DISCUSSION PAPER SERIES

DP15362

HISTORICAL DATA: WHERE TO FIND THEM, HOW TO USE THEM

Paola Giuliano and Andrea Matranga

ECONOMIC HISTORY

ISSN 0265-8003

HISTORICAL DATA: WHERE TO FIND THEM, HOW TO USE THEM

Paola Giuliano and Andrea Matranga

Discussion Paper DP15362
Published 12 October 2020
Submitted 11 October 2020

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

HISTORICAL DATA: WHERE TO FIND THEM, HOW TO USE THEM

Abstract

The use of historical data has become a standard tool in economics, serving three main purposes: to examine the influence of the past on current economic outcomes; to use unique natural experiments to test modern economic theories; and to use modern economic theories to refine our understanding of important historical events. In this chapter, we provide a comprehensive analysis of the types of historical data most commonly used in economic research and discuss a variety of issues that they raise, such as the constant change in national and administrative borders; the reshuffling of ethnic groups due to migration, colonialism, natural disasters, and many other forces. We also point out which methodological advances allow economists to overcome or minimize these problems.

JEL Classification: N0

Keywords: Historical data, geographical data, ethnographic data, censuses

Paola Giuliano - paola.giuliano@anderson.ucla.edu
UCLA and CEPR

Andrea Matranga - andreamatranga@gmail.com
Chapman University

Acknowledgements
Prepared for the Handbook of Historical Economics, edited by Alberto Bisin and Giovanni Federico. We thank seminar participants at the Handbook of Historical Economics Conference at New York University. All remaining errors are ours.

Historical Data: Where to Find Them, How to Use Them[1]

Paola Giuliano, UCLA, NBER, CEPR and IZA
Andrea Matranga, Chapman University

August 2020

Abstract

The use of historical data has become a standard tool in economics, serving three main purposes: to examine the influence of the past on current economic outcomes; to use unique natural experiments to test modern economic theories; and to use modern economic theories to refine our understanding of important historical events. In this chapter, we provide a comprehensive analysis of the types of historical data most commonly used in economic research and discuss a variety of issues that they raise, such as the constant change in national and administrative borders; the reshuffling of ethnic groups due to migration, colonialism, natural disasters, and many other forces. We also point out which methodological advances allow economists to overcome or minimize these problems.

[1] Prepared for the *Handbook of Historical Economics*, edited by Alberto Bisin and Giovanni Federico. We thank seminar participants at the Handbook of Historical Economics Conference at New York University. All remaining errors are ours.

1. Introduction

The use of historical data has become a standard tool in economics. The use of these data poses several challenges. Historical data are collected by different people, at different times, using different units of analysis. Especially when looking at long periods of time, not only the quality can vary but also the geographical unit needs to be adjusted, for example because borders across geographical units are not constant. The composition of populations can also change due to migration, natural disasters and many other forces.

This chapter describes the main data sources that have been profitably used in economics research, and their most prominent applications. We broadly classify them in geographical data, ethnographic data and Censuses. We also describe other commonly used historical data. For each group, we outline the issues they raise and also point out which methodological advances allow economists to overcome or minimize these problems.

The chapter starts with the description of geographical data. Exploiting the geographical element of the data to match different variables over time, is the most common task economic historians have to perform before even starting to analyze their data. We start by describing how original sources should be evaluated: before going through the time and effort of digitizing a map, for example it is important to do a preliminary step and considers capabilities and motives of the individual or entity who compiled the map. We proceed by showing how to use a geographical information systems software (GIS) to convert historical maps in modern format and how to analyze and integrate the data obtained using spatial locations. After digitizing different maps, one of the difficulties is that both the number and the geographical boundaries can change over time. Take the case of the United States: there were only 250 counties in 1790, whereas the number in 2000 is well above 3000. The chapter describes in details different potential ways of linking counties across the different census years. We finally describe other geographical data commonly used in economics, which, unlike old maps that need to be digitized, are already available and ready to use, simply outlining potential issues in using them.

The chapter proceeds by describing the use of ethnographic data. A prolific strand of the economic literature has documented a strong persistence of economic outcomes over time, including economic growth, political development and a variety of cultural traits (Putternam and Weil, 2010, Spolaore and Wacziarg, 2013, Michalopoulos and Papaioannou, 2013, Alesina, Giuliano and Nunn, 2013, Voigtlaender and Voth, 2012). The most comprehensive information about societal characteristics going back to pre-industrial times is the *Ethnographic Atlas,* assembled by Murdock

(1967a), containing information on political, economic and cultural traits of societies. Economists vastly used this dataset to document persistence in political, economic or cultural outcomes (Gennaioli and Rainer, 2007; Michalopoulos and Papaioannou, 2013, Giuliano and Nunn, 2013; Alesina et al., 2013; Becker, 2019 to quote just few studies). The challenge faced by researchers when using this important source is how to connect historical characteristics of pre-industrial societies to current outcomes: the further back into the past one goes, the more the population composition of a given place tends to diverge from people who currently live there, because ethnicities moved over time or even disappeared. For example, the ethnicities reported in the *Ethnographic Atlas* for the case of the United States were Native-American populations, mostly involved in hunting, fishing and horticultural communities and organized in small, pre-state political units. By contract, if one looks at the actual composition of the population in the United States today, a large fraction comes from ethnicities that lived in settled agricultural societies organized in large states. The chapter describes different ways used in literature to link the past to the present, using historical geographical locations (Alsan, 2015; Mayshar et al, 2015), current distribution of languages or ethnicities across the world (Alesina et al., 2013, Giuliano and Nunn, 2018), a migration matrix (Putternam and Weil, 2010) or individual data at the ethnicity level (Alesina, Giuliano and Nunn, 2013; Becker, 2019).

Recent advances in economic history came from the use of historical complete census data (mostly for the United States), whose biggest advantage has been to link individuals' names over time, therefore allowing researchers to provide new and more precise answers to topics such as social mobility (Abramitzky et al., 2014). Contrary to modern administrative data, which contains social security numbers, historical censuses can only be linked using name, presenting a huge challenges especially for individuals with common names. The chapters present the different methods used to match individuals over time (and their limitations): from direct match (Ferrie (1996) and Abramitzky et al., (2012, 2014, 2019a)) to a more sophisticated machine-learning (Feigenbaum, 2016) or fully automated probabilistic algorithm (Abramitzky et al., 2019b, 2019c)

The paper is organized as follows: Section 2 describes the use of geographical data. Section 3 is dedicated to ethnographic data. Sections 4 and 5 describe Census data and other historical data, whereas Section 6 concludes.

2. Geographical Data

Researchers using historical datasets often need to exploit the geographical element of their data to match the different variables to each other. Economic historians are much more likely to use

data that were collected by different people at different times, using different units of analysis. For example, an economic historian who wants to study the effect of wars on contemporary outcomes might be using the amount of wars engaged in by German principalities in the 15th century, but might need to cross-reference them with the population levels of German provinces a few centuries later, which have completely different borders. The only way to do so is exploiting the geographic overlap in the two sets of jurisdictions.

The common way of dealing with geographical data is a geographic information system (GIS) which allows to analyze and integrate different types of data, using spatial locations. Any GIS data one can find will be of one of two kinds, shapefiles and rasters. Most kinds of data can be represented in either format, and it is possible to convert between the two, but different types of data and collection processes generally call for one or the other.

i) *Shapefiles* are vector formats, that give precise coordinates for specific points, which might represent distinct features such as villages or wells, or be strung together in lines or polygons to describe borders, roads, canals, etc. For example if we were interested in using the border between Colorado and Wyoming (a straight line along the 41st parallel North) for an RDD along the lines of Card and Krueger (1994), we could summarize it in the following very short shapefile giving only the two endpoints (41, -109.5; 41 -102.05). Even relatively complicated borders (or other linear features) can be summarized by files that are quite small. Shapefiles can be similarly used to describe point features, such as villages, wells, or homicides.

ii) The second way of representing geographical data is through *rasters*. Rasters are image files that have been georeferenced, i.e. have been tagged with information on the geographic location of the corners of the image, and the precise way the flat image should be stretched to conform to the curved surface of the earth. For the case of the border between Wyoming and Colorado, one would have to create a raster covering the entire states of Wyoming and Colorado with cells small enough to record the data with the precision required. The difficulty with creating a border using a raster file is the creation of a large number of records coded as "Not a Border" for all the points in the interior of the two states, which could easily result in a file be several gigabytes in size.

Generally different types of data are better suited for storage using one of the two methods. Data originally created as a rectangular grid of points, for example scanned maps, aerial photographs,

or satellite imagery and its many derived products (e.g. cloud cover, rainfall, landcover) are normally saved as raster files, whereas data indicating just specific locations, such as borders, wells, or villages, are more conducive to the use of shapefiles. All GIS packages provide many tools for converting data from raster to shapefile and vice versa.

Below we provide a concrete example on how to convert historical data using the GIS tools. Often economic historians need to work with scanned maps, whose quality can vary. Suppose that the researcher needs to convert a paper map of medieval borders in Europe, which might show each country with a different color. Once scanned each country will in fact be composed of pixels that slightly differ in shade, because of imperfections in the paper and printing technique, as well as random fluctuations in the acquisition process.

Esri ArcGIS provides the "Classify" tool that allows researchers to collapse all these similar colors to a raster of unique values, which can then be used as is, or converted to polygons. Generally, such a dataset will still need to be cleaned to remove objects within each country that are of a different color, such as rivers, cities, and printed text.[2]

2.1. Assessing the Suitability of Geographic Data

Before going through the time and effort of digitizing a paper map, it is necessary to do a preliminary step and consider the capabilities and motives of the individual or entity who compiled the map. To warrant use, the map should have been compiled by people that had the technical and resource capability to collect the data, and the incentives to do so accurately (unless of course reporting biases are themselves being studied!).

How was the map created?

It is often useful to have at least a rough understanding of how the map was compiled. For example, if a 19th cartographer was asked to produce a map of UK counties by population density, the best procedure will involve two steps. In the first step one could use the map to simply trace out the borders of the British counties. In the second step, the researcher could use data from the UK census to convert the population densities into shades, and color in the counties appropriately. Since this tracing process inevitably introduces errors, digitizing this map directly would import those errors into the dataset. It would therefore be better to instead digitize the county borders from a large scale

[2] This process can often be partially automated by running the "Zonal Statistics" tool with the Median option, and a radius larger than the thickness of the largest feature that should be ignored. For example if the problem is a network of rivers that are 5 pixels wide, a radius of 10 pixels will usually work well. These routines are sufficiently powerful to generally pick out roads, forests, or other similar uniformly colored features from satellite imagery.

map directly into a shapefile, and then assign to each county the correct values from the population density map.

Other issues can arise when maps are compiled by cartographers based on accounts from explorers. The report might have said simply that after crossing to the left bank of River X, she encountered population Y. In absence of further details, the cartographer might have generalized this observation by assigning the entire left bank of the river to population Y, which may or may not have been true. If the area under study is an entire continent, it is perhaps acceptable to use a map with such generalizations, since over a large sample size many sources of error will even out. But if we are interested in estimating the long run persistence of the effect of the institutions of the single population Y, such a map could not be reasonably be taken at face value. Tracking down the exact history of how a map was made could be too laborious (particularly if it is not the variable of interest, but one of dozens of controls), but a diligent researcher will seek contact with experts on the history of population Y, that could point out what controversies might exist over the precise borders of their territory.

How to deal with interpolated data

Another common issue in digitizing a map is the use of interpolated data, especially common when one has to deal with weather data. Suppose one is interested in converting point data for rainfall to a raster file. The first step before using such a point dataset, would be to go to the original point dataset and interpolate the data yourself with a known method rather than using the historical interpolation. If the variable in question can be expected to change very discontinuously, it might be advisable to eliminate observations that are too far away from an actual sampling point. For example, if a researcher was using data from pollution that was interpolated based on data from a collection of measurement stations, it would be wise to check whether (as it is likely) these stations are more commonly placed in populated centers, which are likely to also be major sources of pollution. If that is the case, the researcher would be wise to eliminate interpolated observations that were more than a fixed distance from each actual measurement.

As a best practice to make the assessment of the original quality of the data transparent, one could add some explanation on the quality of the original data. For example "The map was compiled by the Engineer Corps of India to facilitate movement of troops, based on an extensive survey conducted over the course of 10 years. Since they had the technical capability, and the incentives to record the land accurately, we can assume that the location of the villages was accurately recorded".

2.2. Projections, inaccurate or sketched maps, distance measures

Map projections describe mathematically how an Earth's curved surface is conceptually flattened for representation in maps or their digital equivalent. To be an ideal representation of geographic reality, a map would have to be conformal, equal area, and equidistant. *Conformal* means that the map is not stretched locally. *Equal area* implies that areas on the map are proportional to areas on the planet. And *equidistant* means that distances are preserved, though unfortunately this is possible only between two points.

Unfortunately, while projections can have any of these properties, no projection can have all of them at the same time. So economic historians have to pick the best compromise depending on the area depicted by the map, and its intended use. For example, if we are interested in using a shapefile of country borders and a raster of total annual rainfall to calculate average rainfall within different countries, it's important to ensure that the calculation is performed using an *equal area* projection. This will ensure that each pixel in the precipitation raster represents the same geographic extent, and will avoid certain areas of each country being incorrectly weighted more.

Universal Transverse Mercator Projection (UTM)

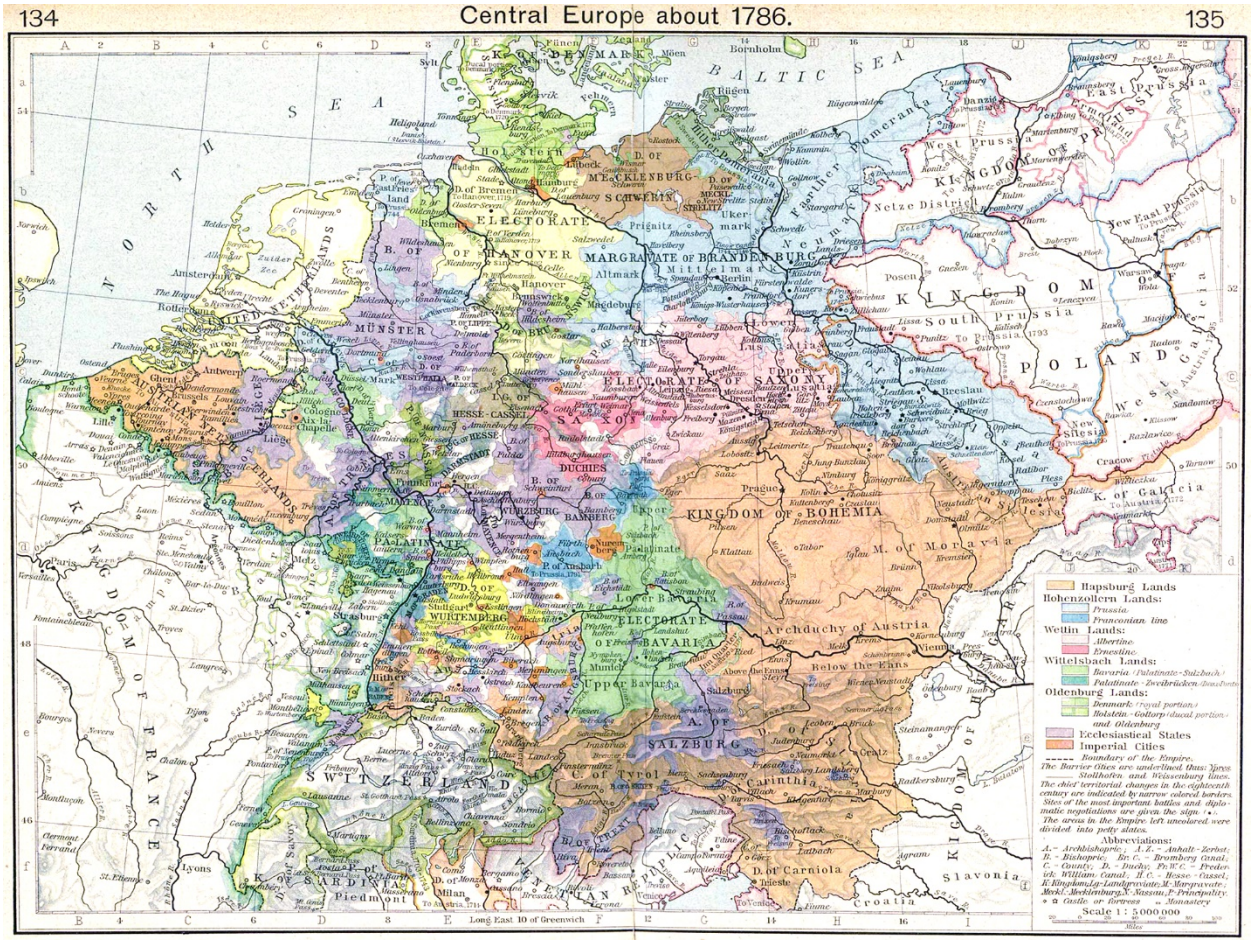
For areas up to a few hundred kilometers in extent, the easiest solution is to use the appropriate Universal Transverse Mercator (UTM) projection. The UTM projection is a family of projections each covering a slice of the Earth that is 6 degrees wide in longitude. The area within each slice can be treated as essentially flat, and coordinates are given in meters East or North of an arbitrary point. Specifically the size of the slices was chosen so that any errors in distance would be at most 0.1% which is essentially negligible for most economic history projects.

Below we describe a workflow to digitize historical maps.

Approaches to digitization

Let's assume a researcher is interested in creating a shapefile with the borders of the Holy Roman Empire in 1786, and the territories of imperial cities, based on the following map, which uses an unknown projection (Figure 1).

Figure 1. The Holy Roman Empire in 1786



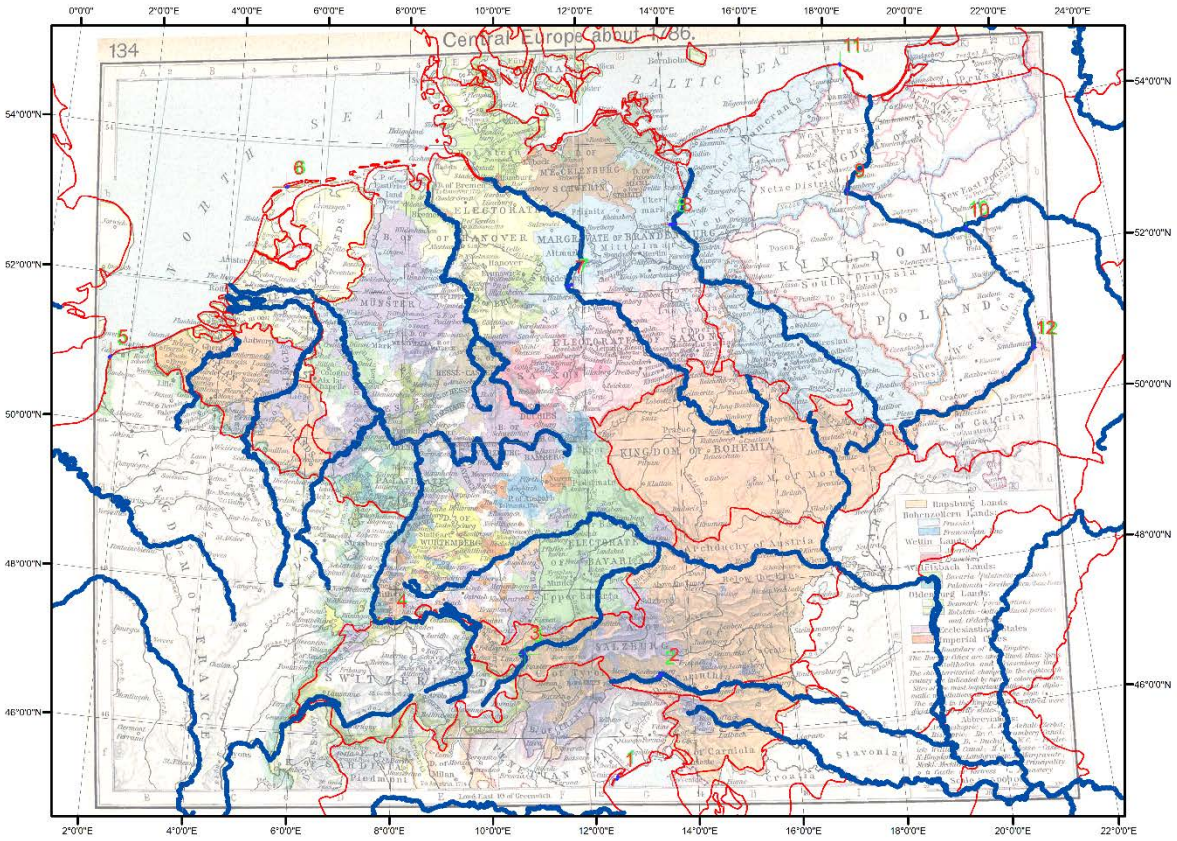
One would first load some already georeferenced data for the same area, to which points from the scanned map can then be matched. For example in this case our base-map consisted of data on country borders (and coastlines) and major rivers. We first need to click on a point to geo-reference on the scanned map, and then on the corresponding spot on our base-map. Two is the minimum number of point pairs, but this will only be precise if one knew or correctly guessed the projection. More points help average out some inaccuracies and will allow for progressively more flexible transformations such as polynomials up to order three.

In our example, given the European setting and the fact that the meridians converge towards North, we first guessed that the map had been projected using some sort of conic projection.[3] We

[3] In this case we chose the Europe Albers Equal Area Conic as the base projection. This projection is a conic, equal area map projection that uses two standard parallels. With this projection, although scale and shape are not preserved, the distortion between the two chosen parallels is minimal. If an incorrect projection had been chosen, the geo-referencing

then matched the locations of a number of prominent points on each map. Note that in this case, since a graticule was available in the scanned map (Figure 2), it would have been more accurate to match the intersections of the latitude and longitude lines, but since these are not always available we wanted to show how points could be selected. Instead in this case the two graticules can be compared to show the inaccuracies introduced by this process (most prominent in the North Sea).

Figure 2. The Holy Roman Empire in 1786, Georeferenced Using Coastlines and Rivers



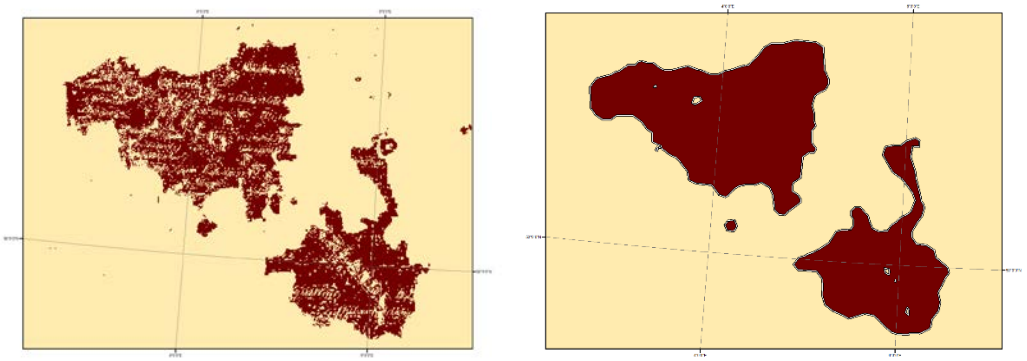
Using more points is generally better for older, hand traced maps, since they average out the inevitable inaccuracies. After selecting each point, one should check if the general fit of the two maps improves, and remove the point if it does not (this usually means the compiler of the map made a tracing error). If older maps that appear to be locally distorted are used, it is better to use the *spline* function. This forces each pair of points to fit exactly, and then stretches the parts in between to make

would still have worked, but it would have been necessary to specify more points, and use a third order polynomial or spline stretching to ensure a good fit.

them fit. One should continue by adding points until all visible features match. Again, if adding a particular point makes the fit visibly worse, that point should be deleted, and other reference points used. This can happen if for example the cartographer was inaccurate in placing a particular village which is being used as a reference point.

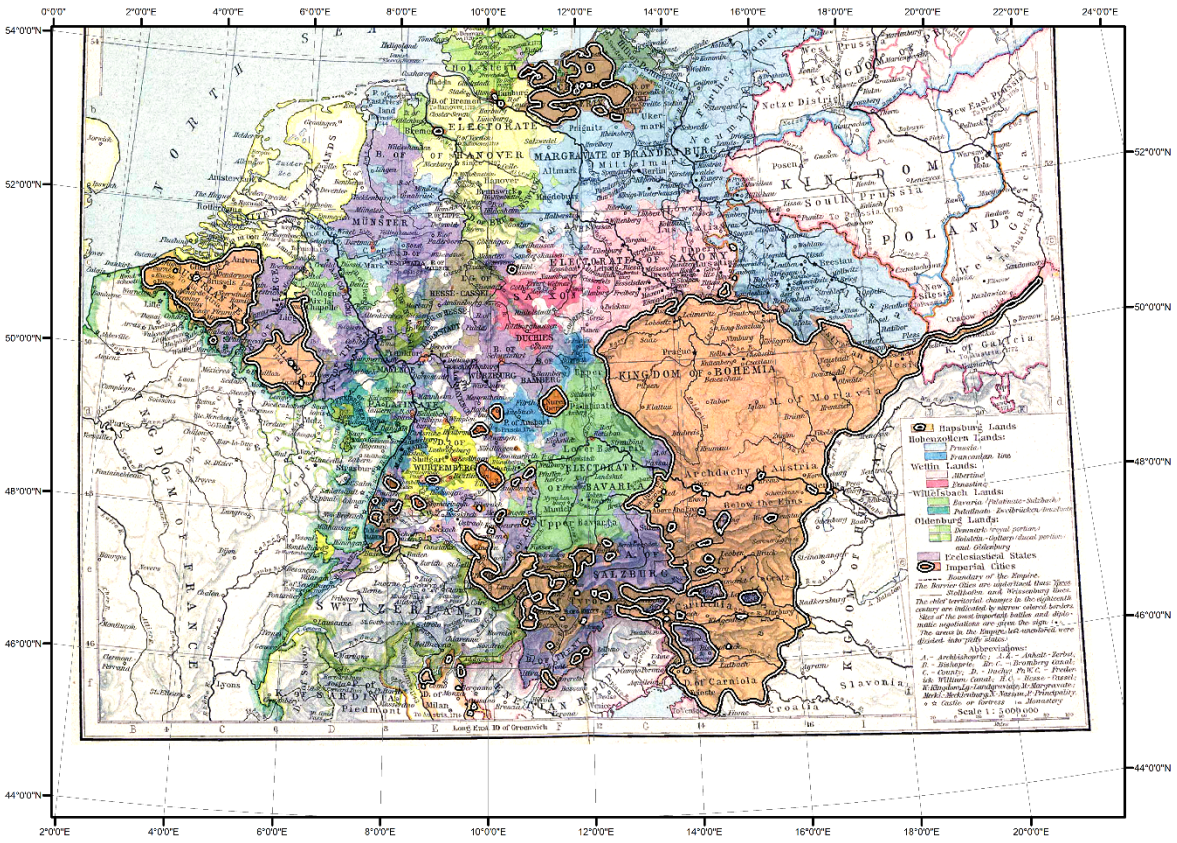
Once the researcher is satisfied with the fit between the map being georeferenced and the reference map, they can either create polygons to represent the features manually, or use the "Classify" tools described above to partially automate the process. To do so, the researcher will have to select some areas of the map that are of the colors that she wants to extract. The algorithm will then select all pixels that have colors similar to those chosen. This will generally create rasters that have holes in them due to the presence of text, which is not the same color as the area of interest. These holes can be filled using for example the *Zonal Statistics* tool with the *Median option*, and a radius large enough to span the holes to be filled (Figure 3).

Figure 3. Classification Tool Results and After Median Smoothing



The following map shows the end result. The black and white line delineates the areas owned by the Hapsburg Monarchy, plus the territories of the Imperial Cities (Figure 4).

Figure 4. The Automatically Generated Borders of the Hapsburg Lands and Imperial Cities



2.3. Reconciling changing unit boundaries

After digitizing different maps, one of the difficulties of using historical data at the geographical level is that both the number and the geographic boundaries can change over time. For the case of the United States, one of the most systematically studied countries, the number and location of the counties changed frequently and significantly since 1790. The difficulty is creating a consistent panel of spatial units, which are consistent over time. The most used source for historical geographical data of the United States is the National Historical Geographic Information System (NHGIS). This website is rich in terms of information, but provides data with geographical identifiers at the time of data collection. The number of counties changed a lot over time. For example, there were only 250 counties in 1790, and the number is higher than 3000 in 2000.

There are different ways of linking counties across all the different census years. Horan and Hargis (1995) published a County Longitudinal template in order to allow for an inter-temporal comparison of fixed county groups between 1840 and 1990. At the time of publication, ARCGIS was

not much used. What the authors did was to aggregate counties, as defined by their boundaries in 1990, into larger units on the basis of earlier historical county boundary configurations. These clusters are larger the further one goes back in time.

The most commonly used approach is the one followed by Eckert et al. (2018). They provide a crosswalk from historical county boundaries in every decade from 1790-2000 to the 1900 county borders. Eckert et al. (2018) used ArcGIS to provide a crosswalk that disaggregate historical counties into 1990 county boundaries based on their land area. They overlaid historical county shape files with the 1990 county shape files and then calculate the share of land of each historical county that forms part of a given recent county. Using such land partitions as weights, researchers can directly aggregate historical information to spatial units as defined by the recent boundaries. The advantage compared to Horan and Hargis (1995) is that the number of counties stay constant over time.

A different approach has been followed by Hornbeck (2010) who started with historical U.S. county boundary files (Carville, Heppen and Otterstrom, 1999) and intersected county borders in later decades with historical county borders. When later counties fall within more than one 1870 county, data for each piece are calculated by multiplying the later county data by the share of its area in the 1870 county. For these later periods, each 1870 county is then assigned the sum of all pieces falling within its area.

2.4. Detailed Data Presentation: A practical guide using GIS

Maps serve many uses in papers.

Presenting summary statistics as maps

One frequent use of maps is to locate the area of interest, particularly when discussing areas that the reader cannot be expected to be immediately familiar with. This is perhaps more frequent in presentations than in papers, since the audience can't easily check reference material in a seminar room.

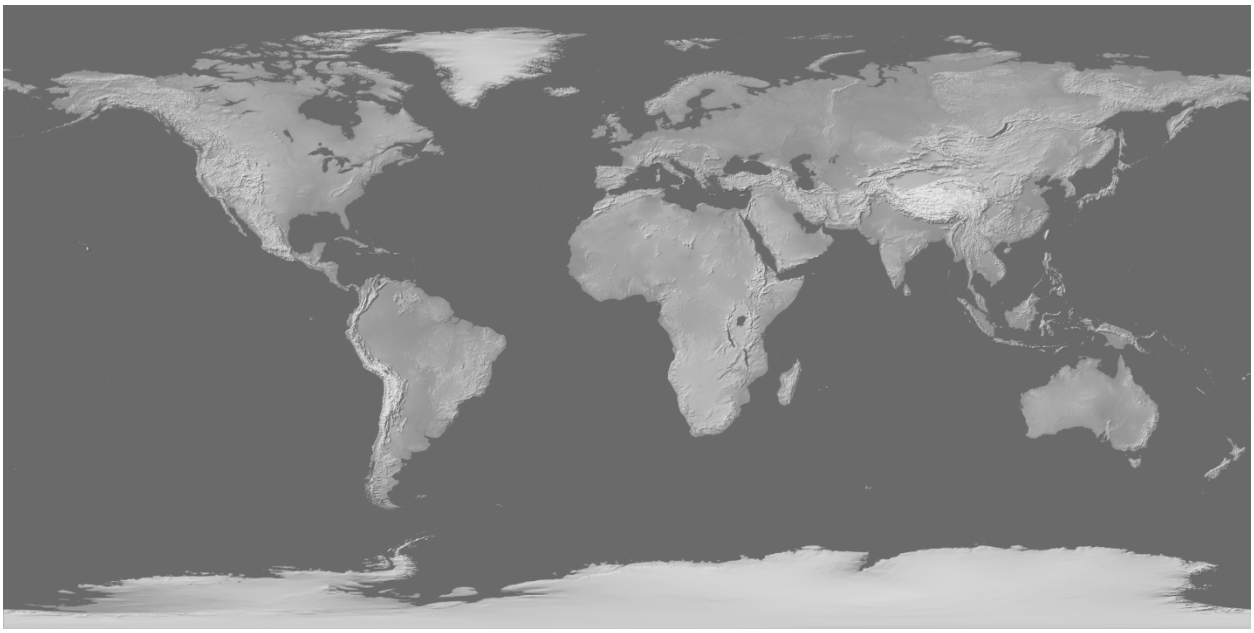
Generally speaking, most economics paper that include historical content could benefit from the inclusion of one summary map. This can help the reader locate the area of the study in the world, familiarize himself with any location names that are important in the story and see visually the relevant sources of geographical variation. If the dataset contains less than a few hundred observations, it is usually feasible to show all the observations at this stage.

If understanding the story requires displaying multiple types of geographic data, it is generally possible to consolidate into a smaller number of combined maps. For example instead of showing

side by side maps of waterways, it might be better to combine them into a single map showing both types of data in different colors.

Where the map is only used to show the geography of the area in question, such as a handful of cities with some associated borders, backgrounds like Gray Earth (pictured below in Figure 5)[4] can be more informative than showing just the coastlines and land borders. The map is ideal for many uses because it exaggerates elevation in flatter areas. This allows e.g. the Hymalayas and the much less prominent Appalachians to be visible on the same map.

Figure 5. The Map of Gray Earth



Showing all of the observations in a dataset as a map is also useful in that it can preempt readers' doubts as to the representativeness of the sample.

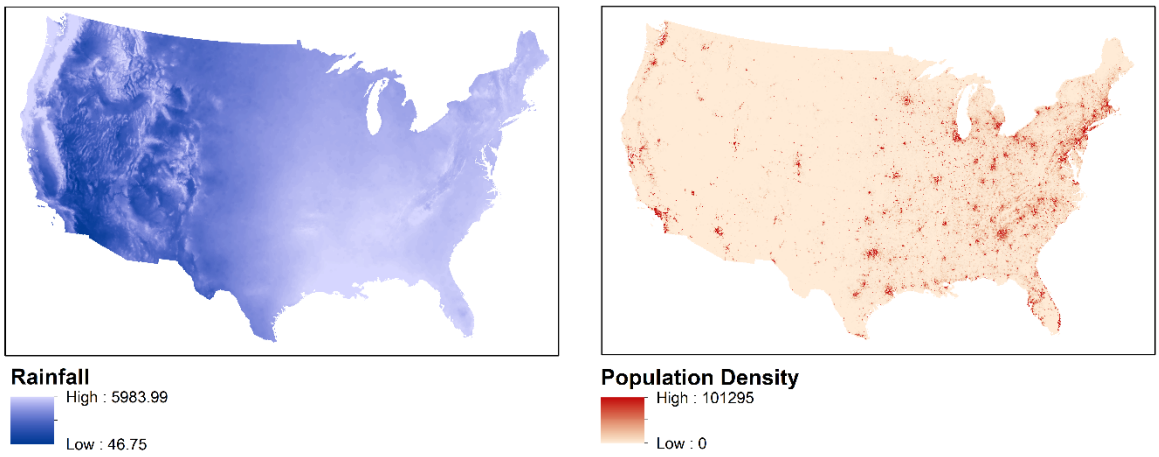
Showing causal effects using maps and integrating maps and scatterplots

Communicating causal relationships through maps is often difficult. This is particularly true in situations where both the independent and dependent variables are most easily represented by a raster, since both can't be simultaneously visible. For example if we wanted to show with a map the relationship between rainfall and population density, the simplest approach would be to show the two

[4] The Grey Earth background, including versions with rivers and sea floor relief, is available from https://www.naturalearthdata.com/downloads/10m-raster-data/10m-gray-earth/

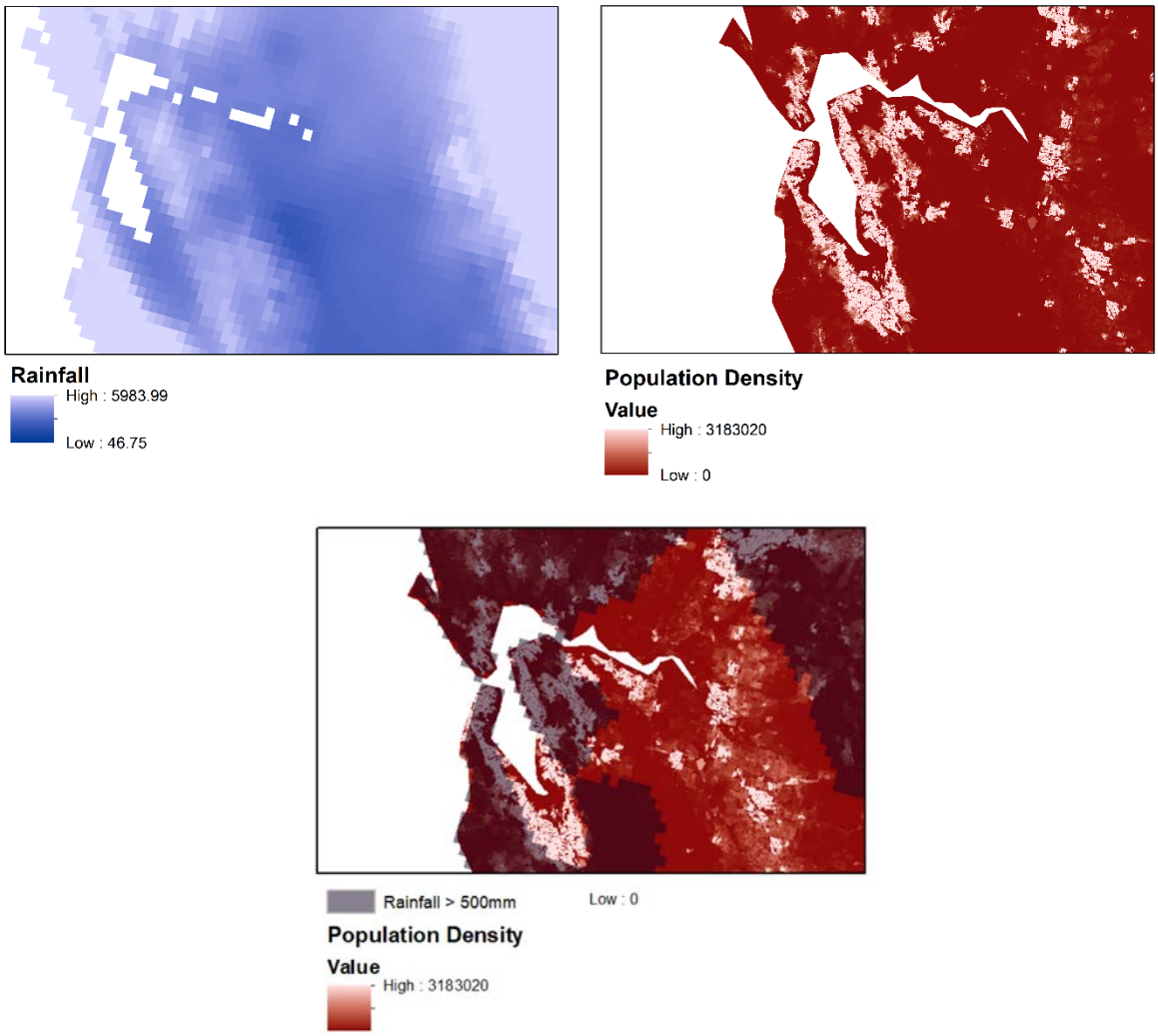
maps side by side (Figure 6). This works reasonably well, since the relationship is generally strong (e.g. in the US the area west of Texas is very dry and very unpopulated until you reach the Pacific Coast.)

Figure 6. Comparison of Rainfall and Population map



However if we did the exercise for only e.g. the San Francisco Bay Area, the relationship might no longer be as obvious, since for example mountain areas will have more rainfall, but less population. As a side-by-side comparison might now be difficult to read, in this case it might be better to transform the rainfall raster into a simple binary raster showing if rainfall is above or below a specific threshold, This could then be overlaid over the vegetation map, to show more clearly how the preponderance populated areas are in the locations with lower precipitations, i.e. population density is lower in the mountains (Figure 7).

Figure 7. Comparison of Rainfall and Population Density for the San Francisco Bay Area



Mapping as a research tool

So far we have assumed that the paper has already been written, and the researcher is simply deciding how to best present her results to the profession. However, making maps can also be a powerful tool at every step of the research process.

For example, while conducting their literature review, researchers can make simple maps, e.g. they can drop pins on every important city mentioned in the text, draw lines along the trade routes mentioned, and start finding and adding potentially important variables to the map. Besides improving their understanding of the historical events being studied, this can help them recognize patterns that might be significant for their research, and will eventually make it easier to transition to the hypothesis testing phase of the project.

Once the results are ready, these same place-markers can be used as needed for any of the maps necessary for the working paper.

2.5. Other geographical data

In this subsections we describe geographical data commonly used in economic research. Unlike old maps that need to be digitized, most of these data are already available and ready to use. We therefore just outline potential issues in using them.

2.5.1. FAO GAEZ

The FAO's Global Agro-Ecological Zones dataset (GAEZ) (Fisher et al, 2002) was created to document the extent to which agriculture is possible, and performed all over the world. It provides freely accessible data in five main areas: land and water resources, agro-climatic resources, suitability and potential yield for 280 crops under a variety of farming technologies, actual production data for the main crops, and information on the gap in yield and production.

As an illustration for the use of these data we can describe the application to the introduction of the potato in Europe (Nunn and Qian, 2011). The idea being that while actual modern production or productivity of potatoes is endogenous to current economic conditions, the soil and climate suitability of different European regions to a crop which had never been cultivated in the Old World before is plausibly exogenous. Similarly, Alesina, Giuliano, and Nunn (2013) used the relative suitability for either plow positive or plow negative crops to instrument for the relative productivity of women's labor. As the FAO GAEZ data covers an enormous variety of crops and cultivation techniques, a wide range of hypothesis can be tested, such as the importance of irrigation in developing state capacity (Bentzen et., al 2017), the role of subsistence versus cash crops (Nunn and Qian, 2011), the implications of improved agricultural technology for structural transformation (Bustos et al., 2016), and agriculture vs pastoralism (Voigtlander and Voth, 2013a).

The model documentation is well written and worth reading, particularly if the GAEZ data is central to the research project. For example, soil suitability for various crops is recorded for low, intermediate, and high input levels. These levels correspond to using traditional methods and seeds, improved methods and seeds, or fully mechanized agriculture.[5]

[5] The documentation tells how these values were calculated: nutrient availability is of utmost importance for low level input farming; nutrient retention capacity is most important for high level inputs; nutrient availability and nutrient retention capacity are considered of equal importance for intermediate level inputs farming; nutrient availability and nutrient retention capacity are strongly related to rooting depth and soil volume available, and oxygen available to roots, excess salts, toxicity and workability are regarded as equally important soil qualities, and the combination of these four soil qualities is best achieved by multiplication of the most limiting rating with the average of the ratings of the remaining three soil qualities.

So if we use the increase in production going from low to high input levels, we will be largely exploiting variation in the ratio of nutrient retention to nutrient availability. If any of these geological factors are connected with our variable of interest in another way, for example greater risk of landslides, higher slope, or mineral deposits, the analysis will naturally be invalid.

The dataset also includes an enormous array of gridded geographic data, such as weather, geology, and soil cover. Furthermore these data have already been harmonized across countries, and derived consistently. They are therefore a natural first place to look for such data when needed. These can be used for data analysis directly, or as derived measures tailored to specific empirical purposes.

One important aspect to consider is that the FAO calculated the crop potentials for the purpose of determining how the world could be fed today, under a variety of scenarios. This means that care must be exercised when using the data in an historical context, as today's crops and techniques are of course very different from those of the past.

The usual workaround is to use the suitability using "low input levels", i.e. only hand power and unimproved variety. While this assumption is often perfectly reasonable, it might become invalid if this modern data is used to proxy for conditions many centuries or even millennia into the past.

Over these extended timeframes, we can expect a certain degree of coevolution between the crops and their farmers, as more efficient cultivation techniques are developed, and the plants themselves adapt to their new conditions through both artificial and natural selection.

For example, imagine we are interested in measuring the effect of the introduction of coffee to a hypothetical Pacific archipelago in the 18th century on present day urbanization. We might be tempted to use the FAO GAEZ productivity level with low inputs as a proxy for the exogenous suitability for coffee before its introduction. Let us imagine that we find that in fact high suitability for coffee production is a significant predictor of urbanization, which we attribute to agglomeration economies and persistence of urban settlements.

However, there is a potential confounding factor. The FAO is calculating present day productivity. Even the traditional cultivars available today are very different from those that would have initially have been brought over on the ship. For example, it could be that the coffee available at the time grew best at higher altitudes, but that due to transportation issue the plantations were initially created close to the harbor used for export. Over time, the local cultivars would have adapted to their new conditions, and farmers would have learned how to coax the most output in the soil and weather they were familiar with. By the time that transportation to the highlands became feasible, coffee cultivation may in fact have become more productive in the lowlands, which is what we measure now.

Therefore, if we had the true soil suitability for coffee in 1700, we would have found the opposite effect --- that greater suitability led to less urbanization. The reason for finding a positive effect was that the presence of a harbor drove both the location of cities, and of the first plantations. In turn, the decision to plant those first coffee trees in a particular location ensured that the local cultivars adapted to those conditions in the long run.

Whether this scenario applies to particular historical instances is something that must be determined by researchers on a case by case basis. What must be avoided are cases in which a) enough time has elapsed since crop introduction to expect significant learning by doing and b) there is a potential factor that could have both influenced the initial decision to plant a particular crop and independently had an effect on the variable of interest.

2.5.2. Rivers

Data on rivers is frequently included in economic history regressions, either as a variable of interest, or as a control. Depending on the historical context, rivers have had remarkably varied impacts on human societies. They are used as sources of irrigation and drinking water, transportation (Bleakly and Lin, 2012), fish, mechanical energy (Caprettini and Voth, forthcoming), and obstacles to be crossed by both traders and armies (Matranga and Nathkov 2019).

Depending on the specific factor being analyzed, different types of data on rivers are most appropriate. If the water source aspect is the primary interest, it is obviously necessary to know how much water is actually in the river. For many regions, the annual water flow of rivers can be obtained only at its mouth, and perhaps at a few of the important forks. If finer grained data is necessary, ground elevation and precipitation data can be combined to calculate the Flow Accumulation, which is the amount of water that should theoretically flow through each pixel of the map given its slope and precipitation totals: While this methodology excludes the role of evaporation and ground water transport, it should in most cases produce results that are at least ordinally consistent, particularly in smaller areas (e.g. the Po's plain in Italy).

From a transportation standpoint, besides the actual river network, the extent of each river's watershed can also be used. Since river transport of bulk goods was generally much cheaper than overland transport by pack or draft animal, the patchwork of watersheds will often be a reasonable description of the areas that can form integrated markets for the major staple crops, or other bulky trade items such as lumber. Historically important examples of such areas are the Po' Valley, the Rhineland, the Yangtze and Yellow River basins, the Mississippi Valley, and the Indus and Ganges Valley.

While the tracks of rivers are generally fairly stable on human timescales, the lower portions of many rivers (such as the Mississippi and the Yellow River) have experienced multiple changes in their tracks during historical times, sometimes swinging by hundreds of miles. It is important therefore to not accept data on river location blindly, but to instead check whether the data reflects the actual location of the rivers at the time of the study.

Data on modern rivers is available from Hydrosheds (Lehner et al., 2008). More detailed data can generally be acquired by the various national geological offices.

2.5.3. Elevation

Elevation is another frequently used variable in economic analysis. Higher elevations areas are generally colder, wetter, and harder to travel in. This has major impacts on the organization of production of the human societies which inhabit it. High altitude areas are also much harder for armies to conduct offensive operations in, giving an advantage to the defenders. In particular, since Nunn and Puga (2012), the de facto standard for measuring the impenetrability of an area is the Terrain Ruggedness Index (TRI) of Riley et al. (1999). The TRI is simply the power mean of the differences in altitude between each cell and the neighboring eight cells. Note that in general, ruggedness calculations will give different results depending on the resolution of the raster being used, therefore consistency is key to obtaining a valid dataset.

2.5.4. Climate data

Many economic history papers have used climate data, as either variable of interest or control. (see Dell et al. 2014 for a review). The two most frequently used variable are temperature and rainfall. These variables have been used as instruments for warfare or agricultural productivity (Miguel et al 2004, Dell et al. 2012, Burke and Leigh 2010)). These two measures are also part of the de facto standard set of controls for regressions in economic history, where available.

If only cross sectional variation is required, it is sometimes appropriate to use modern climatic averages as proxies for past average climatic conditions. This is easier to justify if the units of observation are widely scattered in various climate zones and if the period of interest is not too remote, or part of a known climatic anomaly. For example modern temperature averages could plausibly proxy for past temperature in world national capitals in 1750. They would, however, likely not be a good proxy for the temperatures of different cities on the shore of Lake of Geneva during a specific year of the Little Ice Age.

3. Ethnographic data

A prolific strand of the economics literature has documented a strong persistence of economic outcomes over time, including economic growth, political development and a variety of cultural traits (Putternam and Weil, 2010, Spolaore and Wacziarg, 2013, Michalopoulos and Papaioannou, 2013, Alesina, Giuliano and Nunn, 2013, Giuliano and Nunn, forthcoming,Voigtlaender and Voth, 2012).

The historical dataset containing the most comprehensive information about societal characteristics going back to pre-industrial times is the *Ethnographic Atlas,* assembled by Murdock (1967a). The original dataset is collected at the ethnicity level and contains information on 1,265 ethnic groups, all observed prior to industrialization or European contact. The earliest observation dates are for groups in the Old World where early written evidence is available. For the parts of the world without a written history, the information is from the earliest observers of these cultures. Overall, 23 ethnicities are observed during the seventeenth century or earlier, 16 during the eighteenth century, 310 during the nineteenth century, 876 between 1900 and 1950, and 31 after 1950. For nine ethnicities an exact year is not provided. The dataset contains information on political, economic and cultural traits of societies, and it has been used in various papers to analyze the long-term effects of historical characteristics.[6] The variable in the original database are names *v1*, *v2*, etc. There are 115 variables in the dataset. The type of variables present in the dataset is provided in Table 1.

[6] One issue with the original *Ethnographic Atlas* is that European groups are under-represented, the information about these groups was available but anthropologists were interested in studying ethnicities that could be actually be observed. Giuliano and Nunn (2018) created an extended version of the *Ethnographic Atlas,* adding several additional ethnicities. Specifically, they use three sources. Two sources are data collections, one containing 17 ethnic groups from Eastern Europe (Bondarenko, Kazankov, Khaltourina and Korotayev, 2005) and the other containing information from Siberia (Korotayev, Kazankov, Borinskaya, Khaltourina and Bondarenko, 2004). The third source, the World Ethnographic Sample, was also assembled by Murdock (1967b), and it contains additional 17 European ethnicities not included in the original *Ethnographic Atlas*. The complete sample contains information on 565 ethnic groups. Giuliano and Nunn (2018) construct three versions to link the past to the present. One using only information contained in the *Ethnographic Atlas,* one adding the Eastern European and Siberian identities, and the third one including the 17 additional ethnicities contained in the World Ethnographic Sample.

Table 1. Ethnographic Atlas, variables

Variables	Societal characteristics
	Economic characteristics
v1-v5	Main form of economic subsistence (agriculture, husbandry, fishing, hunting and gathering
v39-v42	Animals and plow cultivation, type of animal husbandry
v44-v65	Sex and age occupational specialization
v79-v88	Type of dwelling
	Political and societal characteristics
v32-v35, v72,v94	Political organization and religion (jurisdictional hierarchy of local community and beyond the local community, the presence of high gods and which types of games the society was practicing), succession to the office of local headman
v66-v69	Class stratification
v70-v71	Type of slavery
v73-v77	Inheritance rules
v90	Political integration
	Cultural characteristics
v6-v27, v43	Various forms of societal organization, mostly related to marriage practices, and type of descent
v36-v38, v78	Societal behavior of boys and girls such as male genital mutilation, post-partum sex taboos, pre-marital sexual behavior
v97-v99	Linguistic Affiliation (by language continent, language phylum, subfamilies)
	Geographical characteristics
v91-v92	Region (Africa, Mediterranean, East Eurasia, Insular Pacific, North and South America) and area within region
v95-v96	Climate
v103-v106	Latitude, longitude
	Miscellaneous
v89, v93	Inclusion in summary atlas volume, ethnographic atlas number
v100-v102	Date (millennium, century, year with century)
v107-v111	Society name (first, second….nine letters)

3.1. Political characteristics

The *Ethnographic Atlas* contains several measures of political and institutional characteristics. One of the most widely used variable is the level of jurisdictional hierarchy beyond the local community. The variable, which is generally interpreted as a measure of political centralization or political sophistication, measures the level of political authority when one moves beyond the local authority. It attributes the value of 0 to groups lacking any form of centralized political organization, 1 to petty chiefdoms, 2 to large paramount chiefdoms/small states and 3 or 4 to large states. For

example, if the local village chief is the highest level of authority, and he or she does not answer to anyone above them, then the variable would take on a value of zero. If above the chief there was a district leader, and above this the paramount chief, then this variable would take on the value of four. Gennaioli e Rainer (2007) documents a strong correlation between the provision of public goods (education, health and infrastructure) in Africa and the centralization of their ethnic groups' precolonial institution.

Michalopoulos and Papaioannou (2013) also use the degree of centralization and study whether it matters for contemporary economic performance, as proxied by satellite images of light density at night. They find a strong correlation, which also holds within pairs of adjacent ethnic homelands, with different legacies of pre-colonial institutions.

Mayshar et a. (2015) argue that hierarchies and states were related to differences in the appropriability of agricultural surplus, rather than to differences in land productivity. They provide empirical evidence supporting the theory that the presence of cereals (a type of crop much easier to appropriate than other crops) is strongly related to political hierarchy, calculated using data from the *Ethnographic Atlas*.

Another measure present in the *Ethnographic Atlas* is the extent of village democracy during the pre-industrial period. This variable reports the traditional form of succession of the local headman (or close equivalent such as clan chief). More specifically, the categories recorded in the data are: patrilineal heir, matrilineal heir; appointment by a higher authority; seniority or age; influence, wealth, or social status; formal consensus (including elections); and informal consensus. A given society has a tradition of democracy if the appointment of the local headman was through either formal consensus or informal consensus. Giuliano and Nunn (2013) show that having experienced local democracy in the past makes it more likely to develop democratic institutions today. The evidence provided by the authors suggest that the persistence comes from the development of more supportive beliefs of national democracy today.

3.2. Economic characteristics

Alesina, Giuliano and Nunn (2013) test the hypothesis that traditional agricultural practices influenced the historical division of labor and the evolution of gender norms. They find that societies that traditionally practiced plough agriculture in the past, today have less equal gender norms, measured using reported gender-role attitudes and female participation in the workplace, politics and entrepreneurial activities. They use a measure from the *Ethnographic Atlas* to calculate the historical

reliance on plough agriculture, and proxy plough agriculture using the agroclimatic suitability for crops that benefit from the plough. The data is provided by the FAO GAEZ project (see Section 2.5.1).

Becker (2019) studies how reliance on pastoralism has been relevant in determining restrictions on women's sexuality, such as female genital cutting, restrictions on women's freedom of mobility, and norms about their sexual behavior. For pastoralists it was hard to check paternity due to extended periods of male absence from the settlement. Using within-country variation across 500,000 women, Becker (2019) shows that women coming from pastoral societies are more likely to have undergone infibulation, adhere to more restrictive norms about women's promiscuity, are more restricted in their freedom of mobility.

She constructs a measure on historical dependence on pastoralism, by combining two variables from the *Ethnographic Atlas:* the degree to which a society depended on animal husbandry (from 0 to 100%) combined with the predominant type of animal in that specific society. For the predominant animals she defines a dummy if the societies has a herding animal (sheep, cattle, horses, reindeer, alpacas, or camels) and 0 otherwise (pigs, dogs or poultry or not animals at all).

Another commonly used variable is the measure of complexity of settlements. Ethnic groups are classified as belonging to categories going from nomadic to having complex settlements.[7] This variable has been used by a number of scholars as a measure of traditional economic development by assigning each non-missing category an integer value from 1 to 8 (Alesina, Giuliano and Nunn, 2013, Giuliano and Nunn, forthcoming, and Michalopoulos and Papaioannou, 2013).

3.3. Cultural characteristics

Enke (2019) studies the long-term effect of different societal organization, based on how tight the kinship structure was. He finds that societies with a historically tight kinship structure regulate behavior through communal moral values, revenge taking, emotions of external shame, and notions of purity and disgust. In loose kinship societies, cooperation is enforced through universal moral values, internalized guilt, altruistic punishment, and the appearance of moralizing religions.

He constructs an index measuring the extent to which people in preindustrial societies were embedded in large, interconnected extended family networks. He follows Henrich (2020) and relies on information on local family structures and descent systems. More specifically he identifies two societal characteristics in the *Ethnographic Atlas* that reflect strong extended family networks: the

[7] Categories in between are: Semi-nomadic, semi-sedentary, compact but not permanent settlements, neighborhoods of dispersed family homesteads, separate hamlets forming a single community, compact and relatively permanent settlements.

presence of extended family systems and post-marital residence with parents (family structure) and the presence of lineages and localized clans (descent systems).

For family structure, he creates a variable that equals 1 if the domestic organization is around independent nuclear families and 0 otherwise. The idea is that living in extended family systems is an indication of the presence of large interconnected family networks. He also creates a variable equal to 1 if the wife is expected to move in with the husband's group or viceversa, and 0 otherwise. Strong kinship are indicated by norms that prescribe residence with the husband or wife group.

On the descent systems, the distinction is between unilineal or bilateral descent, and between segmented communities and localized clans. Unilineal descent systems track descent primarily through one line (maternal or paternal) as opposed to through both lines, and induce strong and cohesive in-groups. For segmented communities and localized clans, he defines a variable equal to 1 if people are part of localized clans that live as segmented communities and 0 otherwise. Clans are important to build very large extended family networks because they allow very distantly related people to feel connected.

Another cultural practice that has received considerable attention in recent research is the practice of bride price, which is a transfer of money and/or other valuable assets that is made at marriage from the groom and/or his parents to the bride's parents. The importance of this tradition for female educational investments has recently been studied by Ashraf, Bau, Nunn and Voena (2020), Corno and Voena (2016) and Corno, Hildrebrandts and Voena (2017).

The *Ethnographic Atlas* categorizes the marriage customs of pre-industrial societies into the following groups: Bride price, which is also known as bride wealth and is a transfer of a substantial consideration in the form of goods, livestock, or money from the groom or his relative to the kinsmen of the bride; token bride price is a small or symbolic payment only; bride service, which is a substantive material consideration in which the principal element consists of labor or other services rendered by the groom to the bride's kinsmen; gift exchange, which is a reciprocal exchange of gifts of substantial value between the relatives of the bride and groom, or a continuing exchange of good and services in approximately equal amounts between the groom or his kinsmen and the bride's relatives; female relative exchange, which is a transfer of a sister or other female relative of the groom in exchange for the bride; dowry, which is a transfer of a substantial amount of property from the bride's relative to the bride, the groom, or the kinsmen of the latter; and no significant consideration, which is an absence of any significant consideration, or giving of bridal gifts only.

3.4. Connecting the past to the present

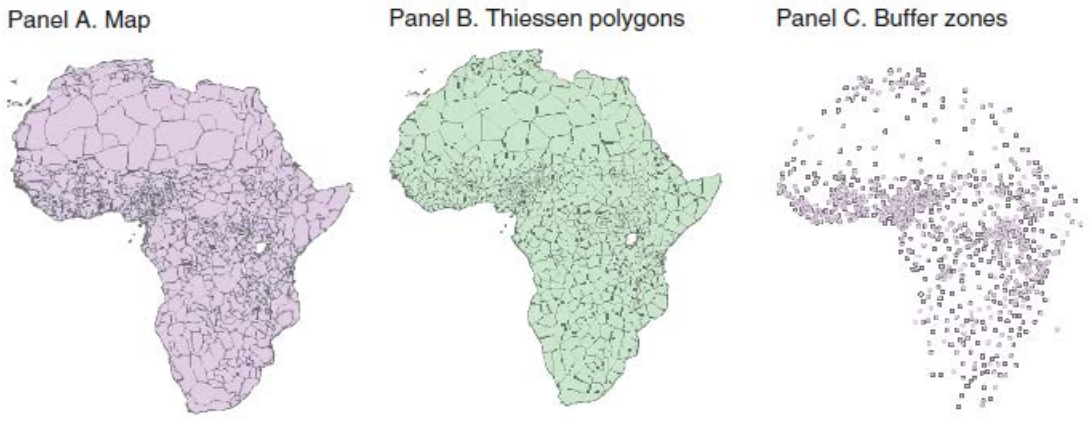
There are various ways in which researchers could link the historical characteristics of pre-industrial societies to current outcomes. We review the main ones used in literature.

Linking historical characteristics using the historical geographical location

The *Ethnographic Atlas* records the centroid of each society (longitude and latitude in degree). The question is how to associate the historical characteristics to current territory. One possibility is to use a circular "buffer zone" of various distances around the centroid. For example, Mayshar et al. (2015) use a circle of 20 mile radius around the centroid. This approach is shown in Figure 8A.

There are several potential problems with this approach. If the buffer zones are chosen too large, they overlap, making it difficult to allocate territory to mutually exclusive ethnic groups. If the buffer zones are too small, they will be a poor approximation of the actual boundaries. An alternative approach (Alsan, 2015) is to construct Thiessen polygons, which more nearly approximate boundaries. The starting point to construct the Thiessen polygons are the centroids of the ethnic groups as reported in the *Ethnographic Atlas*. For a set of points *S* in Euclidean space, a Thiessen polygon (also known as a Voronoi diagram) is one such that every point in the constructed polygon is closer to one such point *p* than to any other point in *S*. Within Africa, Thiessen polygons have a higher correlation with the boundaries of ethnicities (that for this continent was provided by Murdock) than the buffer zone technique.

Figure 8. Thiessen Polygons and Buffer Zones (from Alsan, 2015)



Linking historical characteristics using the current distribution of languages or ethnicities across the world

Both the buffer zone and the Thiessen polygons techniques are valid if the current location of the historical society is very similar to the current location of the society today. The further back into the past one goes, the more the economic history of a given place tends to diverge from the economic history of the people who currently live there. For example, the ethnicities reported in the *Ethnographic Atlas* for the case of the United States were Native-American populations, mostly involved in hunting, fishing and horticultural communities and organized into small, pre-state political units. By contrast, if one looks at the actual composition of the population in the United States a large fraction comes from ethnicities that lived in settled agricultural societies organized in large states. For the case of the United States, therefore, if one would link current outcomes using the geographical method this would lead to a misleading representation as the current distribution of population is very different than the one represented in the *Ethnographic Atlas*.

A better approach has been to link the historical information of the *Ethnographic Atlas* to the current population distribution. We describe in detail the approach followed by Giuliano ad Nunn (2018).

To link the historical ethnicities with the current distribution of ethnicities, Giuliano and Nunn (2018) use the sixteenth edition of the *Ethnologue: Languages of the World* (Gordon, 2009) a data source that maps the current geographic distribution of over 7,000 different languages and dialects, which were manually matched to one of the ethnic groups from the ethnographic data sources.

The *Ethnologue* provides a shape file that divides the world's land into polygons, with each polygon indicating the location of a specific language/dialect as of the data of publication. The *Ethnologue* shapefile is combined with data on the global distribution of the world's population taken from the Landscan 2007 database. The source reports estimates of the world's population in 2007 for 30 arc-second by 30 arc-second (roughly 1 km by 1 km) grid-cells globally. Combining these two sources of data provides an estimate of the distribution of populations' mother-tongue and, hence, the ancestral characteristics of populations across the globe today at a 1-km resolution. By combining these data sources, the authors construct country-level estimates of the average ancestral characteristics of populations for each modern country. The procedure can also be used to construct average ancestral characteristics at the subnational level.

We use the example of the authors to illustrate their procedure. In Alesina, Giuliano and Nunn (2013) the authors' research question was to look at the historical persistence of plough agriculture on current female labor force participation.

The first step is to look at the distribution of languages in a given country. Figure 9A shows a map of Ethiopia with the land inhabited by different ethnic groups, i.e. groups speaking different languages. Each polygon represents the approximate borders of a group as found in the *Ethnologue*. The map also shows the *Landscan* estimate of the population of each cell within the country. A darker share indicates greater population.

The second step in their procedure is to manually match each of the 7,612 *Ethnologue* language groups to one of the 1,265 *Ethnographic Atlas* Ethnic groups. From the *Ethnographic Atlas,* they then know whether a given ethnic group used the plough historically. Figure 9B shows whether the ancestors of a given language group engaged in plough agriculture or not.

The third step consists in overlaying political districts and construct, using the *Landscan* population data (Figure 9C), an finally estimate for each district (or country) the fraction of the population living that descends from ancestors that traditionally engaged in plough agriculture (Figure 9D).

Figure 9A, Map of Ethiopia from Alesina et al. (2013), Ethnologue

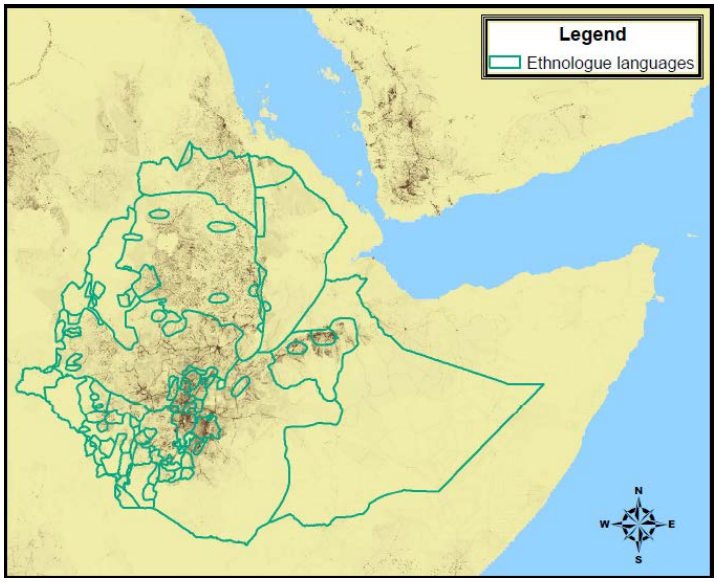


Figure 9B, Map of Ethiopia with Plough/No Plough

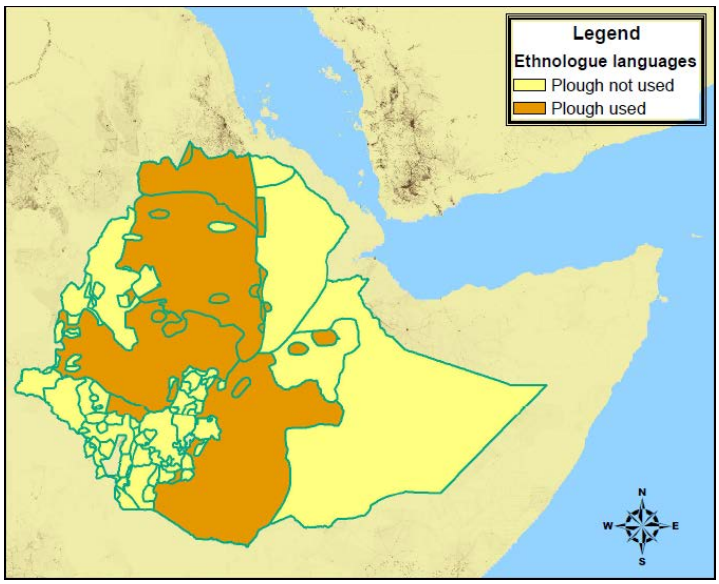


Figure 9C, Map of Ethiopia with Political Districts

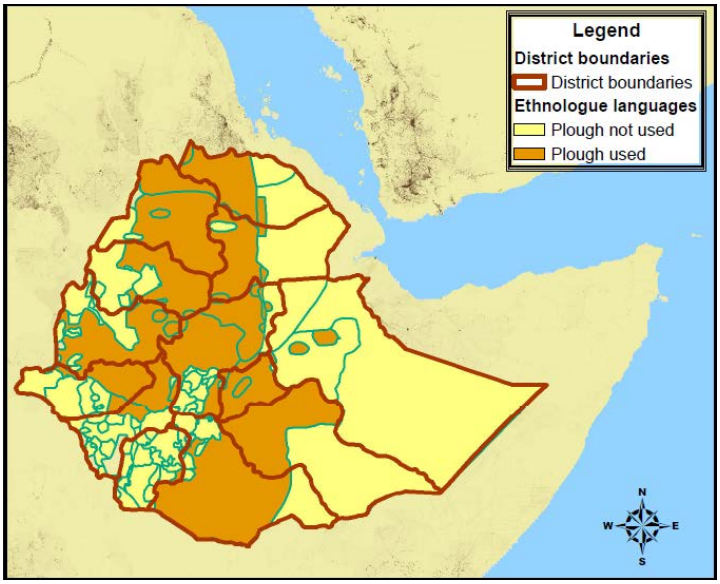
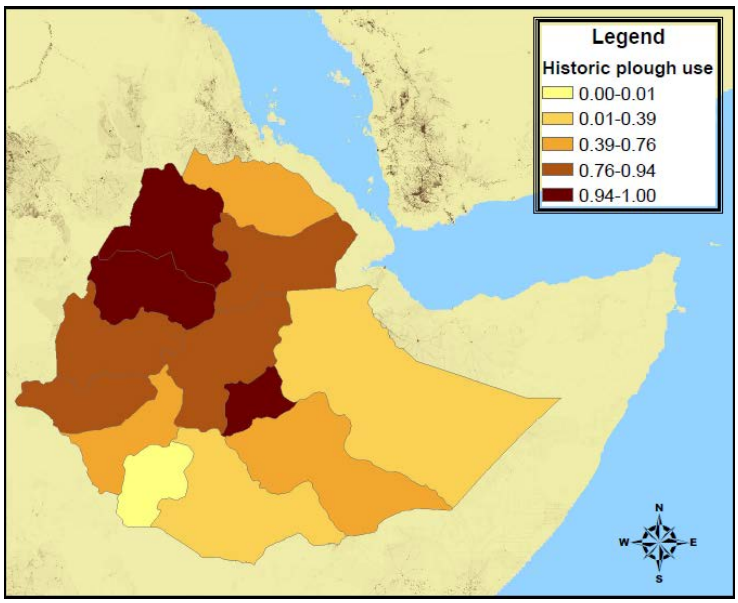


Figure 9D, Map of Ethiopia, Plough Use at the District Level



While the use of the *Ethnologue* constitute an improvement compared to the matching using a geographical radius, it has its own drawbacks. The first one is that the information is missing for some part of the world. This is due to uncertainty or a lack of information about the boundaries of some language groups, a problem particularly pronounced for Latin America.

For those countries or districts in which the information is missing, a potential strategy is to construct alternative measures, by making different assumptions regarding the missing language data. The first is to assume that the inhabitants of unclassified territories all speak the official national language of the country. The second strategy is to impute the missing data using information on the spatial distribution of the ethnic groups taken from the *Geo-Referencing of Ethnic Group (GREG)* database (Weidmann et al. 2010). Like the *Ethnologue* , the *GREG* database provides a shape file that divides the world's land into polygons, with each polygon indicating the location of a specific ethnicity. The shortcoming of the GREG database is that ethnic groups are much less finely identified relative to the *Ethnologue* database.[8] The big advantage if this dataset is that provides a better distribution of ethnicities for Latin America, where the *Ethnologue* provide only the official language.

Linking historical characteristic using a migration matrix

[8] The GREG database identifies 1,364 ethnic groups, while the *Ethnologue* identifies 7,612 language groups.

An alternative approach to account for migration is to follow Putternam and Weil (2010). The authors use this measure to re-examine the hypothesis that early development of agrarian societies (and states, their socio-political correlate) conferred development advantages that remain relevant today. The authors formally address the issue of migration by constructing a matrix detailing the year-1500 origins of the current population of almost every country in the world. Before them, Acemoglu, Johnson and Robinson (2001, 2002) calculated the share of the population that is of European descent for 1900 and 1975. Putternam and Weil (2010) was an improvement because they broke down ancestor populations much more finely than Europeans and non-Europeans.

Putterman and Weil (2010) use 1500 as a rough starting point for the era of European colonization of other continents. To estimate the proportion of the ancestors of today's inhabitants they use a wide range of secondary sources. A crucial challenge in their methodology is the attribution of mixed populations (for instance mestizos) to their original source countries. The authors, whenever possible, use genetic evidence as the basis for dividing the ancestry of modern mixed groups that account for large fractions of their country's population. In cases where genetic evidence on the ancestry of mixed groups was not available, the authors rely on textual accounts and generalizations from countries with similar histories for which genetic data were available.

The authors construct a matrix of migration since 1500. The matrix has 165 rows, each for a present-day country, and 172 columns (the same 165 countries plus seven other source countries with current populations of less than one half million). Using the authors' example: Malaysia has five entries, corresponding to the five source countries for the current Malaysian population: Malaysia (0.6), China (0.27P, India (0.075), Indonesia (0.04) and the Philippines (0.025).

Linking historical characteristics using individual level data at the ethnicity level

The final way of linking current outcomes to ancestral characteristics is to use individual level information on the ethnicity or language spoken by the person. Each current ethnicity/language can be manually matched to the historical ethnicity name. This procedure has the advantage that one can control for county and also finer geographical fixed effects. The direct link has been used in different papers, including Alesina, Giuliano and Nunn (2013) and Becker (2019).

4. Censuses

4.1. The use of complete count population censuses

Research in economic history has advanced very recently with the use of historical complete count population censuses, mostly for the United States. Besides the enormous sample size, an

important advantage of these datasets is that they provide individuals' names, allowing researchers to link information across time, creating large panel datasets, which can then be used to answer questions on intergenerational mobility or immigrant assimilations, often with illuminating results. For example, cross-sectional work had concluded that immigrants to the United States started from lower-paid positions than U.S. born workers but converged over time. Abramitzky et al. (2014) instead used individual-linked, and find that even recent immigrants to the U.S. earned just as much as the native born, and further that their wages then grew at similar rates.

Linking historical data across censuses present various challenges. For modern data, administrative datasets contain social security numbers (SSN), allowing researchers to link individuals using SSN, names and place of birth. Social security numbers are however not present in historical records. Without SSN linking individuals with census frequency is extremely challenging especially for individuals with common names. Various methods have been developed to improve the matching. Abramitzy et al. (2019b) provide a detailed overview. Here we briefly describe each them, together with their advantages and disadvantages.

One approach is to try to match individuals by using their name and last name directly, as initially done by Ferrie (1996), and later improved by Abramitzky, Boustan and Erikson (2012, 2014, 2019a). The second is a machine-learning approached developed by Feigenbaum (2016). The third one is a fully automated probabilistic algorithm (Abramizky, Mill and Perez (2019c)).

4.2. Linking historical information using names

This method has been followed by Ferrie (1996) and more recently by Abramitzky, Boustan and Eriksson (2012, 2014, 2019a). The algorithm consists in uniquely identifying a name in a given Census Year. In their case the name is identified using first and last name, place and date of birth. The information is further restricted to individuals who are unique. They then search for these unique individuals in the following Census-year. If there is unique match, this is considered a match. If there are multiple matches, this information is discarded. If there are no matches, the algorithm searches for matches within two years of reported birth (one year before or one year after) first, and if this is unsuccessful within 4 years (two years before and two years after). Only unique matches are accepted. If there are no unique matches the observation is discarded. The procedure is then repeated for each record in the second database. Finally the intersection of the two matched samples is taken.[9]

[9] The authors also propose additional strategies to improve the match. For example, to limit the possibility of misspelling, they use the New York State Identification and Intelligence System (NYSIIS) standardized names, rather than actual names (this system standardize names based on their pronunciation). Another

Abramitzky, Mill and Perez (2019c) developed a probabilistic approach. For each observation in a given dataset, they identify a set of potential matches in the second dataset. The potential matches are identified by looking for individuals with the same place of birth, the same estimated year of birth (plus/minus five years), and the same first letter in their first and last names. For each pair of potential matches, they compute a measure of similarity in the reported year of birth and name, where they use the Jaro-Winkler score for the first and last names. Similarity in age is calculated using the absolute value of the difference in reported years of birth. They then look at the combination of distances in reported names and ages, which is a good approximation of the probability that both records belong to the same individual.

4.3. Machine learning algorithm

Feigenbaum (2016) uses a machine learning algorithm which train the algorithm using hand linked data. The procedure has the following steps. First, for each observation in dataset A, he identifies a set of potential matches in dataset B (see Feigenbaum for the specific rule followed). Second, a human researchers helps with the construction of a training dataset on a small share of the possible links. This training dataset is used for the matching algorithm, by using a probit model, taking the value of 1 if the human researcher matched these records and 0 if not. The fitted model is then applied to the full data and generate a predicted probability of being a match for each pair of records in A and B (see Feigenbaum, 2016, for details.)

Abramitksy et al. (2019c) describe in details the accuracy of the different methods and the trade-off that they have between type I and type II errors, where the trade-off is between low discrepancy rates (their false positive) at the cost of having a relatively low true match rate. The other option is to have higher true match rates but also higher discrepancy rates. Finally, the authors also discuss how different automated linking methods can affect inferences.

5. Other historical data

In this section we describe some other historical data commonly used by economic historians.

5.1. Military history

Understanding the causes and consequences of wars, and their outcomes, has been one of the perennial objective of historiography since its very inception. A number of economic historians

possibility is to use the Jaro-Winkler string distance adjustment which gives a measure of the similarity of two strings, placing more weight on characters at the beginning of the strings.

have contributed to this area, some focusing on the causes (Gennaioli and Voth, 2015), and others on the effects (Voigtländer and Voth, 2013).

If the number of conflicts is needed, one approach is to digitize the battles catalogued in reference material, such as Chandler (1987) or Jacques (2006). In more localized settings, specialized sources should be sought where possible. For example, in their paper on the long run effects of Sherman's March to The Sea in the US Civil War, Feigenbaum et al. (2019) digitized maps of the US Army's routes of advance prepared by the US Department of War. Indeed the US government edited an entire War of the Rebellion Atlas, comprising over 1,000 such maps, detailing the war at both the strategic and tactical scale.

In reading contemporaneous accounts of specific campaigns, or the memoirs of their combatants, researchers should always maintain some level of skepticism. The participants of every conflict are keen to embellish their successes and hide or dismiss their defeats. Before using such data, researchers should ask themselves whether the source would have had an interest to misreport in ways that could be significant to the analysis, and whether it would have been likely to be caught if they had tried to do so.

5.2. Transportation networks

The economic history of transportation was of course one of the first fields to receive modern cliometric techniques (Fogel 1964). For more recent times, data is often already available digitized, or at least contemporaneous maps are available to be scanned and geocoded. For example, the US Military has mapped enormous areas of the world at a high level of detail, and many of the maps from the WWII era are freely available online from the Perry Castadena Library Map Collection at the University of Texas, Austin. All these maps include many features of potential economic interest, such as roads (often separated by surface type, canals, railways, ports and airfields.) For many important European and North American areas, maps of similar quality were being made already in the 19th century, but for earlier periods researchers will typically have to rely on maps that were prepared well after the fact, usually by historians or archaeologists basing themselves on the lay of the land, archaeological remains, narrative knowledge of which cities were connected, and other contextual factors.

Often the researcher has to guess about the historical nature of roads. For example, if the compiler knew that cities A and B were connected by a road, but did not know its exact lay, she might guess that it probably followed more or less the track of the modern road. This guess would be fairly

accurate only when the location of transportation routes are extremely persistent, for example, the Roman road network is known with very high precision for the vast majority of its extent.

6. Conclusions

The use of historical data has become a standard tool in economics. This chapter describes the main sources of data used by economic historians, specifically looking at geographical data, ethnographic data and censuses. For each group, we describe where to obtain these data, how to use or manipulate them and the main methodologically advances which allow economists to overcome or minimize the problems in using them. We finally discuss a variety of issues that they raise, such as the constant change in national and administrative borders; the reshuffling of ethnic groups due to migration, colonialism, natural disasters, and many other forces.

References

Abramitzsky, R., Boustan, L. and K. Eriksson, 2012, "Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration", *American Economic Review,* 102 (5), 1832-1856

Abramitsky, R., Boustan, L. and K. Eriksson, 2014, "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration", *Journal of Political Economy,* 112 (3), 467-717

Abramitzky, R., Boustan, L. and K. Eriksson, 2019a, "To the New World and Back Again: Return Migrants in the Age of Mass Migration", *Industrial and Labor Relations Review,* 72 (20), 300-322

Abramitzky, R., Boustan, L., Erikson, K., Feigenbaum, J. and S. Perez, 2019b, "Automated Linking of Historical Data", Stanford University, mimeo

Abramizky, R., Mill, R. and S. Perez, 2019c, "Linking Individuals Across Historical Sources: A Fully Automated Approach", *Historical Methods.*

Acemoglu, D., Johnson, S. and J. Robinson, 2001, "The Colonial Origins of Comparative Development: An Empirical Investigation", *American Economic Review,* 91(5), 1369-1401

Acemoglu, D., Johnson, S. and J. Robinson, 2002, "Reversal of Fortunes: Geography and Institutions in the Making of the Modern World Income Distribution", *Quarterly Journal of Economics,* 117 (4), 1231-1294.

Alesina, A., P. Giuliano and N. Nunn, 2013, "On the Origins of Gender Roles: Women and the Plough", *Quarterly Journal of Economics,* 128 (2), 469-530.

Alsan, M., 2015, "The Effect of the TseTse Fly on African Development", *American Economic Review,* 105 (1), 382-410

Ashraf, N., Bau, N., Nunn., N. and A. Voena, 2020, "Bride Price and Female Education", *Journal of Political Economy*, 128 (2): 591-641

Becker, A., 2019, "On the Economic Origins of Restrictions on Women's Sexuality", Harvard mimeo

Bentzen, J.S., Kaarsen, N. and Wingender, A.M., 2017, "Irrigation and Autocracy", *Journal of the European Economic Association*, 15 (1), 1-53

Bleakley, Hoyt and Jeffrey Lin, 2012, "Portage and path dependence," *The Quarterly Journal of Economics*, 127 (2): 587-644.

Bondarenko, D., Kazankov, A., Khaltourina, D. and A. Korotayev, 2005, "Ethnographic Atlas XXI: People of Eastermost Europe", *Ethnology,* 44. 261-289

Burke, Paul J., and Andrew Leigh. 2010. "Do Output Contractions Trigger Democratic Change?" American Economic Journal: Macroeconomics 2 (4): 124–57.

Bustos, P., Caprettini, B. and J. Ponticelli, "Agricultural Productivity and Structural Transformation: Evidence from Brazil", *American Economic Review,* 106 (6), 1320-1365.

Caprettini, Bruno and Hans-Joachim Voth. "Rage against the machines: labor-saving technology and unrest in England, 1830-32."*American Economic Review: Insights,* forthcoming.

Card, D. and A. Krueger, 1994, "Minimum wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania", *American Economic Review,* 84 (4), 772-793.

Carville, E., Heppen, J. and S. Otterstrom,1999, *HUSCO 1790-1999: Historical United States County Boundary Files,* Baton Rouge: Louisiana State University

Chandler, David G., 1987, *Dictionary of Battles: The World's Key Battles from 405 BC to Today*, Ebury Press.

Corno, L. and A. Voena, 2016, "Selling daughters: Age of marriage, income shocks, and the bride price tradition", IFS W16/08, Institute for Fiscal Studies

Corno, L., Hildebrandt, N. and A. Voena, 2017, "Age of Marriage, weather Shocks, and the Direction of Marriage Payments", NBER WP 23604.

Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken. 2012, "Temperature Shocks and Economic Growth: Evidence from the Last Half Century," *American Economic Journal: Macroeconomics* 4 (3): 66–95

Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken, 2014, "What do we learn from the weather? The new climate-economy literature," *Journal of Economic Literature*, 52 (3): 740-98.

Eckert, F., Gvirtz, A. and M. Peters, 2018, "A Consistent County-Level Crosswalk for US Spatial Data Since 1790", Yale University, mimeo.

Enke, B., 2019, "Kinship, Cooperation, and the Evolution of Moral Systems", *Quarterly Journal of Economics,* 134 (2), 953-1019.

Feigenbaum, J. J., 2016, "Automated Census Record Linking: A Machine Learning Approach", WP https://open.bu.edu/handle/2144/27526.

Feigenbaum, J. J., Lee, J. and F. Mezzanotti, "Capital Destruction and Economic Growth: The Effects of Sherman's March, 1850-1920, NBER WP 25392.

Ferrie, J. P., 1996, "A new sample of males linked from the public use microdata sample of the 1850 US federal census of population to the 1860 US federal census manuscript schedules", *Historical Methods,* 29 (4), 141-156.

Fischer, Gunther, Harrij van Nelthuizen, Mahendra Shah, and Freddy Nachtergaele, Global Agro-Ecological Assessment for Agriculture in the 21st Century: Methodology and Results (Rome: Food and Agriculture Organization of the United Nations, 2002)

Fogel, Robert William, 1964, *Railroads and American economic growth*, Baltimore: Johns Hopkins Press, 1964.

Gennaioli, N. and I. Reiner, 2007, "The modern impact of precolonial centralization in Africa", *Journal of Economic Growth,* 12 (3), 185-234.

Gennaioli, N. and H.J. Voth, 2015, "State Capacity and Military Conflict", *The Review of Economic Studies,* 82 (4), 1409-1448.

Giuliano, P. and N. Nunn, 2013, "The Transmission of Democracy: From the Village to the Nation-State", *American Economic Review: Papers and Proceedings,* 103 (3): 86-92.

Giuliano, P. and Nunn, "Understanding Cultural Persistence and Change", *The Review of Economic Studies,* forthcoming.

Giuliano, P. and N. Nunn, 2018, "Ancestral Characteristics of Modern Populations", *Economic History of Developing Regions,* 33 (1): 1-17

Gordon, R.G., 2009, "Ethnologue: Languages of the World", 16th ed. SIL International, Dallas.

Henrich, J., 2020, *The WEIRDest People in the World: How the West Became Psychologically Peculiar and Particularly Prosperous,* Princeton, Nj, Princeton University Press

Horan, P. M. and P. G. Hargis, 1995, "County Longitudinal Template, 1840-1990", *Ann Arbor, MI: Inter-University Consortium for Political and Social Research*

Hornbeck, R., 2010, "Barbed Wire: Property Rights and Agricultural Development", *Quarterly Journal of Economics,* 125 (2): 767-810.

Jaques, T., 2007, *Dictionary of Battles and Sieges*: Vol. 3: P-Z., Greenwood Publishing Group.

Korotayev, D., Kazankov, A. Borinskaya, S., Khaltourina, D. and D. Bondarenko, 2004, "Ethnographic Atlas XXX: People of Siberia", *Ethnology,* 43, 83-92

Lehner, B., Verdin, K., and Jarvis, A, 2008, "New global hydrography derived from spaceborne elevation data," *Eos, Transactions, American Geophysical Union*, 89(10): 93–94.

Matranga, Andrea and Timur Nathkov, 2019, "All Along the Watchtower: Linear Defenses and the Introduction of Serfdom in Russia", Chapman University mimeo

Mayshar, J., Moav, O., Neeman, Z. and L. Pascali, 2015, "Cereals, Appropriability and Hierarchy", CEPR WP 10742

Michalopoulos and Papaioannou, 2013, "Pre-colonial Ethnic Institutions and Contemporary African Development", *Econometrica,* 81, 113-152.

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. "Economic Shocks and Civil Conflict: An Instrumental Variables Approach." Journal of Political Economy 112 (4): 725–53.

Murdock, G. P., 1967a, *Ethnographic Atlas,* University of Pittsburgh Press.

Murdock, G. P., 1967b, "World Ethnographic Sample", *American Anthropologist,* 59, 664-687.

Nunn, N. and D. Puga, 2012, "Ruggedness: The Blessing of Bad Geography in Africa", *Review of Economics and Statistics,* 94 (1): 20-36

Nunn, N. and N. Qian, 2011, "The Potato's Contribution to Population and Urbanization: Evidence from a Historical Experiment", *Quarterly Journal of Economics,* 126 (2), 593-660.

Putternam, L. and D. Weil, 2010, "Post-1500 Population Flows and the Long-Run determinants of Economic Growth and Inequality", *Quarterly Journal of Economics,* 125, 1627-1682

Riley, S. J., De Gloria, S. D. and R. Elliot, 1999, "A Terrain Ruggedness Index That Quantifies Topographic Heterogeneity", *Intermountain Journal of Science,* 5, 1-4, 23-27

Spolaore and Wacziarg, 2013, "How Deep Are the Roots of Economic Development?", *Journal of Economic Literature,* 51, 325-369.

Voigtländer, N. and J. Voth, 2012, "Persecution Perpetuated: The Medieval Origins of Anti-Semitic Violence in Nazi-Germany", *Quarterly Journal of Economics,* 127, 1339-1392.

Voigtländer, Nico, and Hans-Joachim Voth, 2013a, "How the West "Invented" Fertility Restriction." *American Economic Review*, 103 (6): 2227-64.

Voigtländer, Nico and Hans-Joachim Voth, 2013b, "The Three Horsemen of Riches: Plague, War, and Urbanization in Early Modern Europe," *The Review of Economic Studies*, 80 (2), 774–811.

Weidmann, Nils B., Jan Ketil Rod, and Lars-Erik Cederman, 2010, "Representing Ethnic Groups in Space: A New Dataset," *Journal of Peace Research*, 47 (4), 491–499