Readme file for Replication Package

Roman Transport Network Connectivity and Economic Integration

by Matthias Flückiger, Erik Hornung, Mario Larch, Markus Ludwig, and Allard Mees

Replication package overview

This replication package includes:

- **DoFiles**: Two Stata files (.do) to replicate results reported in figures and tables of the main text and the appendix
  - **Main.do** contains the Stata code with which the results presented in the main part (Tables 1–8) can be replicated.
  - **Appendix.do** contains the Stata code with which the results presented in the Appendix can be replicated.
- **Estimating datasets**: Four Stata files (.dta) used in the regression analysis in the main text and the appendix
  - **Roman.dta**: Dataset covering all grid-cell pairs that lie within Western Europe and are intersected by the Roman transport network.
  - **NonRomanData.dta**: Dataset covering grid-cell pairs outside of the Roman Empire. Used in the falsification test.
  - **AllGridCells.dta**: Dataset covering all grid-cell pairs that lie within Western Europe, including those not intersected by the Roman transport network. Used in robustness tests.
  - **ProductionSiteData.dta**: Dataset at the production site × grid cell level. Used in robustness tests.
- **Programs**: One self-written Stata command (.ado) to estimate instrumental variable regressions with PPML and high dimensional fixed effects
  - **ppmlhdfe_iv.ado**: Program that performs an MM estimation of a poisson model with an endogenous covariate and multi-dimensional fixed effects.
  - **ppmlhdfe_iv_funcs.ado**: Sub-program called by ppmlhdfe_iv.ado to create regression table.
- **Rcode**: Code for reproduction of Figure 4.
- **Results**: Outputs for tables (.xls) displayed in the main text and the appendix as well as Figures 4 and A.4 (.pdf).

- **Shapefiles**: Six shapefiles (especially .shp) with geospatial vector data providing georeferenced information on the Roman transport system to calculate the Roman effective distance measure.
  - **Artificial_Roads**: Shapefile with polylines of artifical roads created to connect the centroid of a grid cell in a straight line to the closest point in the Roman transport network in the same cell
  - **Grid**: Shapefile with polygons dividing Western Europe into grid cells of 0.5×0.5 degrees longitude/latitude
  - **ProductionSites**: Shapefile with points of the respective terra sigillata production sites
  - **Rivers**: Shapefile with polylines of navigable river segments
  - **Roads**: Shapefile with polylines of Roman road segments
  - **Sea**: Shapefile with polylines of coastal routes segments

Data availability and provenance statements

The paper uses data extracted from multiple data sources. In what follows, we list the variable, its source, and its availability. We start with the two main dependent and the main explanatory variable for which we provide additional information on data cleaning and construction.

Instructions on how to obtain and construct the terra sigillata data

Variables: NumberTerraSigillataFinds, ShareTerraSigillata, IVShareTerraSigillata, ExtensiveTerraSigillata

These variables were generated from publicly available information contained in the the Samian Research Database, a database provided by the Romano-Germanic Central Museum in Mainz and the Universities of Reading and Leeds at https://www.rgzm.de/samian/. The database is an ongoing project and is continuously updated. The version of the database used for our analysis was downloaded on November 5, 2020.

Information on the precise location of the production sites is included in the shapefile 'ProductionSites.shp'. Updates may be requested from researchers at Römisch-Germanisches Zentralmuseum in Mainz.

Follow the steps outlined below to construct the bilateral, grid-cell level terra sigillata dataset from the raw data.

1. Download the terra sigillata data from https://www.rgzm.de/samian/. Navigate to Linked Open Data→Open Access→Web Feature Service (https://www1.rgzm.de/Samian/Home/wfs.html) to load latest version into QGIS or ArcGIS.
2. Associate the individual terra sigillata finds with the respective grid cell identifier (from the shapefile 'Grid.shp') → the data should now include a variable 'Grid_ID' that contains information on the destination grid cell. Rename this variable to 'grid_id_j' and save the data.
3. Associate the individual production sites from the shapefile 'ProductionSites.shp' with the grid cell identifier (from the shapefile 'Grid.shp') → the data should now include a variable 'Grid_ID' that contains information on the origin grid cell. Rename this variable to 'grid_id_i' and save the data.
4. Clean the terra sigillata finds dataset.
   (a) Drop entries without production site information.
   (b) Drop uncertain entries: entries that include hyphen+blank mark potential duplicate entries, often from lost collections. Particularly, drop those with '-'.

(c) Drop uncertain entries: entries marked as 'FL' are firing lists inscribed on a single samian vessel placed in the kiln with the rest of a load, giving the names of workmen involved and the types of vessels to be fired. Particularly, drop dies with 'FL'.

(d) Homogenize the production site names (i.e., spelling).

(e) Combine observations with multiple listed production sites into one, if and only if all production sites fall within the same grid.

5. Add the grid-cell information for the production sites (grid_id_i) to the terra sigillata dataset
6. Aggregate to the grid-cell pair level (origin-destination)
7. Remove all within grid-cell trade. Particularly, drop entries where grid_id_j is equal to grid_id_i

Instructions on how to obtain and construct the ownership data

Variables: NumberOwnership, NumberOwnershipManufacturing, NumberOwnershipService, ShareOwnership, ShareOwnershipManufacturing, ShareOwnershipService, IndustryDissimilaritySD, ShareOwnershipFalsification, ShareOwnership50Pct, ExtensiveOwnership, IVShareOwnership

These variables are derived from the commercial Orbis database by Bureau van Dijk, a provider of company and business data. The version of the database used for our analysis was downloaded from February to April 2018 (Update numbers 168 to 169) from http://orbis.bvdep.com. Follow the steps outlined below to replicate our bilateral, grid-cell level ownership variables from the raw data.

1. Using the Orbis interface restrict the database to firms with the following characteristics: an annual operating revenue of more than 2 million U.S. dollars, existing address information and that where located within Western Europe. For these firms we extracted the following variables: "Country ISO Code", "BvD ID number", "City", "NACE Rev. 2 Core code (4 digits)", "Shareholder – Name", "Shareholder - Country ISO code", "Shareholder – City", "Shareholder - Direct %", "Shareholder - Total %".
2. Manually geocode the location of the company as well as the location of the shareholders based on the information in the variables "Country ISO Code", "City", "Shareholder - Country ISO code", "Shareholder – City", using either the Google Geocoding API or Google Maps directly.
3. Determine an ownership link (stake of at least 25% in our main specification) based on the variables "Shareholder - Direct %" and "Shareholder - Total %" (i.e., we use the information in "Shareholder - Total %" if the information in "Shareholder - Direct %" is missing).

4. Use the geo-coordinates from Step 2 to assign firms and their shareholders to the respective grid identifier using the shapefile 'Grid.shp'. Shareholders should then be assigned with the variable "grid_id_i" and firms should be assigned with the variable "grid_id_j".
5. Aggregate ownership links to the grid-cell pair level (origin-destination).
6. Remove all within grid-cell ownership links. Particularly, drop entries where grid_id_j is equal to grid_id_i

Instructions on how to obtain and construct the measure of Roman effective distance

Variable: LNRomanEffectiveDistance, LNNetworkDistance, LNDirectedRomanEffectiveDistance, LNNetworkTime, SeaLCP, RiverLCP, RoadLCP, SLIntersectRiver, SLIntersectCoast, LNDifferenceFromMessina, LNIVRomanEffectiveDistance, LNFalsificationEffectiveDistance, LN0ArtRomanEffectiveDistance, LN100ArtRomanEffectiveDistance, LN150ArtRomanEffectiveDistance, LN200ArtRomanEffectiveDistance

These variables were constructed combining the shapefiles provided in the replication package that we created for the purpose of this analysis. The sources underlying the shapefiles are publicly available: The road network is extracted from the digitised version of the Barrington Atlas of the Greek and Roman World (Talbert and Bagnall, 2000). The river network represents river sections that were navigable during Roman times. We compiled this data ourselves (see Table A.2 in the Appendix for details and sources).

The battle nodes used in the construction of the instrument (LNIVRomanEffectiveDistance) were extracted from Adamson (2020). The road and river network used in the falsification test (LNFalsificationEffectiveDistance) is drawn from ESRI (2020) and `http://bit.do/WiseRivers`, respectively.

Follow the steps outlined below to construct the variable Roman Effective Distance. We used ArcGIS (Version 10.6) and the shapefiles Artificial_Roads, Roads, Rivers, Sea (see folder shapefiles) that correspond to our Roman transport network.

1. Within your working directory, create a "File Geodatabase" and then a "Feature Dataset" using ArcCatalogue. Import the transport network shapefiles into the "Feature Dataset".
2. Create a "Network Dataset" within the "Feature Dataset" using the transport network shapefiles as source files. Allow for any vertex under the connectivity policy. Under "Attributes" create a new attribute (Roman Effective Distance). For the evaluators of the new attribute specify first Type=Field and then use the evaluator properties (expression) to specify costs for each mode of transport. Specifically, set the costs for (artificial) road=52* LengthSha, river=7.5* LengthSha, sea=LengthSha.

3. Build the network and calculate bilateral least cost routes (and associated costs: Roman Effective Distance) using for example an OD Cost Matrix in the Network Analyst.
4. To build the measure of network distance (LNNetworkDistance), set the costs for (artificial) road=1* LengthSha, river=1* LengthSha, sea=LengthSha.
5. To build the measure of network time (LNNetworkTime), set the costs for (artificial) road=1.85* LengthSha, river=0.7825* LengthSha, sea=LengthSha.

Instructions on how to obtain data or underlying sources for all remaining variables

Please see the Online Appendix for details on the construction of these variables.

Alpine jade (ShareAlpineJade) This variable was generated using information from the Alpine jade database as part of the project "JADE: Social inequalities in Neolithic Europe: the circulation of long axeheads of Alpine jades" (Pétrequin et al., 2012). The database contains precise information on the find site of jade axeheads and can be downloaded at http://jade.univ-fcomte.fr.

Burial traditions (BothDolmen, ChamberedCairns, Menhirs, RoundBarrows) These variables were generated using publicly available information on the location of megalithic structures from the Megalithic Portal at www.megalithic.co.uk, last accessed on 23 October 2020. In particular, dolmen from `http://bit.do/burialchambers`, chambered cairns from `http://bit.do/chamberedcairns`, and menhirs from `http://bit.do/menhirs`, and round barrows from `http://bit.do/roundbarrows`.

Access to a waterway (BothIntersectedWaterway) This variable was generated in R using the Rivers.shp shapefile included in this replication package.

Both Mediterranean Sea (BothMediterranean) This variable was generated in R using the Sea.shp shapefile included in this replication package.

Same country (CountryPairFE) This variable was generated by identifying whether two firms are located in the same country. The location of firms was geocoded manually (see above); the country polygons were drawn from `gadm.org`.

Absolute difference in agricultural suitability (DifferenceCSI) This variable was generated using publicly available information from `https://ozak.github.io/Caloric-Suitability-Index/`.

Absolute difference in elevation (DifferenceElevation) This variable was generated using publicly available information on elevation from WorldClim (v. 2.1). This database is described in Fick and Hijmans (2017) and available at `https://www.worldclim.org/data/worldclim21.html`.

Absolute distance latitude (DifferenceLatitude) This variable was generated in R using the Grid.shp shapefile included in this replication package.

Absolute difference in longitude (DifferenceLongitude) This variable was generated in R using the Grid.shp shapefile included in this replication package.

Absolute difference in precipitation (DifferencePrecipitation) This variable was generated using publicly available information on average precipitation in millimetres over the 1970–2000 time horizon from WorldClim (v. 2.1). This database is described in Fick and Hijmans (2017) and available at `https://www.worldclim.org/data/worldclim21.html`.

Absolute difference in ruggedness (DifferenceRuggendess, RuggednessAlongSL) These variables were generated combining publicly available information on elevation from WorldClim (v. 2.1) and the index devised in Riley, Degloria and Elliot (1999). WorldClim database is described in Fick and Hijmans (2017) and available at `https://www.worldclim.org/data/worldclim21.html`.

Absolute difference in access to waterways (DifferenceStrahler) This variable was generated using publicly available information on river systems obtained from Vörösmarty et al. (2000).

Absolute difference in temperature (DifferenceTemperature) This variable was generated using publicly available information on temperature in degrees Celsius over the 1970–2000 time horizon from WorldClim (v. 2.1) This database is described in Fick and Hijmans (2017) and available at `https://www.worldclim.org/data/worldclim21.html`.

Preferences (PreferencesSD, DistanceTrustSD, DistanceAltruismSD, DistanceNegativeReciprocitySD, DistancePositiveReciprocitySD, DistanceRiskSD, DistanceTimeSD) These variables were generated using publicly available data from the Global Preferences Survey (GPS, Falk et al., 2018).

Attitudes and values (ValuesSD, DistanceLifeSD, DistanceWorkSD, DistanceFamilySD, DistancePoliticsSocietySD, DistanceReligionSD, DistanceNationalismSD) These variables were generated using publicly available data from the European Values Study EVS (2016). Population data at the NUTS2-level were taken from Schiavina, Freire and MacManus (2019).

Joint duration under Roman rule (JointDurationRomanRule) This variable was handcoded from information contained in the publicly available source Shepherd (1923).

Geodesic distance (LNDGeodesicDistance) This variable was generated in R by computing the geodesic distance between the centroids of grid cells in the Grid.shp shapefile.

ln Δ geodesic distance to Messina (LNDGeodesicFromMessina) This variable was generated in R by computing the geodesic distance between the centroid of the grid cell in which Messina is located and the centroids of all other grid cells in the Grid.shp shapefile.

The topography-based least-cost path (LNHMISea) This variable was generated based on the least cost path identified on the basis of the Human Mobility Index with Seafaring (HMISea) by Özak (2018). The underlying GeoTiff files are publicly available and can be downloaded from https://human-mobility-index.github.io/.

Energy expenditure (LNPandolfGivoniGoldman LNLVanLeusen LNLloberaSluckin LNHerzog) These variables were constructed using the R-package 'movecost' developed by Alberti (2019). All energy expenditure functions used in our analysis are listed in Alberti (2019).

Travel time (LNLangmuir) This variable was generated in R using the travel time function of Langmuir (2003, pp. 39 ff.).

Google driving distance (LNDrivingDistanceSD) This variable was generated using publicly available data from Google Maps. It was generated using the Distance Matrix API, last used on 28 October 2020.

Rome2Rio travel time (LNRomeToRioSD) This variable was generated using publicly available data from `rome2rio.com`. It was generated using an API on the production servers on 28 September 2018. Access was made available free of charge by the good people at Rome2Rio Labs.

Social connectedness (LNSCISD) This variable was generated using limited access information on social ties from facebook. Accessed on 24 April 2020. In the meantime, this data was externally shared and is publicly available under https://data.humdata.org/dataset/social-connectedness-index.

Timing of Black Death onset (LagOnsetPlague) This variable was generated using information on the timing of onset of the Black Death contained in the publicly available source Christakos et al. (2005, pp. 214–282).

Correlation nighttime lights (LightCorrelation, LightCorrelationBaxterKing) This variable was generated using night-time luminosity data from the Defense Meteorological Satellite Program-Optical Line Scanner (DMSP-OLS) sensor. This data is publicly available for the years 1992–2013 at a spatial resolution of 1×1 kilometres from. It can be accessed here: `https://www.ngdc.noaa.gov/eog/dmsp/downloadV4composites.html`.

Oppida settlements (Oppida) This variable was generated using publicly available information on the location of Celtic settlements (Oppida) during the La Tène culture in Gaul from the Oppida portal, an initiative of the Marc Bloch University, see `http://bit.do/oppida`.

Same biome (SameBiome) This variable was handcoded from information contained in the publicly available source Olson et al. (2001).

Same watershed (SameWatershed) This variable was generated using publicly available information on watersheds from the ECRINS (v. 1.1) data set. This database is provided by the European Environment Agency and available at `https://www.eea.europa.eu/ds_resolveuid/b0bbe232ca62427288bdcdee22ca95e4`

All goods (ShareAllPreRomanGoods ShareAllPreRomanGoodsAssigned) These variables were generated by combining the information on alpine jade, British jade, and metal goods. Sources: see Alpine, British jade, and Metal goods.

British jade (ShareBritishJade) This variable was generated using the publicly available database provided in Schauer et al. (2020).

Metal goods (ShareMetalGoods) This variable was generated using hand-collected data combining information reported in 36 separate publications (listed in Table E.1 in the Appendix). Based on information reported in the original publications, we georeferenced the find site and provenance of each artefact.

Super grid cells (SuperGrid_i and SuperGrid_j) This variable was generated in R by aggregating the grid cells in Grid.shp at the 1×1 degree longitude/latitude level.

Price correlation (WheatPriceCorrelation) This variable was generated from computing grid-cell-pair-level price correlations using data compiled in Federico, Schulze and Volckart (forthcoming). This database is not publicly available and we are grateful to Giovanni Federico for sharing the data.

Summary of availability

Except for the commercial Orbis database from which we created the bilateral firm-ownership data, all data or their underlying sources are publicly available without pay.

Statement of rights

The author(s) of the manuscript have or had legitimate access to and permission to use the data used in this manuscript.

Description of programs/code and instructions for replication of regression tables

Results reported in the paper were produced using STATA 16. In order to replicate results, set the working directory to the replication folder, then execute the code.

To run the replication, you may need to install the following packages from the SSC archive: ftools, outreg2, ppmlhdfe, putexcel, reghdfe

To estimate the instrumental variable regressions for Table 5 with PPML and high dimensional fixed effects, the .do file calls the program 'ppmlhdfe_iv.ado' located in the folder 'Programs'. The syntax to call the program is:

ppmlhdfe_iv depvar [varlist1] (var_endog = var_iv) [if] [in] , [**fe**(fe_syntax)] [further options]

Please type for further details "help ppmlhdfe_iv" in Stata after you loaded the program with "qui do ppmlhdfe_iv.ado".

References

Adamson, Jordan. 2020. "Political Institutions, Resources, and War: Theory and Evidence from Ancient Rome." *Explorations in Economic History* 76: 101324.

Alberti, Gianmarco. 2019. "movecost: An R Package for Calculating Accumulated Slope-dependent Anisotropic Cost-surfaces and Least-cost Paths." *SoftwareX* 10: 100331.

Christakos, George, Ricardo A. Olea, Marc L. Serre, Lin-Lin Wang, and Hwa-Lung Yu. 2005. *Interdisciplinary Public Health Reasoning and Epidemic Modelling: The Case of Black Death.* Springer.

ESRI. 2020. "World Roads." http://bit.do/worldroads.

EVS. 2016. "European Values Study 2008." 4th wave, Integrated Dataset. GESIS Data Archive, Cologne, Germany, ZA4800 Data File Version 4.0.0.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. "Global Evidence on Economic Preferences." *The Quarterly Journal of Economics* 133 (4): 1645–1692.

Federico, Giovanni, Max-Stephen Schulze, and Oliver Volckart. forthcoming. "European Goods Market Integration in the Long Run." *Journal of Economic History.*

Fick, Stephen E., and Robert J. Hijmans. 2017. "WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas." *International Journal of Climatology* 37 (12): 4302–4315.

Langmuir, Eric. 2003. *Mountaincraft and Leadership.* Vol. 3 edition. Manchester & Aviemore: British Mountaineering Council.

Olson, David M, Eric Dinerstein, Eric Wikramanayake, [...], Wesley Wettengel, Prashant Hedao, and Kenneth Kassem. 2001. "Terrestrial Ecoregions of the World: A New Map of Life on Earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity." *BioScience* 51 (11): 933–938.

Özak, Ömer. 2018. "Distance to the Pre-industrial Technological Frontier and Economic Development." *Journal of Economic Growth* 23 (2): 175–221.

Pétrequin, Pierre, Serge Cassen, Michel Errera, L. Klassen, Alison Sheridan, and Anne Marie Pétrequin. 2012. *Jade: Grandes haches alpines du Néolithique européen, Ve et IVe millénaires av. J.-C.* Centre de Recherche Archéologique de la Vallée de l'Ain.

Riley, Shawn, Stephen Degloria, and S.D. Elliot. 1999. "A Terrain Ruggedness Index that Quantifies Topographic Heterogeneity." *Internation Journal of Science* 5: 23–27.

Römisch-Germanisches Zentralmuseum in Mainz. "`http://www.rgzm.de/samian/`."

Schauer, Peter, Andrew Bevan, Stephen Shennan, Kevan Edinborough, Tim Kerig, and Mike Parker Pearson. 2020. "British Neolithic Axehead Distributions and Their Implications." *Journal of Archaeological Method and Theory* 27: 836–859.

Schiavina, Marcello, Sergio Freire, and Kytt MacManus. 2019. "GHS-POP R2019A - GHS Population Grid Multitemporal (1975-1990-2000-2015)." European Commission, Joint Research Centre (JRC).

Shepherd, William R. 1923. *Historical Atlas.* New York: Henry Holt and Company.

Talbert, Richard, and Roger Bagnall. 2000. *Barrington Atlas of the Greek and Roman World.* Princeton, N.J.: Princeton University Press.

Vörösmarty, C. J., B. M. Fekete, M. Meybeck, and R. B. Lammers. 2000. "Global System of Rivers: Its Role in Organizing Continental Land Mass and Defining Land-to-ocean Linkages." *Global Biogeochemical Cycles* 14 (2): 599–621.