Readme for the replication files

Hinterlands, city formation and growth: Evidence from the U.S. westward expansion

Dávid Krisztián Nagy

December 9, 2022

Overview

The Matlab and Stata programs as well as the data files provided in this package generate the figures and tables of the paper, as well as the numbers mentioned in the text. The replicator should expect the entire code to run for 3-14 days.

Data availability and provenance statements

Statement about rights

- ☒ I certify that the author of the manuscript has legitimate access to and permission to use the data used in this manuscript.
- ☒ I certify that the author of the manuscript has documented permission to redistribute/publish the data contained within this replication package.

Summary of availability

- ☐ All data **are** publicly available.
- ☒ Some data **cannot be made** publicly available.
- ☐ **No data can be made** publicly available.

Data sources

The paper uses IPUMS NHGIS data (Manson et al., 2017) on U.S. county-level population and crop output, as well as the shape files of counties between 1790

and 1860. The raw data can be downloaded from *https://www.nhgis.org/*. IPUMS NHGIS allows for the redistribution of data without permission "to meet journal requirements for accessing data related to a particular publication" (*https://www.nhgis.org/citation-and-use-nhgis-data*). The NHGIS data files provided in this replication package are all in .csv format. They can be found in the folders *Files* and *Files/NHGIS*, and are described in detail below.

The paper also uses publicly available census data on the population of urban places between 1790 and 1860. The raw data can be downloaded, for instance, from *https://www.census.gov/library/working-papers/1998/demo/POP-twps0027.html* (Tables 2 to 9 under "Detailed tables"). The data files *logsettlementsize_logsettlementgrowth_decennial.csv*, *urban_cells_with_sizes.csv* and *urbanization.csv* have been created using these data. They are described in detail below.

The paper also uses FAO GAEZ data (FAO IIAS, 2015) on the agricultural productivity of the six major U.S. crops and five variables capturing natural amenities. The raw data can be downloaded from *https://gaez.fao.org/*. The FAO "encourages the use, reproduction and dissemination of material in this information product. Except where otherwise indicated, material may be copied, downloaded and printed for private study, research and teaching purposes, or for use in non-commercial products or services" (*https://gaez.fao.org/pages/disclaimer*). The data files provided in this replication package are all in .tif format. They can be found in the folders *Files/FAO_GAEZ/Agrprod* (agricultural productivity variables) and *Files/FAO_GAEZ/Amenities* (amenity variables), and are described in detail below.

Furthermore, the paper uses the ESRI Map of U.S. Major Waters, railroad maps from the website *oldrailhistory.com*, a map of U.S. territorial expansion from the National Atlas, and two maps provided in Donaldson and Hornbeck (2016). Most of these data are not in the public domain. Therefore, they are not provided in this replication package. However, the derived data files in folder *Files/Instantaneous_costs* (all in .csv format), as well as the files *land_1845.csv*, *land_finegrid.csv* and *seashore_east.csv* in folder *Files* are provided, and described in detail below.

List of data files

Below is a description of each data file, along with its name and location within the folder *Files*.

IPUMS NHGIS data

1. **NHGIS/conversion_1860.csv**: provides a conversion of 1860 county codes in the NHGIS data (the values in the single column) into the county codes that are used in *grid_1860.csv* (the number of each row in *conversion_1860.csv*; that is, row 1 is county code "1" in *grid_1860.csv*, row 2 is county code "2," and so on).

2. **NHGIS/crops_output.csv**: provides the output of each crop in 1860 (columns 3 and higher) in each 1860 county. The counties are identified jointly by columns 1 and 2, which map into the county codes in *NHGIS/conversion_1860.csv* according to the following formula: county code = 10,000 * (column 1) + (column 2). The six major crops used in the analysis are: cereals (sum of columns 3 to 6), tobacco (column 8), cotton (column 9), white potato (column 12), sweet potato (column 13), and sugar cane (column 30).

3. **NHGIS/grid_1800.csv**, **NHGIS/grid_1810.csv**, **NHGIS/grid_1820.csv**, **NHGIS/grid_1830.csv**, **NHGIS/grid_1850.csv** and **NHGIS/grid_1860.csv**: provide the county code of each 5 by 5 arc minute grid cell in the years 1800, 1810, 1820, 1830, 1850 and 1860, respectively. These files have been created by rasterizing the NHGIS shape files of counties in the corresponding years.

4. **NHGIS/pop_1790.csv**, **NHGIS/pop_1800.csv**, **NHGIS/pop_1810.csv**, **NHGIS/pop_1820.csv**, **NHGIS/pop_1830.csv**, **NHGIS/pop_1840.csv**, **NHGIS/pop_1850.csv** and **NHGIS/pop_1860.csv**: provide the imputed population of each 20 by 20 arc minute grid cell in the years 1790, 1800, 1810, 1820, 1830, 1840, 1850 and 1860, respectively. These files have been created by rasterizing the NHGIS shape files of population by county in the corresponding years, using the procedure described in Appendix B of the paper.

5. **regions.csv**: provides the large region to which each 20 by 20 arc minute grid cell belongs; a value of 1 indicates the Northeast, 2 the South, 3 the Midwest, and 4 the West. This file has been created by rasterizing the NHGIS shape file of counties in the year 1860.

Census data on urban places

6. **logsettlementsize_logsettlementgrowth_decennial.csv**: provides the log population (column "citysize") and the decennial change in log population (column "citygrowth") of urban places between 1790 and 1860.

7. **urban_cells_with_sizes.csv**: provides the grid cell (vertical coordinate: column 1, horizontal coordinate: column 2) and the population in 1790 (column 3), 1800 (column 4), 1810 (column 5), 1820 (column 6), 1830 (column 7), 1840 (column 8), 1850 (column 9) and 1860 (column 10) of each urban place above 10,000 inhabitants that came into existence between 1790 and 1860.

8. **urbanization.csv**: provides the total population of urban places above 10,000 inhabitants in 1790 (column "urbanpop_1790"), 1800 (column "urbanpop_1800"), 1810 (column "urbanpop_1810"), 1820 (column "urbanpop_1820"), 1830 (column "urbanpop_1830"), 1840 (column "urbanpop_1840"), 1850 (column "urbanpop_1850") and 1860 (column "urbanpop_1860"), as well as total population in these years (columns "totalpop_1790") to "totalpop_1860") in the four large U.S. regions (Northeast: row 1, South: row 2, Midwest: row 3, West: row 4); as well as the population living in cities that were already urban places in 1790 (row 5) and the population living in cities that exceed 0.25% of U.S. population in any given year (row 6).

FAO GAEZ data

9. **FAO_GAEZ/Agrprod/res03_crav6190l_sxlr_cer.tif**, **FAO_GAEZ/Agrprod/res03_crav6190l_sxlr_cot.tif**, **FAO_GAEZ/Agrprod/res03_crav6190l_sxlr_spo.tif**, **FAO_GAEZ/Agrprod/res03_crav6190l_sxlr_suc.tif**, **FAO_GAEZ/Agrprod/res03_crav6190l_sxlr_tob.tif** and **FAO_GAEZ/Agrprod/res03_crav6190l_sxlr_wpo.tif**: provide the potential yield of cereals, cotton, sweet potato, sugar cane, tobacco and white potato, respectively, in each 5 by 5 arc minute grid cell.

10. **FAO_GAEZ/Amenities/res01_hist_lt2_1960.tif**, **FAO_GAEZ/Amenities/res01_hist_lt3_1960.tif**, **FAO_GAEZ/Amenities/res01_hist_prc_1960.tif**, **FAO_GAEZ/Amenities/res01_hist_tmp_1960.tif** and **FAO_GAEZ/Amenities/res01_hist_crav6190.tif**: provide the number of days with minimum temperature above 5°C, the number of

days with minimum temperature above 10°C, the annual precipitation, the mean annual temperature and the annual temperature range, respectively, in each 5 by 5 arc minute grid cell.

Additional data

11. **Instantaneous_costs/rail_before1835.csv**, **Instantaneous_costs/rail_1835.csv**, **Instantaneous_costs/rail_1840.csv**, **Instantaneous_costs/rail_1845.csv**, **Instantaneous_costs/rail_1850.csv** and **Instantaneous_costs/rail_1860.csv**: provide the set of 20 by 20 arc minute grid cells that were part of the railroad network in the given year; these cells have a value of 1, while the remaining cells have a value of 0. This file has been created by georeferencing the maps in *oldrailhistory.com* on the ESRI Map of U.S. Major Waters.

12. **Instantaneous_costs/water_1790_1820.csv**, **Instantaneous_costs/water_1825.csv**, **Instantaneous_costs/water_1830.csv**, **Instantaneous_costs/water_1835.csv**, **Instantaneous_costs/water_1840.csv**, **Instantaneous_costs/water_1845.csv**, **Instantaneous_costs/water_1850.csv**, **Instantaneous_costs/water_1855.csv** and **Instantaneous_costs/water_1860.csv**: provide the set of 20 by 20 arc minute grid cells that had a navigable waterway in the given year; these cells have a value of 1, while the remaining cells have a value of 0. This file has been created by combining the ESRI Map of U.S. Major Waters with two maps in Donaldson and Hornbeck (2016): a map showing the location of natural waterways (Figure II.A in Donaldson and Hornbeck, 2016) and a map showing the location of natural waterways and canals (Figure II.B in Donaldson and Hornbeck, 2016). Canals appear after their construction dates. The construction dates of individual canals are documented in Table 10 of the paper.

13. **land_1845.csv**: provides the set of 20 by 20 arc minute grid cells that were part of U.S. territory in 1845; these cells have a value of 1, while the remaining cells have a value of 0. This file has been created manually, by combining *NHGIS/grid_1830.csv*, *NHGIS/grid_1850.csv* and the National Atlas map of U.S. territorial expansion (available for download at *https://commons.wikimedia.org/wiki/File:U.S._Territorial_Acquisitions.*

png).

14. **land_finegrid.csv**: provides the set of land-covered 5 by 5 arc minute grid cells; these cells have a value of 1, while the remaining cells have a value of 0. This file has been created using the FAO GAEZ data.

15. **newcities_diff_32_32.csv** and **newcities_diff_316_335.csv**: provide the set of 20 by 20 arc minute grid cells that have a city forming in them by 1860 according to the slow migration model of Appendix C. Columns 1 and 2 contain the latitude and the longitude of the cell, respectively. Column 3 has a value of 1 if the cell has a city forming in it, and it has a value of 0 otherwise. *newcities_diff_32_32.csv* corresponds to the equal migration cost calibration of the model, while *newcities_diff_316_335.csv* corresponds to the heterogeneous cost calibration.

16. **seashore_east.csv**: provides the set of 20 by 20 arc minute grid cells that were on the Eastern seaboard; these cells take a value of 1, while the remaining cells take a value of 0. This file has been created by selecting these cells manually, using *land_finegrid.csv*.

Data used to quantify the DNR model (Appendix D)

17. **DNR/a.csv**: provides the distribution of amenities across 20 by 20 arc minute grid cells in the DNR model. By assumption, the value is 1 in every grid cell that is part of U.S. territory.

18. **DNR/l.csv**: provides the period 0 (1790) distribution of population across 20 by 20 arc minute grid cells, based on the IPUMS NHGIS data.

19. **DNR/pop1.csv**: provides the period 1 (1795) distribution of population across 20 by 20 arc minute grid cells, imputed using the IPUMS NHGIS population data available for 1790 and 1800.

20. **DNR/ubar.csv**: provides the distribution of initial wellbeing across 20 by 20 arc minute grid cells in the DNR model. By assumption, the value is 0 in every grid cell.

Computational requirements

Software requirements

- Stata/SE 14.2

  - Package *outreg2* is necessary to run *empirical_exercises.do*. This package is installed automatically by *empirical_exercises.do* (line 7 of the code).

- Matlab (the programs were run with Matlab release 2019a)
  - Toolbox Fast Marching is necessary to run the programs *fast_marching.m*, *endrail_baseline_from1790.m*, *endrail_H1790_from1790.m* and *endrail_autarky_from1790.m*. This toolbox is available with a Matlab license from Matlab's toolbox website (Peyre, 2022). Once the toolbox has been downloaded, one needs to set a path in Matlab to both the folder *toolbox_fast_marching* and the subfolder *toolbox* that is inside the former.
  - Toolbox Opti is necessary to run the programs *DNR/gd_tau_vect.m* and *DNR/gd_tau_vect_matchuspop.m*. This toolbox is freely available from *https://github.com/jonathancurrie/OPTI*.

Memory and runtime requirements

Summary Approximate time needed to reproduce the analyses on a standard 2022 desktop machine:

☐ <10 minutes
☐ 10-60 minutes
☐ 1-2 hours
☐ 2-8 hours
☐ 8-24 hours
☐ 1-3 days
☒ 3-14 days
☐ > 14 days
☐ Not feasible to run on a desktop machine, as described below.

Details The code was last run on an **8-core, 4.6 GHz Intel-based laptop with Windows 11**.

Description of programs

The folder *Files* and its subfolders contain both the data and the programs that need to be run to replicate the results of the paper.

The programs are

- 44 Matlab .m files, some of which call additional Matlab .m files as functions,
- and 1 Stata .do file.

The following section describes the order in which the programs need to be run in order to replicate the results.

Instructions to replicators

To replicate the results of the paper, the programs in folder *Files* need to be run in the following order.

I. Codes shaping the data

1. *land_20by20.m*
2. *conf.m*
3. *agrprod.m*
4. *Instantaneous_costs/instantaneous_trade_costs.m*
5. *fast_marching.m* (fast marching toolbox needed)
6. *reduce_trmult.m*
7. *reduce_trmult_land1790.m*
8. *empirical_dataset.m*

II. Empirical analysis

9. *empirical_exercises.do*

III. Main model simulations

10. *baseline_1790.m*
11. *baseline_from1790.m*
12. *norail_from1790.m*
13. *H1790_from1790.m*
14. *autarky_from1790.m*

15. *uniformgrowth_baseline_from1790.m*
16. *uniformgrowth_norail_from1790.m*
17. *DNR/gd_tau_vect.m* (opti toolbox needed)
18. *DNR/gd_m2_vect.m*
19. *DNR/main.m*

IV. Robustness

20. *delta003_baseline_from1790.m*
21. *delta003_norail_from1790.m*
22. *delta003_H1790_from1790.m*
23. *delta003_autarky_from1790.m*
24. *delta009_baseline_from1790.m*
25. *delta009_norail_from1790.m*
26. *delta009_H1790_from1790.m*
27. *delta009_autarky_from1790.m*
28. *gam0_baseline_from1790.m*
29. *gam0_norail_from1790.m*
30. *gam0_H1790_from1790.m*
31. *gam0_autarky_from1790.m*
32. *endrail_baseline_from1790.m* (fast marching toolbox needed)
33. *endrail_H1790_from1790.m* (fast marching toolbox needed)
34. *endrail_autarky_from1790.m* (fast marching toolbox needed)
35. *DNR/gd_tau_vect_matchuspop.m* (opti toolbox needed)
36. *DNR/gd_m2_vect_matchuspop.m*
37. *DNR/gd_m2_vect_noprodgrowth.m*
38. *DNR/main_matchuspop.m*
39. *DNR/main_noprodgrowth.m*

V. Quantitative results

40. *maps.m*
41. *maps_DNR.m*
42. *maps_endrail.m*
43. *graphs.m*
44. *tables.m*
45. *numbers.m*

List of figures, tables, and numbers in the text

All figures, tables, and numbers in the text are included in this replication package, in the folder *Figures_numbers_tables*. The programs provided in the replication package reproduce:

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

Figures

Below is a description of how the figures in the paper can be re-generated.

- **Figure 1**: generated by *maps.m*.
- **Figure 2**: generated by *graphs.m*.
- **Figure 3**: generated by *graphs.m*.
- **Figure 4**: a figure showing the timing of events in the model; drawn manually.
- **Figure 5**: generated by *maps.m*.
- **Figure 6**: generated by *graphs.m*.
- **Figure 7**: generated by *maps.m*.
- **Figure 8**: generated by *maps.m*.
- **Figure 9**: generated by *maps.m*.
- **Figure 10**: generated by *maps_endrail.m*.

- **Figure 11**: generated by *maps_DNR.m*.
- **Figure 12**: generated by *maps_DNR.m*.
- **Figure 13**: generated by *maps_DNR.m*.
- **Figure 14**: generated by *maps_DNR.m*.
- **Figure 15**: generated by *graphs.m*.
- **Figure 16**: a figure showing the location of U.S. cities forming between 1790 and 1860; drawn manually, based on the location data included in *urban_cells_with_sizes.csv*.
- **Figure 17**: generated by *graphs.m*.
- **Figure 18**: generated by *graphs.m*.

Tables

Below is a description of how the results in the paper's tables can be re-generated.

- **Table 1**: generated by *empirical_exercises.do*.
- **Table 2**: generated by *empirical_exercises.do*.
- **Table 3**: generated by *tables.m*.
- **Table 4**: a table containing the list of calibrated structural parameters; created manually, using the values of these parameters and the targeted moments reported in the text of the paper.
- **Table 5**: generated by *tables.m*.
- **Table 6**: generated by *tables.m*.
- **Table 7**: generated by *tables.m*.
- **Table 8**: generated by *tables.m*.
- **Table 9**: generated by *tables.m*.
- **Table 10**: a table containing the list of canals constructed between 1790 and 1860; created manually, based on Donaldson and Hornbeck (2016).
- **Table 11**: generated by *empirical_exercises.do*.
- **Table 12**: generated by *empirical_exercises.do*.

- **Table 13**: generated by *empirical_exercises.do*.
- **Table 14**: generated by *empirical_exercises.do*.
- **Table 15**: generated by *tables.m*.

Numbers in the text

The numbers in the text of the paper are listed in *Figures_numbers_tables/ numbers.docx*. All these numbers can be re-generated by running the program *numbers.m*. Below is a list of where the exact numbers can be found in the output of *numbers.m*.

- Section 2, Fact 4: cities' average distance from their closest neighbor in the South is stored in variable *avgdistmin_d_south*, while cities' avaverage distance from their closest neighbor in the Midwest is stored in variable *avgdistmin_d_midwest*.
- Section 4.5: the difference between the average decennial population growth rates of cities and towns is stored in variable *avgdiff_d*.
- Section 5.1: the 1860 populations of the five initial cities in the data are stored in vector *incsize_d*, while the corresponding numbers in the model are stored in vector *incsize_m*. In each vector, the order of the five cities is as follows: NYC, Philadelphia, Boston, Charleston, Baltimore.
- Section 5.2: the share of cells in which shipping costs are directly decreased by railroads in 1840 is stored in variable *costloss_rail40*. The corresponding numbers for 1850 and 1860 are stored in variables *costloss_rail50* and *costloss_rail60*, respectively.
- Section 5.2: the 1860 population decrease of Philadelphia in the absence of railroads is stored in variable *poploss_rail_philadelphia*. The corresponding numbers for NYC, Boston and Chicago are stored in variables *poploss_rail_nyc*, *poploss_rail_boston* and *poploss_rail_chicago*, respectively.
- Section 5.2: the decrease in the number of people living in cities in the absence of railroads is stored in variable *urbloss_rail*.
- Section 5.2: the 1830 to 1860 annual real GDP growth rate of the U.S. economy with railroads is stored in variable *growth_30_60*, while the corresponding number in the absence of railroads is stored in variable *growth_30_60_norail*.

- Section 5.2: the decrease in 1860 U.S. real GDP in the absence of railroads is stored in variable *cf_norail*.
- Section 5.2: the decrease in 1860 U.S. real GDP in the absence of railroads in the model with uniform productivity growth is stored in variable *cf_norail_u*.
- Section 5.3.1: the decrease in 1860 U.S. real GDP without changes in political borders is stored in variable *cf_H1790*.
- Section 5.3.2: the increase in the 1860 population of the Midwest under autarky is stored in variable *midwestgain_autarky*.
- Section 5.3.2: the 1860 population loss of Boston under autarky is stored in variable *poploss_autarky_boston*. The corresponding numbers for Providence, RI and Syracuse, NY are stored in variables *poploss_autarky_providence* and *poploss_autarky_syracuse*, respectively. The population gains in Philadelphia and Baltimore are stored in variables *popgain_autarky_philadelphia* and *popgain_autarky_baltimore*, respectively.
- Section 5.3.2: the 1860 urbanization rate under trade is stored in variable *urb_trade*, while the corresponding number under autarky is stored in *urb_autarky*.
- Section 5.3.2: the decrease in 1805 to 1860 annual U.S. real GDP growth under autarky is stored in variable *growthloss_autarky*.
- Section 5.3.2: the decrease in 1860 U.S. real GDP under autarky is stored in variable *cf_autarky*.
- Section 6.2: the fraction of grid cells with railroads correctly predicted by the endogenous railroad placement framework is stored in variable *predicted*.
- Section 6.2: the overall share of these cells is stored in variable *predicted_overallshare*.

References

Donaldson, D. and Hornbeck, R. (2016): Railroads and American economic growth: A "market access" approach. *Quarterly Journal of Economics*, vol. 131(2), 799–858. *https://doi.org/10.1093/qje/qjw002*

FAO and IIASA (2015): Global Acro Ecological Zones [dataset]. Accessed March 28, 2015. *http://www.fao.org/gaez/*

Manson, S., Schroeder, J., van Riper, D. and Ruggles, S. (2017): IPUMS National Historical Geographic Information System: Version 2.0 [dataset]. University of Minnesota, Minneapolis.

Peyre, G. (2022): Toolbox Fast Marching. MATLAB Central File Exchange. *https://www.mathworks.com/matlabcentral/fileexchange/6110-toolbox-fast-marching*