts_hh.dta should not be merged at the block-group level. The data file contains the following variables:
 - **state**. State code.
 - **county**. County code.
 - **tract**. Census Tract code.
 - **blkgr**. Census Block Group code.
 - **state_name**. State name.
 - **county_name**. County name.
 - **blkgr_name**. Block Group name.
 - **msa**. MSA/CMSA/NECMA FIPS code (1999 definitions).
 - **msa_name**. MSA/CMSA/NECMA name.
 - **msa_maincity**. Main city in MSA/CMSA/NECMA.
 - **msa_lat**. MSA/CMSA/NECMA centre latitude.
 - **msa_lon**. MSA/CMSA/NECMA centre longitude.
 - **blkgr_lat**. Census Block Group latitude.
 - **blkgr_lon**. Census Block Group longitude.
 - **blkgr_area**. Census Block Group land area (square metres).
 - **dist2cbd**. Distance to city centre (km).
 - **college_town**. College town indicator (population < 1 million & > 10% college students 2010).
 - **population**. Population.
 - **households**. Households.
 - **hunits**. Housing units.
 - **yearbuilt_median**. Median year built.
 - **pc_yearbuilt_2010_later**. % housing units built 2010 or later.
 - **pc_yearbuilt_2000_2009**. % housing units built 2000 to 2009.
 - **pc_yearbuilt_1990_1999**. % housing units built 1990 to 1999.
 - **pc_yearbuilt_1980_1989**. % housing units built 1980 to 1989.
 - **pc_yearbuilt_1970_1979**. % housing units built 1970 to 1979.
 - **pc_yearbuilt_1960_1969**. % housing units built 1960 to 1969.
 - **pc_yearbuilt_1950_1959**. % housing units built 1950 to 1959.
 - **pc_yearbuilt_1940_1949**. % housing units built 1940 to 1949.
 - **pc_yearbuilt_1939_earlier**. % housing units built 1939 or earlier.
 - **pc_structure_1detached**. % housing units in structure with 1 unit, detached.
 - **pc_structure_1attached**. % housing units in structure with 1 unit, attached.
 - **pc_structure_2**. % housing units in structure with 2 units.
 - **pc_structure_3_4**. % housing units in structure with 3 or 4 units.
 - **pc_structure_5_9**. % housing units in structure with 5 to 9 units.
 - **pc_structure_10_19**. % housing units in structure with 10 to 19 units.
 - **pc_structure_20_49**. % housing units in structure with 20 to 49 units.
 - **pc_structure_over50**. % housing units in structure with 50 or more units.
 - **pc_structure_mobile**. % housing units in mobile home.
 - **pc_structure_other**. % housing units in boat, RV, van, etc..
 - **pc_bedrooms_0**. % housing units with no bedrooms.
 - **pc_bedrooms_1**. % housing units with 1 bedroom.
 - **pc_bedrooms_2**. % housing units with 2 bedrooms.
 - **pc_bedrooms_3**. % housing units with 3 bedrooms.
 - **pc_bedrooms_4**. % housing units with 4 bedrooms.
 - **pc_bedrooms_5plus**. % housing units with 5 or more bedrooms.
 - **contractrent_median**. Median contract rent for renter-occupied housing units paying cash rent (2012 \$).

- **value_median.** Median value for owner-occupied housing units (2012 \$).
 - **pc_pop_hispanic.** % Hispanic population.
 - **pc_pop_black.** % Black population.
 - **pc_pop_asian.** % Asian population.
 - **school_rank.** Local school performance in standardised tests relative to metro area.
 - **by_coast.** By the coast or Great Lakes.
 - **by_river.** By the riverside.
 - **tri.** Terrain Ruggedness Index.
- data/processed/ipums_cps.dta. This data file has individual education data from the Annual Social and Economic (ASEC) supplement of the Current Population Survey (cps) and contains the following variables:
- **serial.** Household serial number.
 - **year.** Survey year.
 - **asecwt.** Annual Social and Economic Supplement Weight.
 - **msa.** MSA/NECMA FIPS code (1999 definitions).
 - **msa_name.** MSA/CMSA/NECMA name.
 - **msa_pop2010.** MSA/NECMA Population 2010.
 - **metro.** Metropolitan central city status.
 - **metarea.** Metropolitan area.
 - **msacmsz.** Metropolitan area size (CMSA/MSA).
 - **cbsasz.** Core-based statistical area size.
 - **age.** Age.
 - **sex.** Sex.
 - **empstat.** Employment status.
 - **educ.** Educational attainment recode.
- data/processed/nlsy_panel.dta. This data file is the general-purpose panel with NLSY79 data from De la Roca, Ottaviano and Puga (2023), re-created using their replication code (<https://diegopuga.org/data/dreams/>), and contains the following variables:
- **person_id.** Individual identifier.
 - **year.** Year.
 - **non_int.** Non-interview indicator.
 - **birth_year.** Birth year.
 - **birth_month.** Birth month.
 - **age.** Age.
 - **sex.** Sex.
 - **race.** Race.
 - **sample_type.** Sample classification.
 - **educ_enrolled.** Educational enrollment status.
 - **educ_highest.** Highest education attained at survey date.
 - **educ_mother.** Mother's years of education.
 - **educ_father.** Father's years of education.
 - **marital.** Marital status.
 - **spouse_wkswk.** Number of weeks worked by spouse in past calendar year.
 - **spouse_hrswk.** Number of hours worked per week by spouse in past calendar year.
 - **any_children.** Has at least one child.
 - **children.** \# of children (since 2000, only for female respondents and only bio children).
 - **agechildren.** Age of youngest child (only individuals with children).
 - **lfstatus.** Labor force status (generated for selected years).
 - **wkswk_li.** Weeks worked since last interview.
 - **worker_type.** Class of worker (job type), CPS definition.
 - **wage.** Hourly real wage, main job, dollars, CPS definition (1982-84 = 100).
 - **tenure.** Tenure with employer (years).
 - **experience.** Cumulative experience (years).
 - **unemployment.** Cumulative unemployment spells (weeks).
 - **outlf.** Cumulative out-of-labor-force spells (weeks).
 - **occupation3d.** Occupation, CPS definition, 1990 Census 3-digit (standardized).
 - **occupation2d.** Occupation, CPS definition, 1990 Census 2-digit (standardized).
 - **sector3d_90.** 3-digit sector (IPUMS consistent long-term classification, 1990 basis).
 - **sector2d_90.** 2-digit sector (IPUMS consistent long-term classification, 1990 basis).
 - **afqt_80.** AFQT percentile (full sample), 1980 (revised in 2006).
 - **self_80_score.** Self-confidence, raw score, 1980.
 - **self_80_zpctl.** Self-confidence, weighted z score percentile (full sample), 1980.
 - **risk_scale.** Risk aversion (10 = fully prepared to take risks).

- **fips_birth**. Residence FIPS code of birth county.
- **fips_at14**. Residence FIPS code at age 14.
- **fips**. Residence FIPS code.

Note: Producing `data/processed/nlsy_panel.dta` requires the restricted-access geocode files with the residence (county) of respondents in the NLSY79, so this file is not included in the replication package. Users without these restricted access data can create a partial version of this data file without the variables `fips_birth`, `fips_at14`, and `fips` by running `code/_hcgrowth_run.do` with the flag `global NLSYGeocodeUnavailable = 1`. Alternatively, they can set the flag `global NLSYGenerateFakeLocations = 1` and get randomly-generated values for these three variables (obviously, researchers should only select this path to check that the replication code runs smoothly since the results generated will be meaningless).

- `data/processed/travel_nhts_hh.dta`. This data file has household-level data from the 2009 US National Household Travel Survey (NHTS) and contains the following variables:

- **msa**. MSA/NECMA FIPS code (1999 definitions).
- **msa_name**. MSA/NECMA name.
- **msa_maincity**. MSA/NECMA main city name.
- **houseid**. HOUSEID.
- **blkgr**. Census Block Group code.
- **county**. County code.
- **tract**. Census Tract code.
- **state**. State code.
- **hh_drivers**. Number of drivers in household.
- **hh_hispanic**. Household respondent hispanic.
- **hh_income**. Derived total household income.
- **hh_size**. Number of household members.
- **hh_renter**. Housing unit rented.
- **hh_adults**. Number of household members aged 18 and over.
- **hh_white**. Household respondent white.
- **hh_black**. Household respondent black.
- **hh_asian**. Household respondent asian.
- **hh_single**. Single-person household.
- **hh_smallchild**. Household with smallest child aged 0-15.
- **hh_urban**. Household address in urban area.
- **hh_disttowk**. Distance to work (km., household average).
- **hh_male_sh**. Share household drivers who are male.
- **hh_educ_college**. At least one college graduate in household.
- **hh_educ_dropout**. No-one completed highschool in household.
- **hh_vmt_annmiles**. Self-reported annualized household mile estimate.
- **hh_vmt_bestmile**. Best estimate of household annual miles.
- **college_town**. College town indicator (population < 1 million & > 10% college students 2010).
- **dist2cbd**. Distance to city centre (km).
- **bg_population**. Block-group population.
- **bg_households**. Block-group households.
- **bg_hunits**. Block-group housing units.
- **bg_pc_pop_hispanic**. Block-group % Hispanic population.
- **bg_pc_pop_black**. Block-group % Black population.
- **bg_pc_pop_asian**. Block-group % Asian population.
- **bg_school_rank**. Local school performance in standardised tests relative to metro area.
- **bg_by_coast**. Block-group by the coast or lakeshore.
- **bg_by_river**. Block-group by the riverside.
- **bg_tri**. Block-group Terrain Ruggedness Index.
- **bg_area**. Block-group land area (square metres).

Note: Producing `data/processed/travel_nhts_hh.dta` requires the restricted-access file with the residence (block group) of households in the 2009 NHTS, so this file is not included in the replication package. Users without these restricted access data can run `code/_hcgrowth_run.do` with the flag `global NHTSBGUnavailable = 1` and get randomly-generated values for households' block groups (obviously, researchers should only select this path to check that the replication code runs smoothly since the results generated will be meaningless).

- `data/processed/travel_nhts_trip.dta`. This data file has trip-level data from the 2009 US National Household Travel Survey (NHTS) and contains the following variables:

- **msa**. MSA/NECMA FIPS code (1999 definitions).
- **msa_name**. MSA/NECMA name.
- **msa_maincity**. MSA/NECMA main city name.

- **college_town**. College town indicator (population < 1 million & > 10% college students 2010).
- **houseid**. HOUSEID.
- **personid**. PERSONID.
- **hh_drivers**. Number of drivers in household.
- **hh_hispanic**. Household respondent hispanic.
- **hh_income**. Derived total household income.
- **hh_size**. Number of household members.
- **hh_renter**. Housing unit rented.
- **hh_adults**. Number of household members aged 18 and over.
- **hh_white**. Household respondent white.
- **hh_black**. Household respondent black.
- **hh_asian**. Household respondent asian.
- **hh_single**. Single-person household.
- **hh_smallchild**. Household with smallest child aged 0-15.
- **hh_urban**. Household address in urban area.
- **hh_disttowk**. Distance to work (km., household average).
- **hh_male_sh**. Share household drivers who are male.
- **hh_educ_college**. At least one college graduate in household.
- **hh_educ_dropout**. No-one completed highschool in household.
- **hh_vmt_annmiles**. Self-reported annualized household mile estimate.
- **hh_vmt_bestmile**. Best estimate of household annual miles.
- **dist2cbd**. Distance to city centre (km).
- **dist2work**. Distance to work (km).
- **bg_population**. Block-group population.
- **bg_households**. Block-group households.
- **bg_hunits**. Block-group housing units.
- **bg_pc_pop_hispanic**. Block-group % Hispanic population.
- **bg_pc_pop_black**. Block-group % Black population.
- **bg_pc_pop_asian**. Block-group % Asian population.
- **bg_school_rank**. Local school performance in standardised tests relative to metro area.
- **bg_by_coast**. Block-group by the coast or lakeshore.
- **bg_by_river**. Block-group by the riverside.
- **bg_tri**. Block-group Terrain Ruggedness Index.
- **bg_area**. Block-group land area (square metres).
- **trip_min**. Trip duration (minutes).
- **trip_km**. Trip distance (km).
- **trip_kph**. Trip speed (km/h).
- **age**. Driver age.
- **male**. Driver gender male.
- **borninus**. Born in USA.
- **educ_dropout**. High-school dropout.
- **educ_hschoo**. High-school graduate.
- **educ_somecol**. Some college.
- **educ_college**. College graduate.
- **educ_postgrad**. Postgraduate.
- **worker**. Worker.
- **tdaydate**. Year and month of the household's travel day.
- **strttime**. Trip begin time in military format.
- **travday**. Day of the week of the household's travel day.
- **tdwknd**. Trip was on weekend (Friday 6pm - Sunday midnight).
- **peak**. Trip during peak hours (weekdays 6-10am, 3-7pm).
- **whytrp90**. Trip purpose.

Note: Producing `data/processed/travel_nhts_trip.dta` requires the restricted-access file with the residence (block group) of households in the 2009 NHTS, so this file is not included in the replication package. Users without these restricted access data can run `code/_hcgrowth_run.do` with the flag `global NHTSBGUnavailable = 1` and get randomly-generated values for households' block groups (obviously, researchers should only select this path to check that the replication code runs smoothly since the results generated will be meaningless).

- `data/processed/travel_nhts_hh.dta`. This data file has macroeconomic data for the entire United States for 1950, 1960, and 1965-2022 and contains the following variables:
 - **year**. Year.
 - **us_gdp_pc**. Real gross domestic product per person, United States (2012 \$).
 - **us_ayschool**. Average years of school completed by people aged >25, United States.

- **cpi**. Consumer Price Index, all urban consumers (1982-1984 = 100).

Results

The main Stata script `code/_hcgrowth_run.do`, after running `code/1_hcgrowth_builddata.do` to create the data files used for the analysis, automatically runs `code/2_hcgrowth_analysis.do` to perform the analysis of the processed data. All the results are placed in the `results/` directory.

Once the code runs, the researcher must compile in L^AT_EX the file `results/hcgrowth_tables.tex` to produce a PDF file with all the tables.

All of the numbers mentioned in the text that are not directly available in the tables are also automatically produced by `code/2_hcgrowth_analysis.do` by calling the Stata script `code/analysis/hcgrowth_text_results.do` and saved as a text file `results/hcgrowth_text_results.txt` that includes all the relevant sentences in the paper.

Figures are saved in Encapsulated PostScript format as `results/hcgrowth_fig3.eps` (figure 3); `results/hcgrowth_fig4.eps` (figure 4); `results/hcgrowth_fig5a.eps`, `results/hcgrowth_fig5b.eps`, `results/hcgrowth_fig5c.eps`, and `results/hcgrowth_fig5d.eps` (the four panels of figure 5); and `results/hcgrowth_fig1.eps` (appendix figure C.1). They are also saved in Portable Network Graphics (PNG) format with the same file names and extension `.png`.

Figures 1 and 2 illustrate the theoretical model rather than empirical results. The curves in those figures are produced by running Mathematica notebooks `code/analysis/hcgrowth_fig1.nb` and `code/analysis/hcgrowth_fig2.nb` (in the second notebook, specifying first `Year = 1980`; and then `Year = 2010`; in the first line to produce the two panels), which save figures 1 and 2 in Encapsulated PostScript format as `results/hcgrowth_fig1.eps` (figure 1); `results/hcgrowth_fig2_1980.eps` and `results/hcgrowth_fig2_2010.eps` (the two panels of figure 1). These figures are also saved in Portable Network Graphics (PNG) format with the same file names and extension `.png`.

References

- Akbar, Prottoy, Victor Couture, Gilles Duranton, and Adam Storeygard. 2023. The fast, the slow, and the congested: Urban transportation in rich and poor countries. Preprint, University of Pennsylvania.
- Baum, Christopher F., Mark E. Schaffer, and Steven Stillman. 2022. [ivreg2: Stata module for extended instrumental variables/2SLS, GMM and AC/HAC, LIML and k-class regression](#).
- Burchfield, Marcy, Henry G. Overman, Diego Puga, and Matthew A. Turner. 2006. Causes of sprawl: A portrait from space. *Quarterly Journal of Economics* 121(2): 587–633.
- Correia, Sergio. 2018. [ivreghdfe: Stata module for extended instrumental variable regressions with multiple levels of fixed effects](#).
- Correia, Sergio. 2019a. [ftools: Stata module to provide alternatives to common Stata commands optimized for large datasets](#).
- Correia, Sergio. 2019b. [reghdfe: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects](#).
- Couture, Victor, Gilles Duranton, and Matthew A. Turner. 2018. Speed. *Review of Economics and Statistics* 100(4): 725–739.
- Crow, Kevin. 2015. [shp2dta: Stata module to converts shape boundary files to Stata datasets](#).
- De la Fuente, Ángel and Rafael Doménech. 2014. Educational attainment in the oecd 1960–2010 (version 3.1). Working Paper 2014-14, Fundación de Estudios de Economía Aplicada.
- De la Roca, Jorge, Gianmarco I. P. Ottaviano, and Diego Puga. 2023. [City of dreams](#). *Journal of the European Economic Association* 21(2): 690–726.
- De la Roca, Jorge, Ingrid Gould Ellen, and Katherine M. O'Regan. 2014. Race and neighborhoods in the 21st century: What does segregation mean today? *Regional Science and Urban Economics* 47: 138–151.
- Dewitz, Jon and us Geological Survey. 2021. National Land Cover Database (NLCD) 2019 Products: Version 2.0, June 2021. Sioux Falls, SD: United States Geological Survey.
- Duranton, Gilles, and Diego Puga. Forthcoming. [Urban growth and its aggregate implications](#). *Econometrica*.
- Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles, and J. Robert Warren. 2018. Integrated Public Use Microdata Series, Current Population Survey: Version 6.0. Minneapolis: University of Minnesota.
- Forstall, Richard L. 1996. Population of States and Counties of the United States: 1790 to 1990. Washington DC: US Bureau of the Census.
- Glaeser, Edward L. and Joseph Gyourko. 2018. The economic implications of housing supply. *Journal of Economic Perspectives* 32(1): 3–30.
- Gyourko, Joseph, Jonathan S. Hartley, and Jacob Krimmel. 2021. The local residential land use regulatory environment across US housing markets: Evidence from a new Wharton index. *Journal of Urban Economics* 124: 103337.
- Gyourko, Joseph, Albert Saiz, and Anita A. Summers. 2008. A new measure of the local regulatory environment for housing markets: The Wharton Residential Land Use Regulatory Index. *Urban Studies* 45(3): 693–729.

- Jann, Ben. 2023. *estout*: Stata module to export estimation results from estimates table.
- Jann, Ben. 2020. *grstyle*: Stata module to customize the overall look of graphs.
- Jann, Ben. 2022. *palettes*: Stata module providing color palettes, symbol palettes, and line pattern palettes.
- Kleibergen, Frank, Mark E. Schaffer, and Frank Windmeijer. 2020. *ranktest*: Stata module to test the rank of a matrix.
- Lokshin, Michael. 2006. Semi-parametric difference-based estimation of partial linear regression models. *Stata Journal* 6(3): 377–383.
- Manson, Steven, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles. 2021. Integrated Public Use Microdata Series, National Historical Geographic Information System: Version 16.0. Minneapolis: IPUMS.
- Nolte, Christoph. 2020. High-resolution land value maps reveal underestimation of conservation costs in the united states. *Proceedings of the National Academy of Sciences* 117(47): 29577–29583.
- Reif, Julian. 2020. *regsave*: Stata module to save regression results to a Stata-formatted dataset.
- Riley, Shawn J., Stephen D. DeGloria, and Robert Elliot. 1999. A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences* 5(1–4): 23–27.
- Schroeder, Jonathan P. 2016. Historical Population Estimates for 2010 us States, Counties and Metro/Micro Areas, 1790–2010. Minneapolis: University of Minnesota.
- us Bureau of Labor Statistics. 2023. Consumer Price Index for all urban consumers: All Items in us City Average. Washington, dc: United States Bureau of Labor Statistics. Retrieved from FRED, Federal Reserve Bank of St. Louis.
- us Bureau of the Census. 2023. CPS Historical Time Series Tables. Washington, dc: United States Bureau of the Census.
- us Geological Survey. 2018. 1 Arc-second Digital Elevation Models – USGS National Map 3DEP Downloadable Data Collection. Reston, va: United States Geological Survey.
- us Geological Survey. 2020. Protected Areas Database of the United States (PAD-US): Version 2.1, December 2020. Reston va: United States Geological Survey.
- Yatchew, Adonis. 1998. Nonparametric regression techniques in Economics. *Journal of Economic Literature* 36(2): 669–721.