scientific data

Check for updates

OPEN

DATA DESCRIPTOR

HISTORECO: Historical Spanish transition database on climate, geography and economics of the 20th-21st century

Guillermo Rodríguez-López[1], Ana Serrano[1], Miguel Martín-Retortillo[2] & Ignacio Cazcarro[3,4] ✉

One of the major difficulties in the social sciences is obtaining comparable data from different sources. Especially, for small-scale data, such as municipalities within a country, and even more when the variables belong to different fields of knowledge. To avoid these problems, we present a database containing 45 geographic, climatic, hydrological and demo-economic variables (64 columns) that covers the 20th and 21st centuries for all municipalities in Spain. To achieve this, we merged seventeen databases/sources, using several methods, software and programming languages (QGIS, R, Python) to homogenize and downscale the variables for Spanish municipalities. Technical validation results are included with aggregation and alternative sources checks. This database is a valuable resource for researchers from different fields of research, as there is no other resource with this temporal breadth and spatial disaggregation in the current literature. The dataset can contribute to comprehensive analyses with temporal and spatial comparisons around several current debates in the literature and policy (local effects of global and climate change, adaptation, demography, land use change, etc.).

Background & Summary

One of the main problems in the social sciences is obtaining comparable data from different sources in order to conduct comparative analysis, such as multimethodology[1]. This problem is even greater when we want to compare over the long term. Current socio-economic and environmental issues, such as depopulation in some areas, the effects of climate change as, for example, desertification, or differences in the development between societies require databases that are comparable in time and space. These problems are deeply rooted in economic systems, the environment, and societies, so it is imperative to analyse the causes taking into account long-term perspectives but also multidisciplinary variables. Moreover, it is even more difficult to obtain comparable data from different sources when they belong to different fields of knowledge. Researchers need these data to carry out this multidisciplinary analysis.

All this is even more difficult when we try to analyse municipalities. This territorial level of statistics, which is more and more common nowadays, is traditionally not offered by the Statistics Offices, or at least through survey methods. For example, in the European Union, the most common level of territorial statistics in the main variables is for NUTS 2 (Autonomous Communities in the case of Spain) or NUTS 3 (Provinces in the case of Spain). Not even the European Union has used a more disaggregated level beyond the NUTS 3. In this same case, Eurostat offers a survey with the main variables of cities with data from 2013, omitting information on smaller ones (less than 50,000 inhabitants)[2].

The construction of municipal datasets has a growing presence in the literature covering several themes. One of the most important efforts to provide geospatial data in multiple topics is the work of UNECA[3]. Other examples of the construction of municipal data related to socio-economic debates are the cases of waste collection in

[1]Department of Economic Analysis, Faculty of Economics and Business Studies, University of Zaragoza, Agrifood Institute of Aragon (IA2), 50005, Zaragoza, Spain. [2]Department of Economics, Faculty of Economics, Business and Tourism, Universidad de Alcalá, Plaza de la Victoria 2, 28802, Alcalá de Henares, Spain. [3]ARAID (Aragonese Agency for Research and Development), Agrifood Institute of Aragon (IA2), Department of Economic Analysis, Faculty of Economics and Business Studies, University of Zaragoza, 50005, Zaragoza, Spain. [4]Basque Centre for Climate Change, Parque Científico de UPV/EHU, 48940, Leioa, Spain. ✉e-mail: icazcarr@unizar.es

Portugal, the implications of belonging to Special Economic Zones in China or the drinking water database in US municipalities[4–6]. Despite their importance, all of these examples do not cover a long period of time. Another example is the Norwegian dataset for analysing the local government, which offers data for 50 years (1972–2022) with demographic, political or economic measures[7].

Taking all these into account, the case of Spain is striking. Spanish statistics make it possible, not without difficulty, to join several databases and obtain comparable variables in a long-term perspective at a highly disaggregated level. Spain is included in some databases analysing the global climate and geographical conditions. Furthermore, Spanish statistics make it possible, or at least infer, some variables for the analysis of the socio-economic characteristics with a perspective of several decades.

Our dataset can include multidisciplinary variables for the socio-economic analyses, using data from twenty different database sources at a harmonised level of disaggregation (INE[8], MITECO[9–12], GDW[13], SEPREM[14], HYDE[15],HID[16], ESYRCE[17], GIA[18], CRU TS version 4.05[19] IPE-CISC[20,21], IGN[22], Goerlich, 2019[23], Albertus 2023[24], Beltrán Tapia *et al.*[25], Esteban-Oliver and Martí-Henneberg, 2023[26,27]) as well as other extensive data sources (i.e. even along historical census data, yearbooks, etc. on paper or/and digitalized as pdfs) and literature. The Supplementary Information file provides a list of data sources, with the spatial aggregation and temporal coverage indicated (Table S1), and a summary of the main limitations and uncertainties of the original sources from which the key variables of the study were calculated, together with the variables affected by them (Table S2). Obviously for some of the variables, such as climate data, there are other large international databases that provide broad and frequent relevant data (e.g. Menne *et al.*[28]; Silvinsky *et al.*[29]). One of the most important added value features of the database is that there is no other dataset that offers such breadth in time and space with different socio-economic and geographical variables. Some projects in the Spanish economic history have measured, in a long view, some variables that are included in our dataset, such as the SPAREL[30], essays of inequality among Spanish regions[31,32] or the publications of Goerlich[33,34] they focus on the population of villages, towns or municipalities, or on regional inequality. Other analyses with more socio-economic variables are the works of Carreras and Tafunell[35] but they do not provide a disaggregation at the municipal level. There are also other academic works that try to provide or analyse data for Spain. A good example is the work by Goerlich[36], which is relevant for the treatment of climatic data in relation to the Spanish provinces. Another interesting article and data by this author, which we use extensively, is the Municipal Database for the 2011 Census[23]. Cazcarro *et al.*[37] analysed the situation of the main hydraulic infrastructures and water resources providing a context for rural development related to agriculture, and particularly to highly water dependent irrigated agriculture. For the present decades, obviously one finds a myriad of national and international databases on these topics or blocks of variables, at the municipal level or at a resolution that would allow an adequate approximation to this level. Recent examples include the Integrated Municipal Data System (SIDAMUN) of the Spanish Ministry for Ecological Transition and the Demographic Challenge[38], or the new demographic database (DEMOSPA0521) with almost 900 million demographic records from Spain for the last two decades (Lledó and Pavía)[39]. But the limitations of historical data remain for many up-to-date databases, tabular or GIS data, particularly for demographic and other socio-economic variables.

Here, we present a database for Spanish municipalities (departing from a 8,205 objects/polygons shapefile on municipalities of the Spanish National Geographical Institute[40], they were homogenised with the 8,116 Goerlich's[23] homogeneous municipal population series to reach our 8122 homogeneous municipalities of the 2016–2017 entities) for decennial data for 45 different variables (apart from the 7 identification columns, 64 independent columns with variables, hence some of which are presented together in the same box below, listing the names in the dataset for several columns, when they refer to the same variable concept). The time span includes data from 1900 to the present with the following cross sections: 1900, 1910, 1920, 1930, 1940, 1950, 1960, 1970, 1981, 1991, 2001, 2011 and 2021. The following sections describe the variables and how we have harmonised them in our database.

Table 1 shows the climatic variables of our dataset. The total precipitation variable is the average of each decade of total annual rainfall in millimetres for the decadal frequency. For the annual frequency, it is the monthly total precipitation in the reference year. The mean temperature is the average of each decade of the annual average temperature in degrees Celsius for the decadal variable, and the monthly average temperature in the reference year in case of the annual frequency. SPEI "is based on a monthly climatic water balance (precipitation minus PET), and it is expressed as a standardised Gaussian variate with a mean of zero and a standard deviation of one"[41]. In this way, the SPEI data is the average of each decade of the annual mean of the SPEI drought index in millimetres. In case of the yearly frequency, the value corresponds to the monthly average in the reference year. The following variable is the precipitation in the vegetation growth period, namely, the average of each decade (or year depending the frequency) of accumulated rainfall in the growth vegetation period (from April to October). The frost days variable includes the average of each decade of the number of days with a minimum temperature below 0°C. The yearly frequency represents the annual number of days with a minimum temperature below 0°C. The dummies of Köppen's climate classification take the value 1 if the municipality has a specific climate according to the Köppen's climate classification, and 0 otherwise. We take into account if the municipality belongs to the dry and hot climate, oceanic climate, Mediterranean climate, Mediterranean climate with cool summers, boreal climate, humid and hot climate, continental climate with cool summers and continental boreal climate.

Table 2 presents the main information on the geographical variables. The distance variables measure the distance in kilometres from the centroid of the municipality to the nearest coast, to the centroid of Madrid and to the centroid of the provincial capital in a straight line. The distance to the nearest municipality with more than 10,000 and 5,000 inhabitants measure the distance in kilometres of the centroid of both municipalities. The latitude and longitude coordinates correspond to the centroid of the municipality in decimal degrees (DD). The

Variables	Name in the dataset	Data source	Timespan	Frequency	Units
Total precipitation	pp	CRU TS. Version 4.05	1900–2020 (1950–2021)	Decadal (& Yearly)	Decadal mean of total annual precipitation (mm)
Mean temperature	t_average	CRU TS. Version 4.05	1900–2020 (1950–2021)	Decadal (& Yearly)	Decadal mean of mean annual temperature (C°)
SPEI	spei	IPE – CSIC	1900–2020 (1950–2021)	Decadal (& Yearly)	Decadal mean annual mean SPEI drought index (mm)
Precipitation in vegetation growth period	grow_period_pp	CRU TS. Version 4.05	1900–2020 (1950–2021)	Decadal (& Yearly)	Decadal mean of the accumulated precipitation in the months from April to October, both included (mm)
Frost days	frost_days	CRU TS. Version 4.05	1900–2020 (1950–2021)	Decadal (& Yearly)	Decadal mean of number of days per year with a minimum temperature < 0 °C
Dummies of Köppen's climate classification	dry_hot_climate dry_cold_climate oceanic mediterranean mediterranean_fresh_summer humid_hot continental_fresh_summer continental_boreal	IGN	1900 - 2020	Decadal	0 – 1: Dummy variable that takes the value 1 if the municipality has a specific climate type, and 0 otherwise

Table 1. List, sources, timespan and units of climatic variables.

Variables	Name in the dataset	Data source	Timespan	Units
Distance to coast	to_coast_km	IGN	Static variable	Kilometres
Distance to Madrid	to_mad_km	IGN	Static variable	Kilometres
Distance to province capital city	to_prov_cap_km	IGN	Static variable	Kilometres
Coordinates (X,Y)	longitude latitude	IGN	Static variable	Latitude and longitude coordinates of the centroid of the municipality (DD)
Altitude	altitude	IGN	Static variable	Average altitude in metres above the sea
Ruggedness	ruggedness	IGN	Static variable	Standard deviation of the altitude of municipality
Area	area	IGN	Static variable	Area of the municipality (km^2)

Table 2. List, sources, timespan and units of geographical variables. Note: Values for several statistics were also consistently checked with data from Francisco Beltrán-Tapia (e.g. used in Beltrán Tapia *et al*., 2021)[25]. The authors want to greatly acknowledge his help in providing his data, revealing a very high consistency of both databases.

altitude variable is the average altitude of the municipality in metres above sea level. The ruggedness is the standard deviation of the altitude of the municipality. The area is the surface of the municipality in square kilometres.

All the land use variables included in Table 3 are measured in hectares. The irrigated surface of neighbouring municipalities includes all the irrigated surface of neighbouring municipalities, i.e., the municipalities sharing a common boundary with the reference municipality.

Table 4 presents the main information on the hydrological variables. The river basin indicates the river basin to which each municipality belongs. The reservoir water volume capacity of the municipality is measured in cubic hectometres. The usable volume capacity of dammed water in cubic hectometres. The nearest main river provides the name of the nearest river basin bigger than 500 square kilometres. The distance to this main river is in kilometres. The nearest watercourse shows the name of the nearest watercourse, and the distance of this watercourse is in kilometres, regardless of its size.

Finally, Table 5 lists the socioeconomic-demographic variables. The population of the municipalities is the number of inhabitants. The variables of the population residing in municipalities with more than x inhabitants (in thousands) located within a radius between i and j km are measured in number of inhabitants. These variables include all the population who lived in municipalities with more than x inhabitants and they are within specific distance radio, defined by i and j kms. The denomination of the variables follows the pattern: $Px_{i,j}$, being p the population (in thousands), x the threshold of the population (in thousands) residing in municipalities with more than this population than the threshold x located in the radius between the i and j kilometres. For example, p50_25_50km shows the population of municipalities with more than 50,000 inhabitants which are situated in a radius of the referenced municipality between 25 and 50 kilometres (inspired also in Beltrán Tapia *et al.*[25]). The following variables are the distance in kilometres to the nearest municipality with more than 5,000 and 10,000 inhabitants. The Simpson[42] regional classification variables indicate the Spanish agrarian region to which each municipality belongs, depending on whether it is divided into 5 or 11 areas. This classification can be useful for the treatment of different variables. The Population_class initiates a classification, in line with INE, of type of the municipality only with respect to the population size (urban, intermediate, rural). It is based on the classification of the Spanish Statistics Institute in the 1950 Population Census[43]. This classification indicates that a municipality is rural, if it has less than 2,000 inhabitants. An urban municipality has more than 10,000 inhabitants. The

Variables	Name in the dataset	Data source	Timespan	Frequency	Units
Dryland surface	dryland	HYDE 3.2, HID, GIA and ESYRCE (Ministry of Agriculture)	1900–2020 (1950–2021)	Decadal (& Yearly)	Hectares
Pastures surface	pastures	HYDE 3.2, HID, GIA and ESYRCE (Ministry of Agriculture)	1900–2020 (1950–2021)	Decadal (& Yearly)	Hectares
Irrigated surface	irrigated	HID, GIA, ESYRCE (Ministry of Agriculture) and HYDE 3.2	1900–2020 (1950–2021)	Decadal (& Yearly)	Hectares
Cultivated area	cultivated_area	HYDE 3.2, HID, GIA and ESYRCE (Ministry of Agriculture)	1900–2020 (1950–2021)	Decadal (& Yearly)	Hectares
Irrigated surface of bordering municipalities	irrig_neighb	HID, GIA, ESYRCE (Ministry of Agriculture) and HYDE 3.2	1900–2020 (1950–2021)	Decadal (& Yearly)	Hectares
Cultivated area of bordering municipalities	cultivated_area_neighb	HID, GIA, ESYRCE (Ministry of Agriculture) and HYDE 3.2	1900–2020 (1950–2021)	Decadal (& Yearly)	Hectares

Table 3. List, sources, timespan and units of land use variables.

Variables	Name in the dataset	Data source	Timespan	Frequency	Units
River basin	Basin	MITECO (2022)	Static variable		Categorical variable
Reservoir area in the municipality	Reservoir_area	MITECO (2011, 2025)	1900–2020 (1950–2021)	Decadal (& Yearly)	Square kilometers
Reservoir water volume capacity	Reservoir_volume	Global Dam Watch (GDW), MITECO (2011, 2025) and SEPREM	1900–2020 (1950–2021)	Decadal (& Yearly)	Hm^3
Usable volume capacity of dammed water	Usable_reservoir_volume	Global Dam Watch (GDW), MITECO (2011, 2025) and SEPREM	1900–2020 (1950–2021)	Decadal (& Yearly)	Hm^3
Main use of the Reservoir	Reservoir_main_use	Global Dam Watch (GDW)	1900–2020 (1950–2021)	Decadal (& Yearly)	Categorical variable
Reservoir water volume capacity by use	Vol_Irrigation Vol_Electricity Vol_Supply Vol_Other	Global Dam Watch (GDW), MITECO (2011, 2025) and SEPREM	1900–2020 (1950–2021)	Decadal (& Yearly)	Hm^3
The nearest main river (basin surface >500 km^2)	nearest_main_river	MITECO (2018)	Static variable		Categorical variable
Distance to the nearest main river	Dist_main_rivers	MITECO (2018)	Static variable		Km
The nearest watercourse	nearest_all_river	MITECO (2018)	Static variable		Categorical variable
Distance to the nearest watercourse	Dist_all_rivers	MITECO (2018)	Static variable		Km

Table 4. List, sources, timespan and units of hydrological variables.

intermediate municipalities have between 2,000 and 10,000 inhabitants. Our dataset shows the denomination of this classification in each decade. Mainly following Albertus[24] (also Monclús and Oyón, 1988[44], Villanueva and Leal[45]) a dummy variable is elaborated on whether a municipality has colonization towns in the municipal area, adding also a column on the decade in which it had it/them and the town population associated to it. A dummy variable is a binary variable (which takes values 0 or 1) used in statistical models to represent categorical data with two levels, such as the presence or absence of a characteristic. It allows qualitative information to be included in quantitative analyses, helping assess the impact of categorical factors on outcomes.

Based on Esteban-Oliver and Martí-Henneberg[26,27] we compute (for each decade and year) distances of the municipality (centroid) to High-Speed Railway (HSR), to Iberian gauge and to narrow gauge stations. HSR lines are sections of recently built track for high-speed rail services, since 1992, corresponding to the European standard gauge (1435 mm). Iberian gauge network is the largest in Spain (formed by lines which opened after 1848 with the Iberian gauge width, 1668 mm) and the narrow-gauge network includes various types (with widths of 1435 mm, 1062 mm, 1000 mm, 915 mm and 750 mm). Similarly, distances to airports in each decade/year are computed using as reference the year in which the passengers' airports initiated their activity (AENA, 2025)[46]. For all those variables, it should be noted that if returned an "inf" value, it means that there was not such a station or airport in the decade within Spain (hence the algorithm returns "an infinite" distance).

Our reference dataset could be interesting for the environmental, historical, social and economics literature, due to the transversality between different types of variables and the long-time span period it covers (more than a century), as well as data with a high level of disaggregation, considering the municipalities. The diversity of the dataset allows to develop research with several variables included in it or combined with other variables obtained by other researchers. Besides, this dataset allows to evaluate and contribute to several debates in the current literature. One of the potential uses of the data is on the causes, effects and consequences of climate change in a Mediterranean country, allowing the analysis of the long-term trends of these variables and the evaluation of causal effects. For example, Spain has one of the highest levels of hydric stress due to climate change in the EU, mainly due to the potential desertification of a high percentage of its territory[47–49]. Indeed, one of the key future works with the database is to further project and understand, based on past trends, how variables such as temperature or drought indicators affected by climate and environmental change, are expected

Variables	Name in the dataset	Data source	Timespan	Frequency	Units
Number of inhabitants	population	Goerlich (2019) and INE	1900–2010	Decadal	Inhabitants
Population residing in municipalities with more than x population (in thousands) that are located within a radius between i and j km away (i < j); Px_i_j_km	p10_25_50km p50_25_50km p100_25_50km p500_25_50km p10_0_25km p50_0_25km p100_0_25km p500_0_25km	Beltrán Tapia *et al.*[25] and INE	1900–2010	Decadal	Inhabitants
Distance to the nearest municipality of more than 10,000 inhabitants	distance_pop_10000	INE	1900–2010	Decadal	Kilometres
Distance to the nearest municipality of more than 5,000 inhabitants	distance_pop_5000	INE	1900–2010	Decadal	Kilometres
Classification of Simpson's region	Simpson_Areas_5 Simpson_Areas_11	Simpson (1995)	Static variable		Categorical variable
Type of the municipality with respect to the population size (urban, intermediate, rural)	Population_class	INE	1900–2020	Decadal	Categorical variable
Dummy of whether a municipality has colonization towns in the municipal area	municip_coloniz_in_municip	Monclús and Oyón (1988)[44], Villanueva and Leal (1991)[45] and Albertus (2023)	1900–2020 (1950–2021)	Decadal (& Yearly)	Boolean (0–1)
Decade of creation of the colonization town in the municipal area	decade_coloniz_in_municip	Monclús and Oyón (1988)[44], Villanueva and Leal (1991)[45] and Albertus (2023)	1900–2020 (1950–2021)	Decadal (& Yearly)	Decade
Population of colonization towns in the municipality (reference)	pop_coloniz_in_municip	Albertus (2023), Goerlich (2019) and INE	1900–2020 (1950–2021)	Decadal (& Yearly)	Inhabitants
Distance to the closest Iberian Gauge railway stations	railwaystation_distance	Esteban-Oliver and Martí-Henneberg 2023	1900–2021 (1950–2023)	Decadal (& Yearly)	Kilometres
Name of the closest Iberian Gauge railway stations	nearest_station_name	Esteban-Oliver and Martí-Henneberg, 2023	1900–2021 (1950–2023)	Decadal (& Yearly)	Categorical variable
Distance to the closest High-Speed Railway (HSR) stations	highspeedstation_distance	Esteban-Oliver and Martí-Henneberg, 2023	1900–2021 (1950–2023)	Decadal (& Yearly)	Kilometres
Name of the closest High-Speed Railway (HSR) stations	nearest_highspeed_name	Esteban-Oliver and Martí-Henneberg, 2023	1900–2021 (1950–2023)	Decadal (& Yearly)	Categorical variable
Distance to the closest Narrow Gauge railway stations	narrowrail_line_distance	Esteban-Oliver and Martí-Henneberg, 2023	1900–2021 (1950–2023)	Decadal (& Yearly)	Kilometres
Name of the closest Narrow Gauge railway stations	narrowrail_line_name	Esteban-Oliver and Martí-Henneberg, 2023	1900–2021 (1950–2023)	Decadal (& Yearly)	Categorical variable
Distance to the closest (of 48) national airport	airport_distance	Based on AENA[46]	1900–2021 (1950–2023)	Decadal (& Yearly)	Kilometres

Table 5. List, sources, timespan and units of socioeconomic-demographic variables.

to affect socio-economic variables and performance. Other types of potential studies can be developed related to the environmental impacts of economic activities, and vice versa[50–53]. Alternative future applications are the analysis of the causal relationship between some geographical or hydrological characteristics of the territory and economic activities[54] and/or demographic/settlement decisions. In this sense, the database has already been used in Cazcarro *et al.*[54] to study the extent to which irrigated agriculture has contributed to population change in Spanish municipalities. Another example of potential use in this line is the analysis of the determinants of population location, agglomeration problems and market potential[31,55,56].

Figure 1 summarises and classifies the process of obtaining each of the variables corresponding to the different thematic blocks of the database.

Methods

The techniques used for the elaboration of the database presented in this article are based on a variety of statistical approaches, GIS and data analysis techniques oriented especially to the homogenisation and spatio-temporal re-scaling of data from primary sources. These techniques vary widely, depending largely on the nature of the original data and the format in which they are originally found. In this section, the methodological process followed to integrate each of the variables into the database is detailed by thematic group of data. Despite the diversity of applied techniques, to ensure spatial-temporal coherence in the analysis, we have used a static database of municipalities with homogeneous administrative boundaries from 2016 (available at Geographical National Institute INE web portal) throughout the study period. This approach is essential to avoid distortions arising from changes in municipal boundaries, which could affect the interpretation of key variables.

Climatic variables. As can be seen in Table 1, the climate data are from CRU TS. Version 4.05. The only exception is the SPEI drought index, which comes from the portal generated by the Pyrenean Institute of Ecology, part of the Spanish National Research Council (IPE-CSIC). The SPEI used is the 1-month time resolution SPEI. From these data, the entire historical series of SPEI for each of the Spanish municipalities has been constructed and aggregated by decades to homogenise it with the rest of the database.

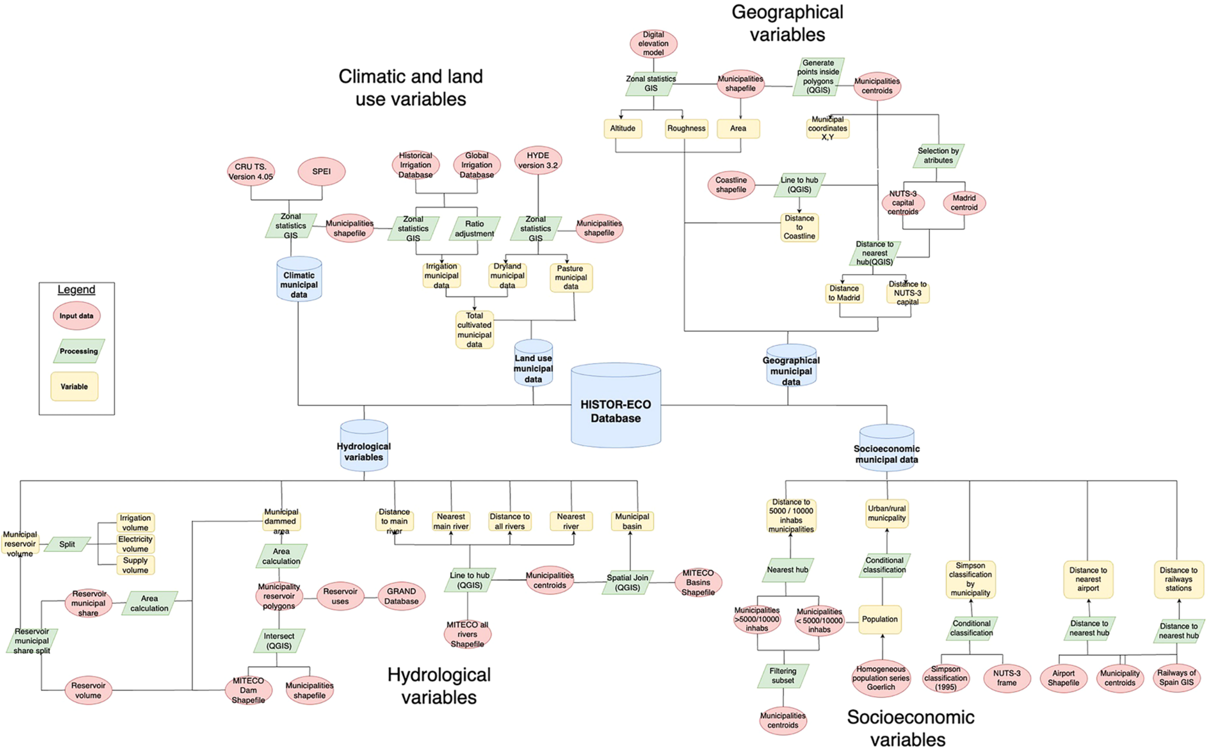

Fig. 1 Structure of the development of variables through each of the 5 blocks of variable types.

In any case, the methods used for the data from both sources are identical, as is the distribution of the data. Both sources distribute the data in a raster format with a cell size of 0.5° × 0.5° with a monthly time resolution. The GIS operation known as zonal statistics was used to re-scale the data to the municipal level, starting from the raster format with the climatic variable in question and a shapefile of the municipalities of Spain. This operation applies, for each municipality, an average of the value of the cells of the raster file contained in the municipal polygon weighted by the proportion of the surface area of the cell included within the polygon. Thus, through Eq. 1, we can approximate the value of each of the climate variables at the municipal level.

$$C_m = \frac{\sum_{i,j \in R_c} \left(\frac{A_{i,j}^m}{A_{i,j}} \cdot r_{c,i,j} \right)}{\sum_{i,j \in R_c} \frac{A_{i,j}^m}{A_{i,j}}} \tag{1}$$

In Eq. 1, $i, j \in R_l$ means all cell i, j of the climatic variable C raster matrix r; $A_{i,j}$ refers to the total area of cell i, j; $A_{i,j}^m$ is the area of the cell contained in municipality m and $r_{c,i,j}$ is the value of the cell (climatic variable c). The application of zonal statistics has been implemented through the R package *exactextract*[57]. In Fig. 2, we can see a schematic of how the zonal statistics operation works as reflected in equation number 1. This figure is also representative of equation number 2, in which we will simply compute the weighted standard deviation instead of the weighted mean.

Once the downscaling to the municipal level has been carried out in this way, we obtain a monthly value. Since we need to aggregate the data at a decadal level in order to homogenise the data at a temporal level, we use the most appropriate statistic (average, sum, etc.) to aggregate them at an annual level, after which the decadal average is carried out.

Although this was the common methodological process, some of the variables have required complementary and/or alternative methodological procedures. The variable *grow_period_pp*, is one of them (see Fig. 3). This variable refers to the total precipitation received during the months of plant growth, i.e. the months from April to October (both included). The temporal resolution of the original data (CRU TS. Version 4.05) is monthly. Therefore, after calculating the total monthly precipitation received in the municipality, the growing months for each of the years were filtered out and the annual total added. To obtain the decadal value, a simple average was taken between the years of the decade. The other variables that required special treatment due to the nature of the original data were those related to the Köppen climate classification. Given that the original data (National Geographic Institute of Spain) are in a polygonal shapefile format, where each polygon represents a specific type of climate, other types of processes were needed to adapt them to a municipal framework. For this purpose, the GIS operation known as *intersect* was used from the shapefiles of the Köppen climate classification and the municipal one. This has created a polygon for each type of climate in relation to each municipality. In this way, we obtain the climate types found in each of the Spanish municipalities and generate a dummy variable for each of them.

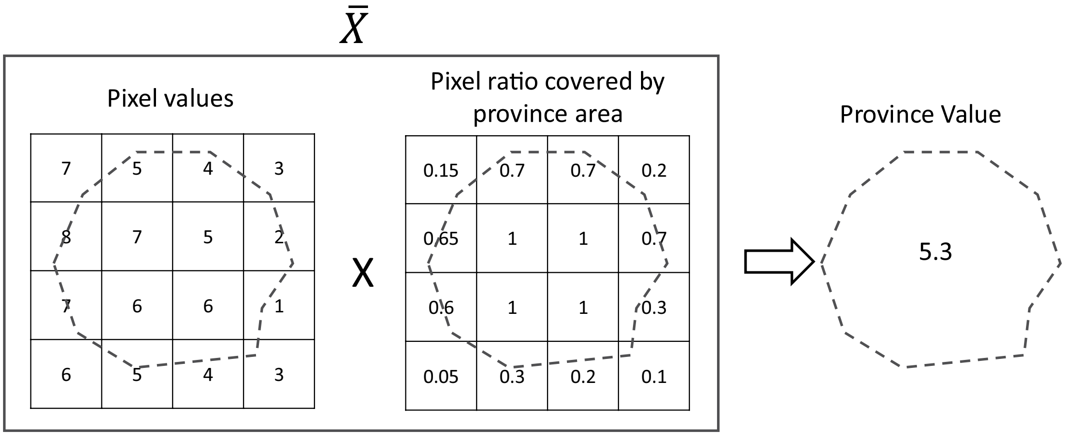

Fig. 2 Graphic scheme of zonal statistics processing operation.

Geographical variables. The geographical variables are basically aspects of distance, location and relief. As we can see in Table 2, we have the distance to Madrid (capital of Spain), the distance to the capital of the province and the distance to the coast. We also have the latitude (y) and longitude (x) coordinates. The first step in calculating these variables is to extract the centroids for each municipality from the shapefile of the Spanish municipalities of the National Geographic Institute[58], using the QGIS algorithm *Generate points (pixel centroids) inside polygons*[59]. From the centroids of the municipalities, the X and Y coordinates were estimated from the *ad hoc* algorithm of the field calculator in QGIS. Once these had been calculated, the distance variables were also calculated from the municipal centroids. The distance variable to Madrid and to the provincial capital was performed by generating 2 independent layers for each of them through a *selection by attributes*, using the R package *dplyr*[60] on the shapefile attribute table loaded in an *sf* object[61]. An *sf* object (short for simple features object) in R is a data structure designed for handling spatial data efficiently while adhering to international standards. It combines geometric information (such as points, lines, and polygons) with attribute data in a format similar to a data frame. This makes it easy to integrate spatial analysis with common R workflows. Additionally, sf objects support coordinate reference systems (CRS), enabling precise spatial analysis and visualization.

After that, the distance of all the municipalities centroids to each of the centroids of the layers was calculated and filtered to match, in the case of the provincial capital, the entity of the same province as the municipality in question. The QGIS algorithm *Distance to nearest hub (points)* implemented in the R package *qgisprocess*[62] was used for this purpose. For the coastline, the procedure was similar, except that the distance between the centroids of the municipalities and the coastline was calculated using the *Distance to nearest hub (line to hub)* algorithm, also implemented in the R package *qgisprocess*. The municipal area was calculated using the *$area* algorithm in the QGIS field calculator applied to the municipalities' shapefile. Finally, altitude and ruggedness were again calculated using the zonal statistics operation implemented in the *exactextract* R package from the shapefile of municipalities (zones) and a digital elevation model. In the case of altitude, the weighted mean is used in the same way as in Eq. 1. Ruggedness, as we know, can be estimated as the standard deviation of the altitude of each municipality, so the zonal standard deviation is applied according to Eq. 2 (see also Fig. 4).

$$\sigma_m = \sqrt{\frac{\sum_{i,j \in R_c} \left(\frac{A^m_{i,j}}{A_{i,j}}\right) \cdot \left(r_{c,i,j} - \underline{H}_m\right)}{\sum_{i,j \in R_c} \frac{A^m_{i,j}}{A_{i,j}}}} \tag{2}$$

The nomenclature is the same as in Eq. 1, adding $\underline{H}_m$, which is the average altitude in each municipality.

Land use variables. The land use variables are related to agricultural uses. Among them, we can distinguish between the area of rainfed crops, the area of irrigated crops and the area of pasture. The area of irrigated crops (equipped for irrigation, to be specific) comes from the Historical Irrigation dataset (HID henceforth) (Siebert *et al.*)[16] and from the Global Irrigation Area (GIA) database (Meier *et al.*)[18] combined with Crop area and yield survey (ESYRCE hereafter[17]) to extend the series to the decade of the 2010s and 2020 s (see Fig. 5). This variable includes the irrigated land under plastic greenhouses. Since the format of the main climatic variables is identical (data in raster format with a spatial resolution of 0.5°, the value of each cell refers to the number of hectares contained in it), the procedure is similar. For the other land uses, the format is the same, but the source of the data is the History Database of the Global Environment (HYDE version 3.2)[15]. In this case, in order to know the total municipal hectares of each land use, a sum weighted by the area of the cell located in the municipal polygon of the hectare value of each cell is made. The calculation of the municipal hectares of land use can be expressed according to Eq. 3. The nomenclature of Eq. 3 is very similar to Eq. 1, only the sub-index *c* is replaced by the sub-indice *l*, which refers to each land use. This method has been used for the variables Dryland (dryland), Pastures (pasture) and Irrigated (irrigated_ha) area.

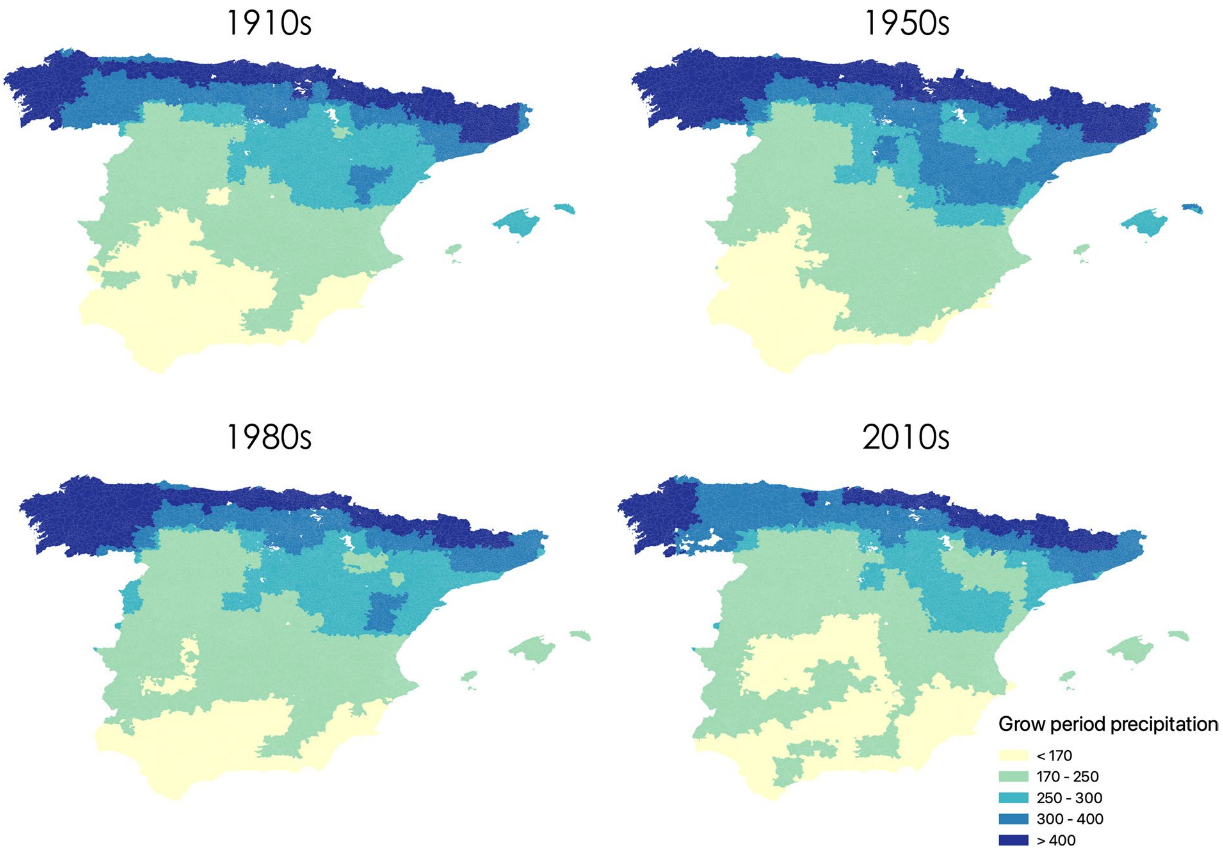

Fig. 3 Evolution of growing season precipitation.

$$L_m = \sum_{i,j\in R_l} \frac{A^{m}_{i,j}}{A_{i,j}} \cdot r_{l,i,j} \tag{3}$$

The variable Cultivated area is derived from the sum of Dryland Surface and Irrigated Surface. The variables Irrigated surface of adjacent municipalities and Cultivated surface of adjacent municipalities, on the other hand, have undergone a different transformation. In both cases, the first step has been to generate an adjacency matrix, i.e., for each of the municipalities, the municipalities that are in contact with it take a value of 1, and 0 otherwise. From this matrix, the value of the Irrigated or Cultivated surface of those municipalities that have a value of 1, i.e. the municipalities that are adjacent to the municipality in question, is added. This process is represented by Eq. 4. In it, C is a contiguity matrix of size $n\ x\ n$, where n is the number of municipalities. Therefore, $C_{m,j}$ will be an element of the matrix that is 1 if municipality m is adjacent to municipality j, and 0 otherwise. On the other hand, V_j will be a vector of size n containing the land use values in hectares for each municipality.

$$S_m = \sum_{j=1}^{n} C_{mj} \cdot V_j \tag{4}$$

Hydrological variables. Hydrological variables are essentially linked to the access of different municipalities to water resources, watercourses or variables related to water infrastructure such as reservoirs (see Fig. 6). These variables are detailed in Table 4. To calculate the basin variable, the GIS *Spatial join* operation was used, implemented in QGIS using the municipality shapefile and the hydrographic basin shapefile (Ministry for the Ecological Transition and the Demographic Challenge)[10]. In this operation, each of the municipalities takes the value of the hydrographic basin in which it is located. If a municipality is located between two basins, it takes the value of the basin that covers the largest area of the municipality. Then, we have the variables related to the reservoirs, i.e., the reservoir area, the reservoir volume and the usable reservoir volume. The difference between total reservoir volume and usable reservoir volume lies in their specific definitions and functions. Total reservoir volume refers to the entire amount of water stored in the reservoir at a given time, encompassing all the water from the bottom to the surface, including portions that may not be accessible for practical use. In contrast, usable reservoir volume represents the portion of water that can be actively utilized for the reservoir's intended purposes, such as irrigation, water supply, hydroelectric power generation or flood control. This excludes the so-called "dead volume", which remains below the outlet level or is retained to prevent sediment resuspension and ensure the reservoir's structural stability. As a result, usable reservoir volume is always less than or equal to the total reservoir volume, with the difference determined by the reservoir's design and operational parameters.

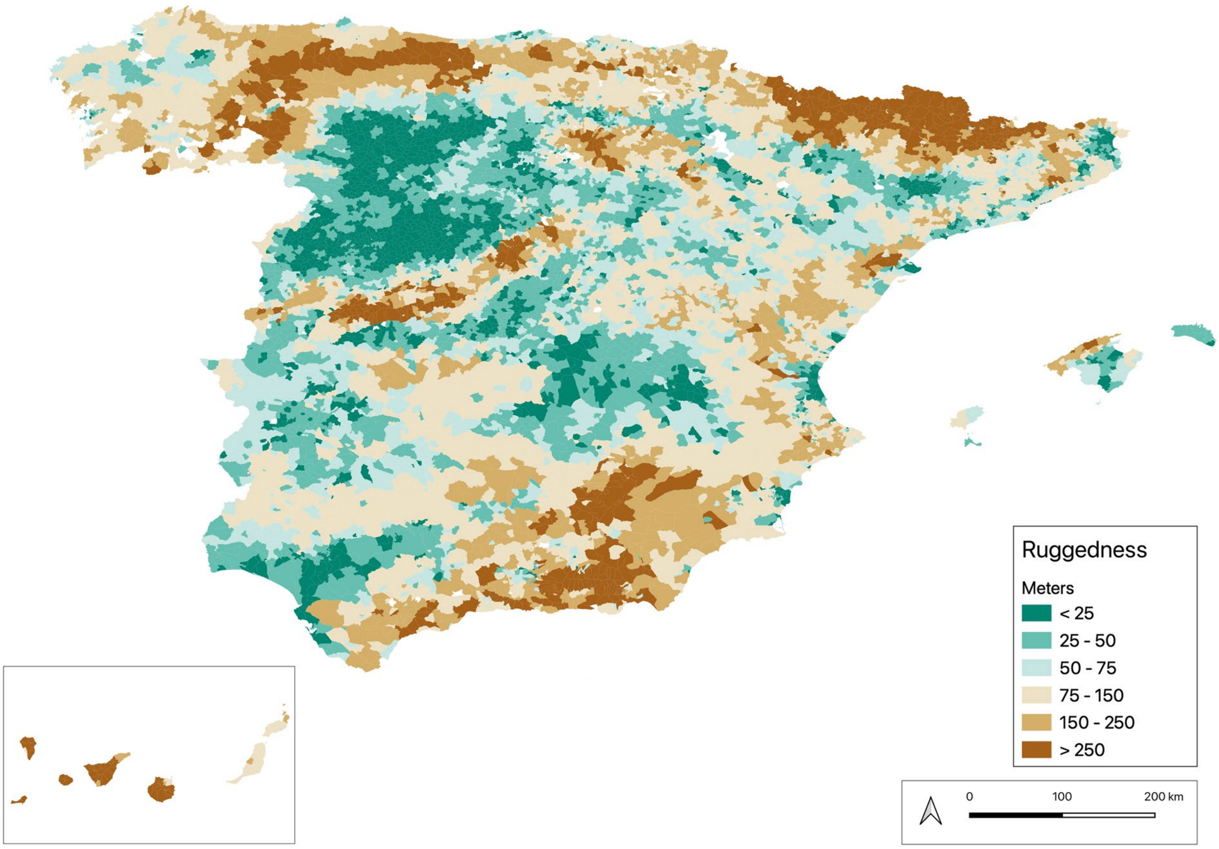

Fig. 4 Ruggedness by municipality.

To calculate these variables, we have used the reservoir shapefile from the inventory of dams and reservoirs of the Ministry for the Ecological Transition and the Demographic Challenge (MITECO)[9] and the inventory of dams of the Spanish Society of Dams and Reservoirs (SEPREM)[14]. The GIS *intersect* operation, implemented in QGIS, was used to calculate the reservoir area of each municipality. As previously mentioned, this operation has obtained a polygon for each part of the reservoir located in each municipality. Once we have the polygons of the reservoirs divided by municipalities, we calculate the area, using the *st_area* function of the R *sf* package[61]. We must take into account that municipalities do not always have the same reservoir capacity over time, since dams and reservoirs are built at different times by different hydrological projects. For this reason, the SEPREM inventory of reservoirs was used to determine the year in which the construction of each reservoir was completed. In order to obtain the decadal value of the reservoir area by municipality, we filter for each decade by the year of construction, group by municipality and add up the total reservoir area (it is possible that a municipality has more than one reservoir). These filtering and summing processes by municipality were executed in R from the *dplyr* package. Considering that the area of each reservoir in a given municipality is expressed in Eq. 5 where P_{mr} is the polygon of intersection between the municipality m and the reservoir r:

$$A_{m,r} = Area(P_{m,r}) \tag{5}$$

The total reservoir area in the municipality m for the decade y is equal to:

$$A_{m,y} = \sum\nolimits_{r \in y} A_{m,r} \tag{6}$$

Once we have calculated the reservoir area, we obtain the values of the volume and useful volume dammed per municipality. MITECO's (2011[9], 2025[12]) inventory of dams and reservoirs contains the data on the volume and useful volume of each reservoir. Therefore, all that remains is to distribute the volume of the reservoir among the different municipalities between which the reservoir is located. This distribution is done in proportion to the previously calculated area of the reservoir that is located in each municipality. Thus, the calculation of the reservoir volume and the municipal useful reservoir volume per reservoir r and municipality m is characterised according to Eq. 7:

$$V_{m,r} = \frac{A_{m,r}}{A_r} \cdot V_r \tag{7}$$

Again, we have to filter by year and make a sum grouped by municipality. Therefore, the calculation of the reservoir volume and the total useful reservoir volume by municipality is as follows:

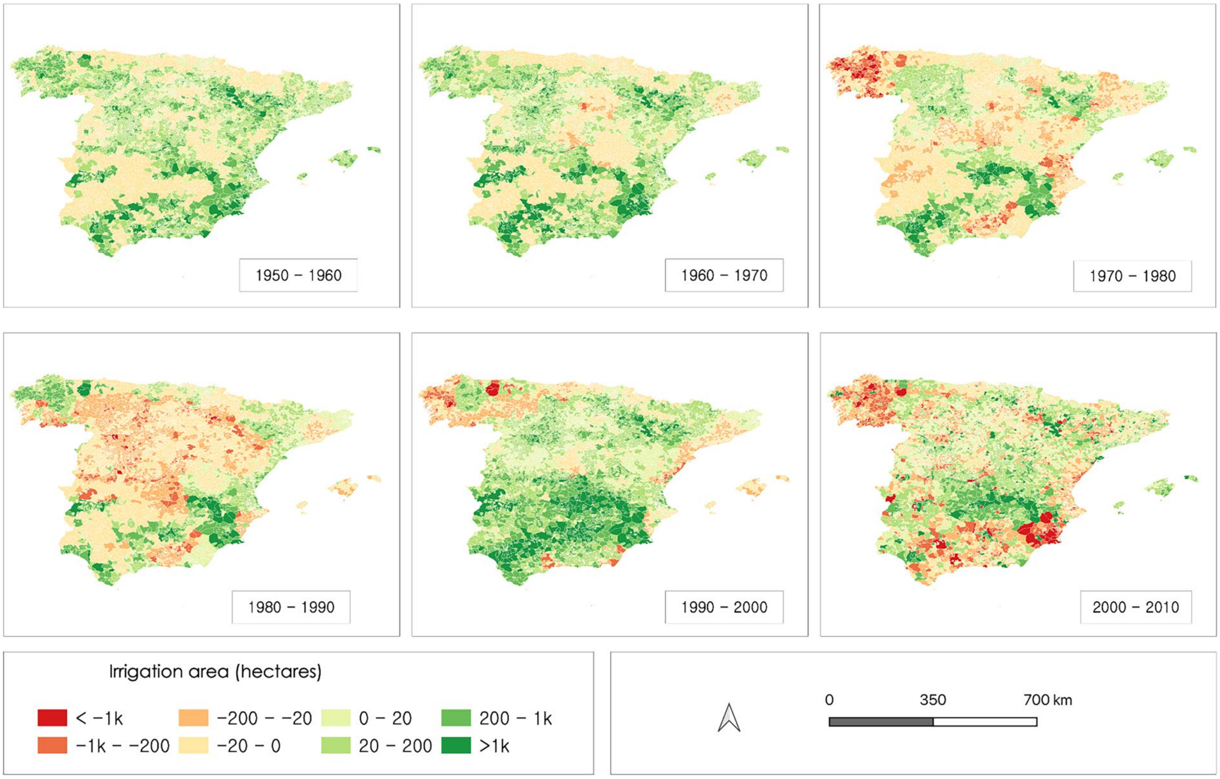

Fig. 5 Irrigated area change (hectares) by decade.

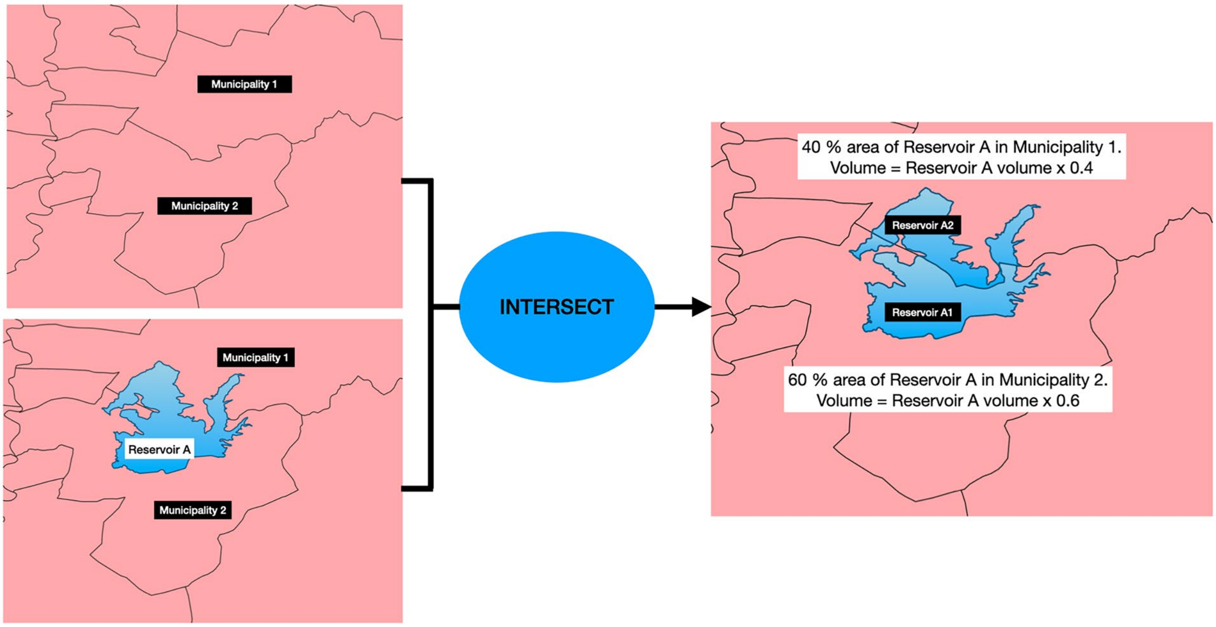
Municipality 1
Municipality 2
Municipality 1
Reservoir A
Municipality 2
INTERSECT
40 % area of Reservoir A in Municipality 1.
Volume = Reservoir A volume x 0.4
Reservoir A2
Reservoir A1
60 % area of Reservoir A in Municipality 2.
Volume = Reservoir A volume x 0.6

Fig. 6 Reservoir volume estimation by municipality.

$$V_{m,y} = \sum\nolimits_{r \in y} V_{m,r} \tag{8}$$

Through this process, we now have the evolution of the reservoir capacity of each of the Spanish municipalities from 1900 to the present day.

The Global Dam Watch (GDW) database version 1.0 has been used to determine the main use of the dammed water in the different municipalities, which provides the vast majority of dams and reservoirs worldwide with different information on the main use. From it, the different uses have been assigned to each of the reservoirs in the inventory of dams and reservoirs of MITECO. Although it has not been possible to assign the main use to all the reservoirs, given that they are not in GDW, the main reservoirs are represented, with a total of 98.8% of the useful reservoir volume and 82.5% of the reservoir area. Once the reservoir uses have been assigned, the procedure follows the same approach as the calculation of impounded volume and area. First, the reservoirs existing in the decade under study are filtered based on their year of construction. Then, a spatial intersection is performed between the filtered reservoirs and the municipalities. Each municipality is assigned the primary use of the reservoir with which it intersects. If a municipality intersects with multiple reservoirs, it is assigned the reservoir that contains the largest volume of water within its boundaries. Figure 6 shows the process of estimating the volume of reservoirs at the municipal level.

Additionally, a distribution of the water attributed for each use by municipalities has been performed following the allocation rules by[63] in its baseline scenario, represented in the variables "Vol_Irrigation", "Vol_electricity", " Vol_supply" and "Vol_Others". These rules state that if a reservoir in the GDW database has only one main use (i.e., no secondary uses), 100% of its volume is allocated to that use. If the reservoir has one secondary use, 75% of its volume is assigned to the main use and 25% to the secondary use. In the case of two secondary uses, 50% of the volume is allocated to the main use, while each secondary use receives 25%. Once the reservoir's volume has been distributed among its different uses, the process for assigning the volume used in each municipality follows the same approach as for the total and usable reservoir volumes. This ensures that the volume data recorded in the inventory of dams and reservoirs is fully accounted for. The type of resulting data is visualized in Fig. 7.

The remaining hydrological variables (nearest_all_river, Dist_all_rivers, nearest_main_river and Dist_main_rivers) refer to the distance of the municipalities to the different watercourses. The main_river concept refers to the main rivers, watercourses with a certain entity of their own, while the *all_rivers* concept includes all the watercourses of the different water networks in Spain. In this way, we start from three main data sources. Firstly, we have the centroids of the municipalities, which we had previously calculated. In addition to these centroids, we have two shapefiles of the water network. One, with the complete water network and the other with the main watercourses from the Ministry for Economic Transition and Demographic Challenge (MITECO, 2018[11]). The variable (and shapefile provided by the Ministry for Economic Transition and Demographic Challenge) Dist_main_rivers refers to the rivers listed as main rivers in Article 3 of the Water Framework Directive, which according to the European guidance 'Guidance No 22 - Updated WISE GIS guidance', are the main rivers whose catchment area is greater than or equal to 500 km. On the other hand, the variable Dist_all_rivers refers to all natural watercourses.

First, we converted the two shapefiles of the hydrographic networks from polylines to points, so that each watercourse became a set of points using the *Points along geometry* QGIS tool. Then, in both cases, we used the QGIS algorithm *Distance to nearest hub (points)* implemented in the R package *qgisprocess*[62]. In this way, for each of the centroids of the municipalities, we have the distance to the nearest point relative to the water network, and the watercourse to which it belongs, to any watercourse in the case of nearest_all_river and Dist_all_rivers or to the main river in the case of nearest_main_river and Dist_main_rivers.

Socioeconomic variables.

Socio-economic variables include mainly demographic variables, access to transport infrastructure and agglomeration economies. First, we have the total population. This was obtained from Goerlich's (2019)[23] homogeneous municipal population series. Spanish censuses exhibit certain problems that Goerlich *et al.*[23,33] manage to resolve, although they themselves consider Spanish census data to be "reasonably reliable" (p. 29). To observe the problems throughout the 20th-century censuses, and how these authors solved them, one should consult Goerlich *et al.*[33]. Regarding the population variable, despite taking the population series of Goerlich[23] as a reference, we have had to make adjustments due to changes in the municipal boundaries with respect to the (most recent found) spatial base that was used (of 2016). In this way, we ensure the homogeneity of the entire series.

In those cases where municipal mergers or segregations have occurred, we have applied harmonization procedures to correctly assign population values at each point in time. Firstly, for mergers prior to the GIS layer used, we have added the values of the original municipalities. For example, in the case of Cerdedo-Cotobade (36902) and Oza-Cesuras (15902), which arose from the merger of the municipalities of Cotobade (36012) and Cerdedo (36011), and of Oza dos Ríos (15063) and Cesuras (15026), respectively, their values have been added from the 1900s to the 2000s to maintain consistency in the series. On the other hand, when a municipality has been segregated in recent years and already appears as an independent entity in the GIS layer used, we have redistributed its previous population based on its relative share in the most recent census. For example, Valderrubio (18914) was separated from Pinos Puente (18158) in 2013, and for previous years it has been assigned a proportion of the total population of the original municipality according to its participation in the 2021 census. A similar process has been applied to Dehesas Viejas (18065) and Domingo Pérez de Granada (18915), separated from Iznalloz (18105) respectively in 2014 and 2015, to Játar (18106), segregated from Arenas del Rey (18020), and to the municipalities of Tiétar (10904) and Pueblonuevo de Miramontes (10905), which became independent from Talayuela (10180) in 2013 and 2015, respectively. Likewise, for municipalities created after the reference date of the GIS layer, they have remained attached to their original municipality throughout the series to avoid the creation of unpopulated areas. This is the case of segregations in 2018 such as Palmar de Troya (41904) from Utrera (41095), Torrenueva Costa (18916) from Motril (18140) and San Martín del Tesorillo (11903) from Jimena de la Frontera (11021); similarly in 2019 of Fornes (18077), separated from Arenas del Rey. Finally, situations have been corrected in which a municipality had population values in 2011 but missing data in previous censuses due to its recent creation. In these cases their previous values have been assigned

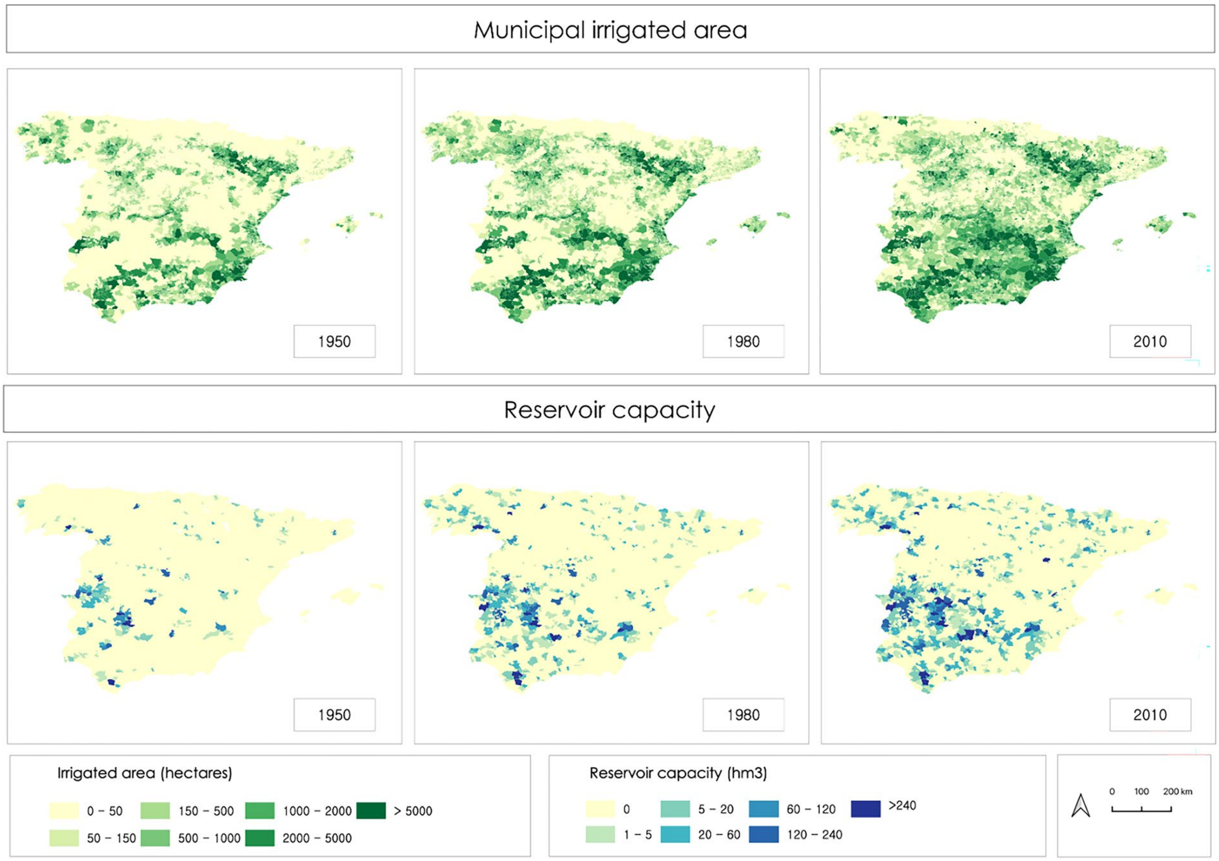

Fig. 7 Evolution of reservoir capacity (bottom row) and irrigated area (top row).

according to the available historical data, namely those independent in 2009: Alagón del Río (10903), from Galisteo, Villanueva de la Concepción (29902), from Antequera in 2009, and Vegaviana (10902), from Moraleja; and the same for Villamayor de Gállego (50903), from Zaragoza in 2006 and Benicull de Xúquer (46904), from Polinyà de Xúcar in 2003.

This procedure guarantees homogeneity in the analysis, preventing administrative changes from introducing biases and allowing the observed dynamics to reflect real processes rather than artifacts derived from modifications in municipal boundaries. All in all, the CSV files comprise our 8122 homogeneous municipalities of the 2016 entities. The shapefile comprises 8,205 objects/polygons given that it also includes 87 objects of communal forests (mainly in Navarre, with the "facerías" regime, Basque Country, Castile and Leon, Cantabria and one case in Castile-La Mancha and another in Asturias).

From the population values, the municipalities were classified as rural, intermediate or urban, resulting in the variable Population_class (see Fig. 8). Let's remind that in this variable, municipalities with less than 2,000 inhabitants are considered rural, those with between 2,000 and 10,000 inhabitants are considered intermediate and those municipalities with more than 10,000 inhabitants are considered rural, following the Spanish Statistics Institute's criterion[64]. The *case_when()* utility of the R package *dplyr* was used to perform this classification.

The variables distance_pop_10000 and distance_pop_5000 collect the distance from each municipality to municipalities with more than 10,000 and 5,000 inhabitants, respectively. As the methodology is the same in both cases, only the number of inhabitants is changed, we refer to the number of inhabitants (5,000 and 10,000) as *p*. To do this, firstly, for each of the decades collected in the databases, we divide the sample into 2 using the centroids of the municipalities. That is, we generate two different shapefiles: one with the municipalities below *p* inhabitants and another with the municipalities above *p* inhabitants. These are again selected by attributes using a filter from the *dplyr* package on the shapefiles loaded in R as *st objects* of the *sf* package. Once generated, for each decade, the two municipal subsets using the value *p* as a threshold, the distance from the municipalities with a population below *p* to the nearest municipality with a population above *p* was calculated. This was done using the QGIS tool *Nearest Hub* implemented in the *qgisprocess* R package. Population residing in municipalities with more than *x* inhabitants (in thousands) that are located within a radius of between *i* and *j* km could be expressed as $p\_x_{i,j}$, being $i < j$ (in the files, it takes the name of "*p_x_i_j*km").

The variables Simpson_Areas_5 and Simpson_Areas_11, correspond to the classification of Spanish areas established by Simpson[42], which divides Spain, grouping provinces, into regions corresponding to diverse variables such as rainfall, crop mix, land distribution, or cultural and linguistic background. Taking into account the Simpson region in which each province is located and the province in which each municipality is located, a direct allocation of the municipal Simpson area (5 and 11, depending on the level of aggregation) is made

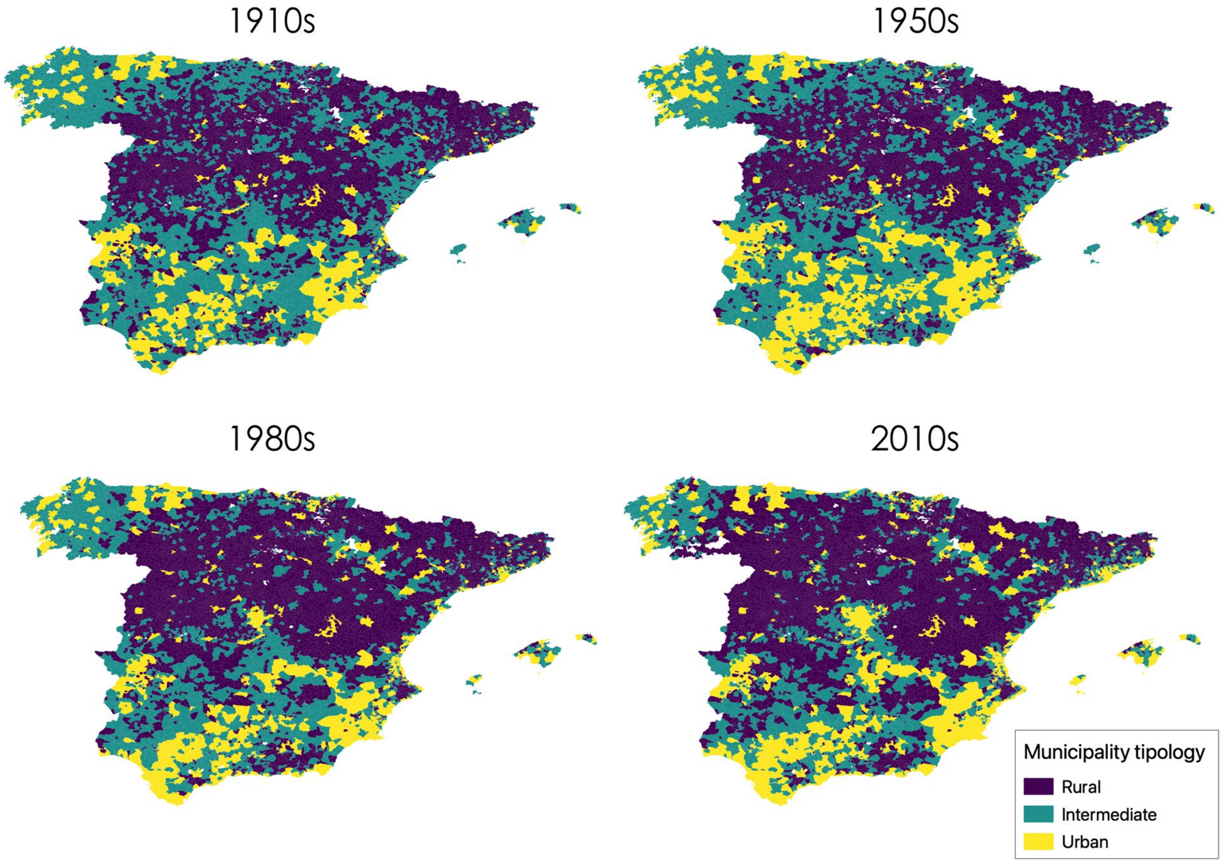

Fig. 8 Evolution of municipalities classified as rural, intermediate and urban.

through a conditional sentence. These aggregations can help different scholars to obtain a criterion to aggregate Spanish provinces in their analysis. The categories in Simpson_Areas_11 are: GA: A Coruña, Lugo, Ourense, Pontevedra; BISCAY: Asturias, Cantabria, Gipuzkoa, Biscay; CL: Avila, Burgos, León, Palencia, Salamanca, Segovia, Soria, Valladolid, Zamora; Cent: Albacete, Ciudad Real, Cuenca, Guadalajara, Toledo; UEBRO: Alava, La Rioja, Navarra; LEBRO: Huesca, Lleida, Teruel, Zaragoza; Med: Alicante, Barcelona, Castellon, Girona, Murcia, Tarragona, Valencia; EX: Badajoz, Caceres; WAN: Cadiz, Huelva, Malaga, Sevilla; EAN: Almeria, Cordoba, Granada, Jaen; and CAN: Santa Cruz de Tenerife y Las Palmas de Gran Canaria. The categories in Simpson_Areas_5 are formed by: North: GA + BISCAY; Interior: EX + UEBRO + LEBRO + CL + Cent; Med: Med; Andalucia: EAN + WAN; and Can: Can. All these categories are explained in the pages 21, 22 and the map 1 from[42].

Data Records

The dataset[65] is available at figshare.

Through the methodological process detailed in the previous section, the HISTORECO database was obtained. It is made up of 45 variables of different types that allow for an integrated analysis of the socio-economic evolution of Spanish municipalities. The database is distributed into two different structures: panel and spatial. The panel format consists of a tabular file with a.csv extension. In this format, we have one observation per municipality and year, and a single column per variable. Thus, there are 7 initial identification columns (NUTS 2 code, NUTS 2 region, NUTS 3 code, NUTS 3 region, municipality code, municipality, year of decade) followed by the set of variables. The spatial format consists of a shapefile that allows the information to be used directly for spatial analysis in a GIS environment. In this shapefile, the spatial base consists of a polygon for each of the Spanish municipalities. These polygons have an associated attribute table (.dbf file) in which we have only one observation per municipality (since we have only one polygon per municipality) and one column per variable-decade combination. Thus, we have only 5 columns related to identifiers (we exclude the decade column) followed by 819 columns, which are the combination of all variables for each decade. The structure of the information in both formats is shown in Fig. 9.

The structure of the data provided in the repository consists of two main parts, the methods folder and the database folder. The methods folder contains a folder called *scripts*, which contains the scripts used to create the database, and two documents, *code_help.pdf* and *code_help.rmd*, which contain a series of hints on how to use the code correctly and explain each of its parts. The database folder consists of 5 files and a folder. Firstly, “Variable_description.xlsx” summarizes the available variables. Then there is the database in long format, called *Panel* in.csv format in the Records section, with one observation per municipality and decade (*Historeco.csv*).

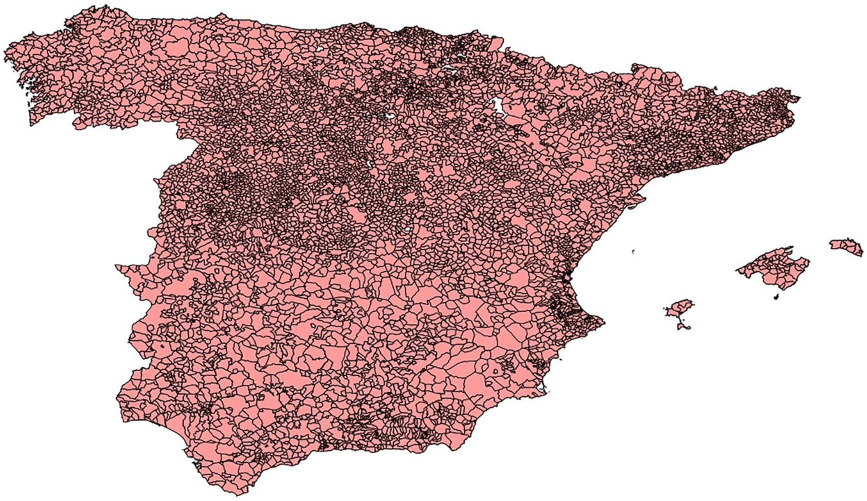

Fig. 9 Schematic form the two structures in which the generated database is contained. Note: This figure shows schematically the two structures in which the generated database is contained. The *Panel data* format is in.csv format. On the other hand, the *Spatial* structure is associated with a shapefile containing the municipal polygons. The data is available in figshare[65] together with this submission.

Also, in.csv format is the database in what is called in the dataset section spatial format, where we have only one row per municipality (*Historeco_wide.csv*). In addition, there is another spatial file (*geopackage*) of the municipal polygons in which, in the table of attributes, the database can be found in spatial format to be used directly in a geographical information system (*Historeco.gpkg*) to carry out other analyses, representation through cartographic products, etc. In the "MunicipalitiesShapefile" folder one finds also only polygons and municipal identifiers (*Municipalities.shp*) in case one wants to join the database in spatial format autonomously. Finally, in the Database folder we can also find a file with the annual version of the database *(Historeco_Year.csv*). The annual database limits its time series to the start in the 1950s for several reasons. Firstly, this version is subject to the availability of yearly data. While the decadal database is semi-definitive, subject to minor corrections and adjustments, the annual database is a 'demo' version, which is still under construction and may be subject to considerable changes. Many of the sources, after certain time thresholds, decrease their temporal resolution (as in the case of land use variables) or we cannot find reliable sources to rely on (as in the case of socio-economic variables), and we deliberatively avoid performing simple interpolations to complete years introducing "fake" trends. The other factor that led us to decide this is computational capacity. The annual version of the database from 1950s has a weight of about 3GB, making it already computationally expensive to handle. Extending it would obviously exacerbate this problem. For all these reasons, and in order to maintain homogeneity, we have decided, for the time being, to cut off the series to the start in 1950s. Figure 10 shows the file structure of the repository.

Technical Validation

In order to validate the accuracy of our database, various tests have been applied to the different variables. Here we describe and present the main aggregated checks, but also the disaggregated comparative data are also available upon readers' request. In general terms, the technical validation was carried out in several directions: firstly, it was checked that the aggregation of the spatially disaggregated data corresponded to the original aggregated values, at different levels: municipality, *comarcal* (a kind of county level), province, regional and national, with priority given to them in that order, with subsequent rebalancing. In other words, all the Root Mean Square Error, RMSE, of those original aggregated values are equal to 0. Secondly, all possible external data were used for comparison, e.g., national/regional surface statistics with the GIS data. In practice, the technical validation was fully hybridised with the construction of the data, as most of the initially found external data were integrated if they were found to be superior. When there were discrepancies between the databases, we established priorities. This was the case with the HID[16], which was preferred due to the fact that it is based on the National Statistics Institute and agrarian statistics, but also covers the XXth century, with other land use databases. Still, there were 2 key complementary computations. One was the effort of consistency with the total land use data estimated from the Spanish municipalities shapefile[40], (in a few selected cases, 11 out of 8122 municipalities, it seemed that the irrigated surface from HID could be higher than the total surface data of the former), deciding to adjust to these figures the total surface area. Secondly, we extended the computation of municipal irrigable area from the dataset (HID)[16] with the Global Irrigation Area (GIA) database[66] and official statistics ESYRCE[17] to consistently extend the series to the 2010s decade. Additionally, when available, the GIS and Database (.csv,.xlsx, etc.) formats data were double-checked. All in all, we have performed tests, essentially on the potential deviations from the objective values or series. Figure 11 summarizes the scheme of data gathering, treatment-processing, technical validation, feedback to computations and results.

One of the tests was oriented to the estimation of irrigated hectares. We compared the estimated area equipped for irrigation (AEI) using GIS techniques from the Historical Irrigation Dataset (HID) with external

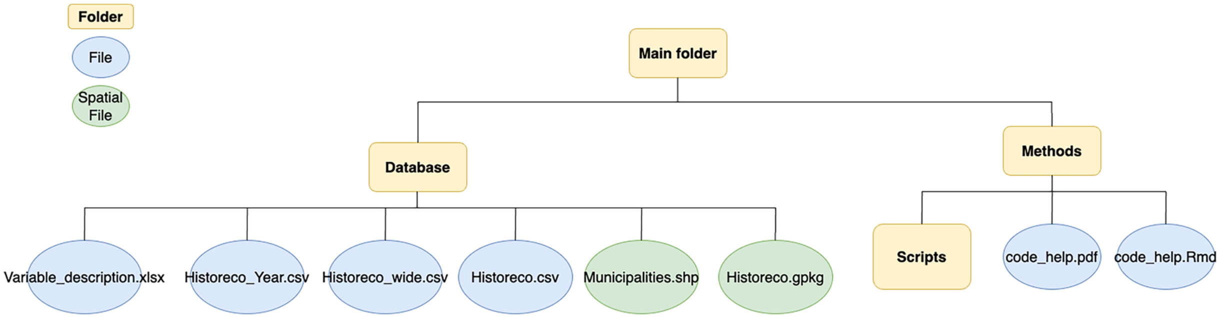

Fig. 10 Repository data structure.

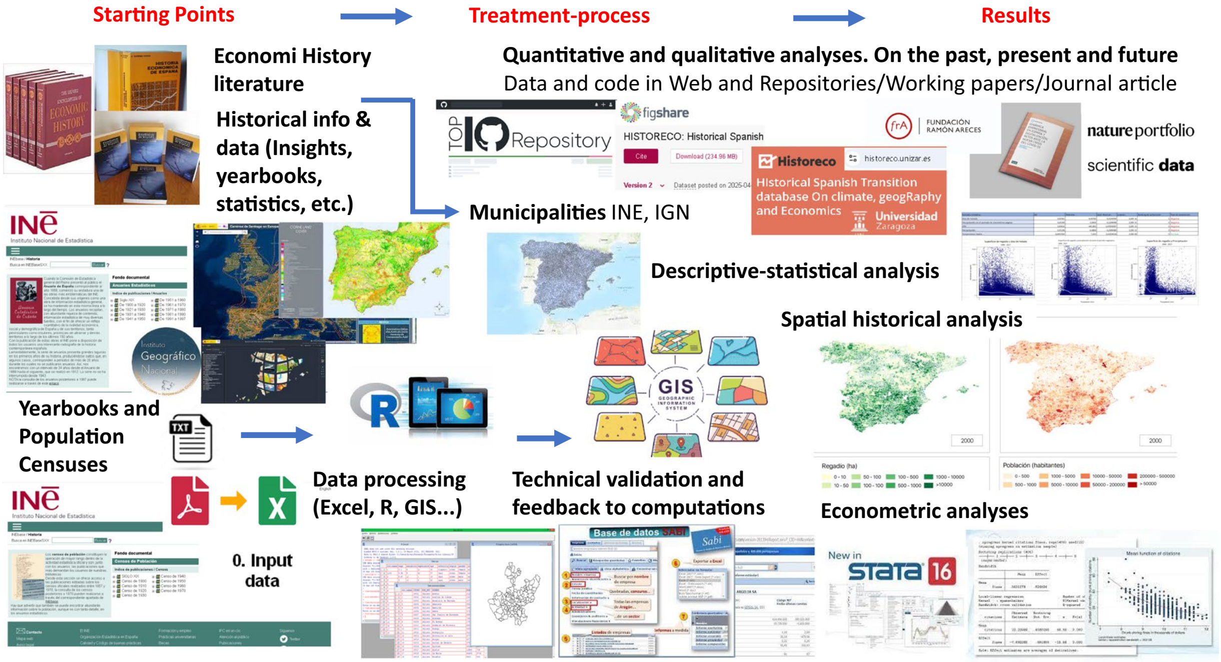
Starting Points
Treatment-process
Results
Economi History literature
Historical info & data (Insights, yearbooks, statistics, etc.)
Quantitative and qualitative analyses. On the past, present and future
Data and code in Web and Repositories/Working papers/Journal article
Municipalities INE, IGN
Descriptive-statistical analysis
Spatial historical analysis
Yearbooks and Population Censuses
Data processing (Excel, R, GIS...)
Technical validation and feedback to computations
Econometric analyses
0. Input data
figshare
HISTORECO: Historical Spanish
Historeco
Historical Spanish Transition database On climate, geogRaphy and Economics
Universidad Zaragoza
nature portfolio
scientific data
INE
Geográfico Nacional
GIS
Base de datos SABI
Stata 16

Fig. 11 Full picture of points of departure for data gathering, treatment-processing, technical validation, feedback to computations and results.

statistics to assess the accuracy of these estimates. The FAO's AQUASTAT (FAO 2025)[67] provides statistics on irrigated area using some official data, specifically from the 1999 National Institute of Statistics agricultural census, but also complementary information to obtain the AEI concept and be more precise by regions (e.g. regional councils, hydrographic confederations and Corine land cover 2000). Using irrigation data for 2000 estimated from the HID (the closest year), we found that the estimated values are close to the FAO'S AQUASTAT values at the regional level. Although the census and complementary data compiled is attributed to 1999 and the estimates of HID to 2000, we find of interest to compare the values. To analyse the similarity of the HID variable incorporated in our database with such data, metrics such as the Root Mean Square Error (RMSE) and the Weighted Absolute Percentage Error (WAPE) have been used, whose estimation is represented in Eq. 9. Although the metrics we use contain the word error, they are not measuring errors as such. We have to remember that FAO Aquastat statistics, despite using census information as a reference, are also estimates, so they cannot be assumed to be true and, therefore, we should talk about discrepancies or differences.

$$WAPE = \left(\frac{\sum_i |y_i - \hat{y}_i|}{\sum_i |y_i|} \right) \tag{9}$$

The RMSE for the regional comparison is 20,353.68, with the largest discrepancy observed in the Region of Murcia, which is smaller in HID by 62,299 ha. The WAPE is 5.27%. These differences might be partly due to surface area changes from 1999 to 2000, a period characterised by the expansion of irrigation. At the provincial level, the relative similarity of the results is confirmed with an RMSE of 6,502.45. The largest difference in higher values of HID compared to FAO-AQUASTAT occurred in the province of Seville (16,084 ha) and the lowest

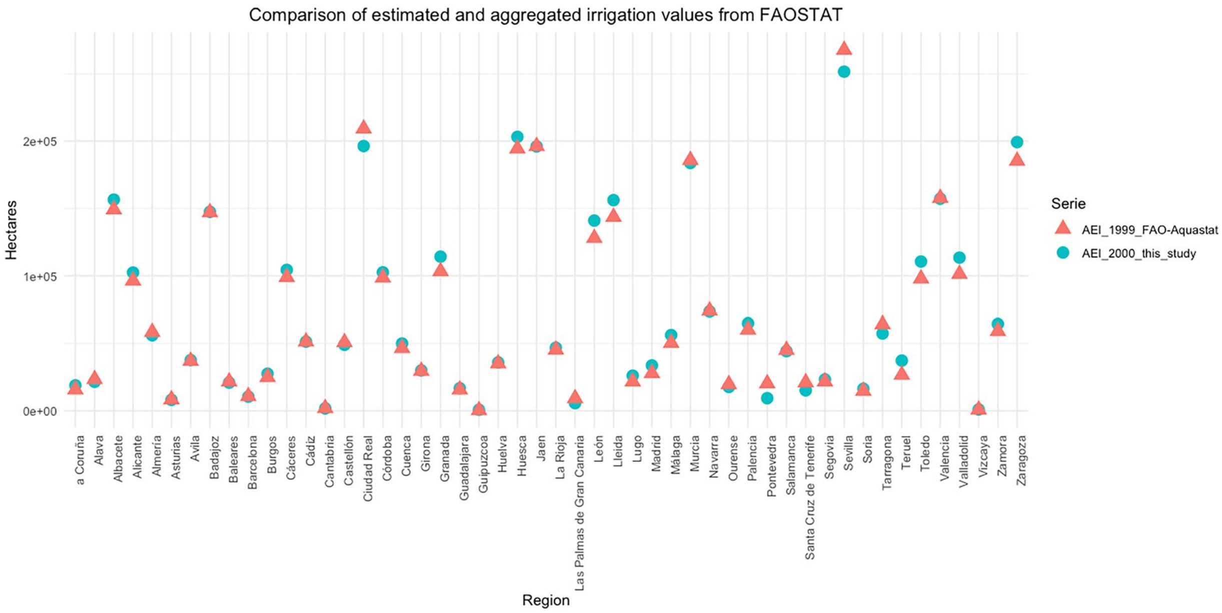

Fig. 12 Accuracy assessment of estimated irrigated area. Note: In this figure we can see the comparison between the FAO-AQUASTAT values for 1999 at NUTS 3 level (red triangle) and the estimated aggregated municipal irrigation in our study at the same level (blue circle) in order to analyse the similarity of our estimates.

values compared to FAO-AQUASTAT in the province of Zaragoza (13,962 ha). The highest percentage discrepancy relative to the irrigated area itself were found in provinces with smaller irrigated areas like Guipuzcoa, which is one of the most humid Spanish provinces. Absolute discrepancies logically occurred in provinces with larger irrigated areas. The WAPE at provincial level is 6.47%, showing a high degree of concordance between the estimated HID and FAO-AQUSTAT values.

Overall, the correlation between the estimated and FAO-AQUASTAT irrigation values is high, with coefficients of 1 at both the regional and provincial levels, indicating the robustness of the GIS-based estimates used in this study. In this respect, Fig. 12 shows the comparison by province of the estimated irrigation values from HID in our database (for the year 2000) and those recorded by FAO-AQUASTAT estimates.

It should be noted that the HID irrigation series only up to the year 2005. Therefore, in order to achieve a high degree of accuracy, different methods were used. Firstly, the Information System on Land Use in Spain (henceforth, SIOSE, *Sistema de Información de Ocupación del Suelo en España*)[68] was initially used as a data source. This database provides comprehensive geospatial data on land use and land cover throughout Spain. This dataset offers detailed and regularly updated information, which makes it an interesting resource for territorial planning and land use change studies. However, when approaching through it and using GIS operations such as Intersect, area calculations, etc., the differences in the original data (vector format vs. raster format, methodological differences in its estimation, etc.) led us to consider other approaches. In this context, the Global Irrigation Dataset (GID) mentioned above had an identical format, and its methodology was very similar. In fact, it is based on the HID series itself, extended with different remote sensing methods. In this way, it was also possible to apply exactly the same methodology (based on the *zonal statistics* GIS operation) to estimate the municipal irrigated hectares for the whole series. Comparing the results obtained with the two data sources, it was possible to appreciate a more homogeneous series with much less quantitative jumps with the decades estimated from the HID source (1900 – 2000). Thus, the approximation using the GID source was chosen to obtain homogeneity and to reduce the error.

Besides, to achieve the highest possible degree of precision in the adjustment, especially in recent decades, official statistics were used. In this case, it is the ESYRCE. For this validation, a relationship is established between the official statistics and the estimated irrigated area. The official ESYRCE[17] statistics provide data on agricultural land, including irrigated areas, at the NUTS 3 level. Our goal is to align the growth rate of our estimated variable at the NUTS 3 level with the ESYRCE statistics. First, we adjust our estimated variable for the year 2000 to match the NUTS 3 totals provided by ESYRCE. Next, we calculate the growth rate between the adjusted irrigation values for 2000 and the ESYRCE data for 2010. Then, we calculate the growth rate between the adjusted irrigation values for 2000 and the ESYRCE data for 2010[17]. We then subtract the 2000 estimate from this new 2010 value, of the absolute change in hectares at the NUTS 3 level, which needs to be distributed across municipalities. To distribute this change, we calculate each municipality's share of the change in irrigation based on the estimated change within its respective Autonomous Community. This proportion is used to allocate the NUTS 3 level change among the municipalities. Finally, we add the municipal-level change (2000–2010) to the irrigation estimate for 2000 derived from the HID series, allowing us to approximate the 2000–2010 change based on official totals while maintaining a consistent structure across all decades. Thus, we find that the growth rates estimated from GIA do not deviate from the official growth rates. To update irrigation to the 2020 s in the

	Estimated			*Carreras & Tafunell (2005)*			Absolute error (%)		
Year	Irrigated	Rainfed	Total crop area	Irrigated	Rainfed	Total crop area	Irrigated	Rainfed	Total crop area
1900	—	—	16060	—	—	17822	—	—	9,9
1930	—	—	19120	—	—	21964	—	—	12,9
1960	2226	18361	20587	1828	18694	20522	21,7	1,8	0,3
1970	2770	17763	20533	2198	18321	20520	26,0	3,0	0,1
1980	2932	17145	20077	2822	17677	20499	3,9	3,0	2,1
1990	2978	16494	19472	3199	16973	20172	6,9	2,8	3,5
2000	3667	14421	18088	3408	14897	18304	7,6	3,2	1,2

Table 6. Crop area accuracy assessment by decade.

database, the same procedure has been used, but taking as a reference the values of 2011 and 2021 by estimating the changes between both years.

As can be seen from the official statistics, the methodology is very close to the totals for the most recent decades. However, these statistics do not provide us with a very long time series, so that the verification of the previous ones is not covered, and furthermore no totals have been verified, neither for the cultivated area nor for the rainfed crops. For this reason, we have used other historical literature sources, such as Carreras & Tafunell[35], particularly on the primary sector. Chapter 4 of this work, on the agricultural and fishing sector[69], contains historical statistics on the cultivated area, with discontinuities, also disaggregated between rainfed and irrigated crops at the national level. Using the previously mentioned metric, *WAPE*, we can estimate the degree of adjustment of our series in comparison with the historical statistics. Before analysing the results of this comparison, we have to take into account that we are comparing slightly different concepts. The estimated irrigation data collect the area equipped for irrigation, regardless of whether or not it was irrigated that year. Carreras & Tafunell (2005), on the other hand, collect the area that was effectively irrigated based on the censuses. These conceptual differences can cause significant discrepancies between the figures of one and the other.

In Eq. 9, y_i is the value estimated in our series and $\hat{y}_i$ is the value of the historical series of[35]. Using this metric, for the whole series we have obtained a difference of 4.3% for the total cultivated area, 2.8% for the rainfed cultivated area and 11.6% in the case of irrigation. Table 6 shows the absolute difference by decade in detail with values expressed in thousands of hectares.

With respect to climate variables, the CRU TS Version 4.05[19] database was used for this study due to its extensive temporal coverage (since 1901) and the wide availability of climate variables at the global level, which allows a detailed historical analysis of the relationship between climate and the demographic and economic development of municipalities. Its spatial resolution of 0.5 degrees facilitates its use in large-scale studies, and its consistency in data collection and processing methodology makes it a reliable source for analysing climate patterns over time. Furthermore, the raster distribution facilitates downscaling through operations such as zonal statistics, as previously explained. However, this same spatial resolution of approximately 50 km per cell can be seen as a weakness when applied at the municipal level, especially in areas with variable topography or microclimates, where the data may not adequately capture local climatic variations. Furthermore, as a global database, it may present limitations in terms of data accuracy in areas with sparse or poorly developed meteorological networks.

Given the problems discussed above, an approximation of the accuracy of our data at different scales and variables was made by comparing them with different databases. We have to be cautious when making comparisons with climate data, as data from different sources come from different measurement techniques/methodologies, each with its own margin of error. At the national level, the report *“Analysis of temperatures in Spain in the period 1961–2018*[70] from the State Meteorological Agency was used. In this report we find an average temperature series for Spain since 1961 on a national scale, estimated based on data from 42 reference stations. Using our database and aggregating the data of the report by decades (simple average), we obtain a *WAPE* of 10%.

For the accuracy analysis of the precipitation variable, a validation at the municipal level was applied. For this purpose, the example of Barcelona was used, given the length of its series (since 1940 with consistent data). Daily data from the *European Climate Assessment dataset* (*ECAD)*[71] for the El Prat Airport station were utilised. These daily data, as in the estimation of our precipitation variable, were summed annually, and the decadal mean was calculated from the annual data. Again, the *WAPE* value reaches 10%. The same process has been carried out, except that the annual aggregation is done by means of the average and not the sum, with the series of average temperature also relative to the Barcelona Airport weather station. In this case, the *WAPE* reaches 8% for the whole series. Taking into account the uncertainties associated with the databases compared, we consider that this level of accuracy guarantees precise and consistent analyses when using climate data.

Regarding the hydrological variables, given the methodology and the sources used (official government statistics), the technical checks carried out were that, after the allocation of the areas and volume of reservoirs, the totals of surface area and volume of the reservoirs recorded in the MITECO reservoir inventory and the national total were met, and in fact, they were met.

By way of summary, the Table 7 gives an overview of the accuracy values obtained in the technical validation process.

It is worth noting that there are many variables (in addition to the non-numerical ones) that are not explicitly validated through WAPE. Variables such as the calculation of distances from the centroid of the municipality to different entities have not been validated using this methodology. As it is a calculation of Euclidean distance

Variable	Reference	Scale of check	WAPE
irrigated	FAO's Aquastat	National	5.27%
irrigated	FAO's Aquastat	Province	6.47%
irrigated	Historical Irrigation Dataset (HID)	National	3,2%
dryland	Carreras & Tafunell (2005)	National	2,8%
cultivated_area	Carreras & Tafunell (2005)	National	4,3%
t_average	State Meteorological Agency	National	10%
t_average	European Climate Assessment dataset (ECAD)	Municipal	8%
pp	European Climate Assessment dataset (ECAD)	Municipal	10%
Usable_reservoir_volume	MITECO & SEPREM Dam inventory	National	0%
Reservoir_volume	MITECO & SEPREM Dam inventory	National	0%

Table 7. Accuracy values obtained in the technical validation process.

between two arbitrarily chosen entities, there is no official source with which to systematically check them. The values of these variables have been checked for consistency with specific examples.

Usage Notes

The data is also freely available on the portal website https://historeco.unizar.es, where additional features and possibilities for filtering, selecting and downloading are available (e.g. queries on the metrics of the different municipalities, comparisons, analysis of trends, etc.).

Code availability

The Code is available on GitHub.

Received: 8 November 2024; Accepted: 23 April 2025;
Published online: 17 May 2025

References

1. Seawright, J. *Multi-Method in Social Science: Combining Qualitative and Quantitative Tools*. (Cambridge University Press, Cambridge, 2016).
2. Eurostat. Regions and cities. https://ec.europa.eu/eurostat/web/cities/database (2024).
3. UNECA. *Geoinformation in Socio-Economic Development: Determination of Fundamental Datasets for Africa*. (2007).
4. Oliveira, V., Sousa, V. & Dias-Ferreira, C. Dataset of socio-economic and waste collection indicators for Portugal at municipal level. *Data Brief* **22**, 658–661 (2019).
5. Wang, J. The economic impact of Special Economic Zones: Evidence from Chinese municipalities. *J Dev Econ* **101**, 133–147 (2013).
6. Hughes, S., Kirchhoff, C. J., Conedera, K. & Friedman, M. The Municipal Drinking Water Database. *PLOS Water* **2**, e0000081 (2023).
7. Fiva, J. H., Halse, A. H. & Natvik, G. J. Local government Dataset. www.jon.fiva.no/data.htm (2023).
8. INE. Demografía y población = Demographics and population. Preprint at https://www.ine.es/dyngs/INEbase/es/categoria.htm?c=Estadistica_P&cid=1254734710984 (2025).
9. MITECO. Embalses = Dams. Last Accessed 9-9-2024. https://wms.mapama.gob.es/sig/agua/Embalses/wms.aspx (2011).
10. MITECO. Cuencas hidrográficas de ríos principales según Pfafstetter modificado = Watersheds of major rivers according to modified Pfafstetter. Last Accessed 8-11-2024. https://wms.mapama.gob.es/sig/Agua/CuencasRiosPfafs/wms.aspx (2022).
11. MITECO. Ríos completos clasificados según Pfafstetter modificado = Complete rivers classified according to modified Pfafstetter. Last Accessed 9-9-2024. https://wms.mapama.gob.es/sig/Agua/RiosCompPfafs/wms.aspx? (2018).
12. MITECO. Home -> Water -> Topics -> Water resources -> assessment -> Hydrological Bulletin -> Download of historical reservoir data since 1988. 02-11-2025. Last Accessed 3-3-2025. https://www.miteco.gob.es/es/agua/temas/evaluacion-de-los-recursos-hidricos/boletin-hidrologico.html (2025).
13. Lehner, B. *et al*. The Global Dam Watch database of river barrier and reservoir information for large-scale applications. *Sci Data* **11**, 1–18 (2024).
14. Sociedad Española de Presas y Embalses (SEPREM). Inventario de Presas Españolas. https://www.seprem.es/presases.php?p=26 (2021).
15. Goldewijk, K. K., Beusen, A., Doelman, J. & Stehfest, E. Anthropogenic land use estimates for the Holocene - HYDE 3.2. *Earth Syst Sci Data* **9**, 927–953 (2017).
16. Siebert, S. *et al*. A global data set of the extent of irrigated land from 1900 to 2005. *Hydrol Earth Syst Sci* **19**, 1521–1545 (2015).
17. Ministerio de Medio Ambiente y Medio Rural y Marino. Encuesta sobre Superficies y Rendimientos Cultivos (ESYRCE). in *Anuario de Estadística* (ed. Secretaría general técnica) vol. 1, 520–910 (Subdirección General de Estadística, Madrid, 2010).
18. Meier, J., Zabel, F. & Mauser, W. A global approach to estimate irrigated areas - A comparison between different data and statistics. *Hydrol Earth Syst Sci* **22**, 1119–1133 (2018).
19. Harris, I., Osborn, T. J., Jones, P. & Lister, D. Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Sci Data* **7**, 109 (2020).
20. Beguería, S., Serrano, S. M. V., Reig-Gracia, F. & Garcés, B. L. SPEIbase v.2.10 [Dataset]: A Comprehensive Tool for Global Drought Analysis. Preprint at https://doi.org/10.20350/digitalCSIC/16497 (2024).
21. Trullenque-Blanco, V., Beguería, S., Vicente-Serrano, S. M., Peña-Angulo, D. & González-Hidalgo, C. Catalogue of drought events in peninsular Spanish along 1916–2020 period. *Sci Data* **11**, 703 (2024).
22. Instituto Geográfico Nacional (IGN). Clasificación climática según Köppen. *Atlas Nacional de España* https://www.ign.es/web/resources/docs/IGNCnig/ANE/Capitulos/04_Climayagua.pdf 103–104 (2010).
23. Goerlich, F. J. A municipal database from the 2011 Spanish census. *Applied Economic Analysis* **27**, 226–238 (2019).
24. Albertus, M. The Political Price of Authoritarian Control: Evidence from Francoist Land Settlements in Spain. *Journal of Politics* **85**, 1258–1274 (2023).

25. Beltrán Tapia, F. J., Díez-Minguela, A. & Martinez-Galarraga, J. The shadow of cities: size, location and the spatial distribution of population. *Annals of Regional Science* **66**, 729–753 (2021).
26. Esteban-Oliver, G. & Martí-Henneberg, J. The expansion of the Spanish railway network (1848–1941): an analysis through the evolution of its companies. *Revista de Historia Industrial* **31**, 87–144 (2022).
27. Esteban-Oliver, G. & Martí-Henneberg, J. Railways of Spain GIS (1848–2023). Preprint at https://doi.org/10.34810/data917.
28. Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E. & Houston, T. G. An overview of the global historical climatology network-daily database. *J Atmos Ocean Technol* **29**, 897–910 (2012).
29. Slivinski, L. C. *et al.* Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. *Quarterly Journal of the Royal Meteorological Society* **145**, 2876–2908 (2019).
30. Tirado-Fabregat, D. *et al.* SPAREL. *sparel.com* (2022).
31. Beltrán Tapia, F. J. & Martinez-Galarraga, J. Inequality and Growth in a Developing Economy: Evidence from Regional Data (Spain, 1860–1930). *Social Science History* **44**, 169–192, https://doi.org/10.1017/ssh.2019.44 (2020).
32. Díez-Minguela, A., Martínez-Galarraga, J. & Tirado-Fabregat, D. *Regional Inequality in Spain, 1860-2015*, (Palgrave Macmillan, Londres, 2018).
33. Goerlich, F. J., Maudos, J. & Mollá, S. *Distribución de La Población y Accesibilidad a Los Servicios En España*. (Instituto Valenciano de Investigaciones Económicas (Ivie) - Fundación Ramón Areces, 2021).
34. Goerlich, F. J. & Mollá, S. Desequilibrios demográficos en España: evolución histórica y situación actual. *Presupuesto y Gasto Público* **102**, 31–54 (2021).
35. Carreras, A. & Tafunell, X. *Estadisticas Historicas de España. Siglo XIX-XX. Fundación BBVA*. https://doi.org/10.1017/CBO9781107415324.004 (2005).
36. Goerlich Gisbert, F. J. Datos climáticos históricos para las regiones españolas. CRU TS 2.1. *Investigaciones de Historia Economica* **8**, 29–40 (2012).
37. Cazcarro, I., Duarte, R., Martín-Retortillo, M., Pinilla, V. & Serrano, A. How Sustainable is the Increase in the Water Footprint of the Spanish Agricultural Sector? A Provincial Analysis of the Years 1955 and 2005. *Sustainability* **7**, 5094–5119 (2015).
38. MITECO. Integrated Municipal Data System (SIDAMUN). Preprint at https://www.miteco.gob.es/en/reto-demografico/temas/analisis-cartografia.html (2024).
39. Lledó, J. & Pavía, J. M. A detailed database of sub-annual Spanish demographic statistics: 2005–2021. *Sci Data* **11**, 79 (2024).
40. IGN. Base de Datos de Divisiones Administrativas de España: Recintos municipales de España = Database of Administrative Divisions of Spain: Municipal precincts of Spain. Last accessed 7-2-2024. https://centrodedescargas.cnig.es/CentroDescargas/catalogo.do; https://centrodedescargas.cnig.es/CentroDescargas/busquedaSerie.do?codSerie=LILIM; https://datos.gob.es/es/catalogo/e00125901-spaignllm (2024).
41. Beguería, S., Vicente-Serrano, S. M. & Angulo-Martínez, M. A Multiscalar Global Drought Dataset: The SPEIbase: A New Gridded Product for the Analysis of Drought Variability and Impacts. *Bull Am Meteorol Soc* **91**, 1351–1356 (2010).
42. Simpson, J. *Spanish Agriculture: The Long Siesta, 1765-1965*. (Cambridge University Press, 1995).
43. INE. *Censo de Población - Population Census*. (1950).
44. Monclús Fraga, F. J. & Oyón, J. L. *Historia y Evolución de La Colonización Agraria En España*. (Instituto de Estudios de Administración Local: Ministerio de Agricultura, Pesca y Alimentación: Ministerio de Obras Públicas y Urbanismo, 1988).
45. Villanueva Paredes, A. & Leal Maldonado, J. Historia y Evolución de La Colonización Agraria En España. La Planificación Del Regadío y Los Pueblos de Colonización. (Ministerio de Agricultura, Pesca y Alimentación (MAPA), Madrid (Spain), 1991).
46. AENA. Aeropuertos y destinos de AENA = AENA airports and destinations. Last Accessed 3-3-2025. https://www.aena.es/es/aerolineas/aeropuertos-y-destinos/nuestros-aeropuertos.html#aeropuertosnacionales (2025).
47. Vargas, E. *Effects of Climate Change in Mediterranean Water Resources and Their Economic Implications*. (2016).
48. Vargas-Amelin, E. & Pindado, P. The challenge of climate change in Spain: Water resources, agriculture and land. *J Hydrol (Amst)* **518**, 243–249 (2014).
49. Estrela, T., Pérez-Martin, M. A. & Vargas, E. Impacts of climate change on water resources in Spain. *Hydrological Sciences Journal* **57**, 1154–1167 (2012).
50. Ortiz-Bobea, A. Chapter 76 - The empirical analysis of climate change impacts and adaptation in agriculture. in *Handbook of Agricultural Economics* (eds. Barrett, C. B. & Just, D. R. B. T.-H. of A. E.) vol. 5, 3981–4073 (Elsevier, 2021).
51. Millner, A. & Dietz, S. Adaptation to climate change and economic growth in developing countries. *Environ Dev Econ* **20**, 380–406 (2015).
52. Tol, R. S. J. The Economic Impacts of Climate Change. *Rev Environ Econ Policy* **12**, 4–25 (2018).
53. Castells-Quintana, D., Lopez-Uribe, M., del, P. & McDermott, T. K. J. Adaptation to climate change: A review through a development economics lens. *World Dev* **104**, 183–196 (2018).
54. Cazcarro, I., Martín-Retortillo, M., Rodríguez-López, G., Serrano, A. & Silvestre Rodríguez, J. Retaining population with water? Irrigation policies and depopulation in Spain. vol. 256 Preprint at https://hdl.handle.net/10419/298604 (2024).
55. Krugman, P. The Role of Geography in Development. *Int Reg Sci Rev* **22**, 142–161 (1999).
56. Head, K. & Mayer, T. Gravity, market potential and economic development. *J Econ Geogr* **11**, 281–294 (2011).
57. Baston, D. Fast Extraction from Raster Datasets using Polygons. Version 0.10.0. https://doi.org/10.32614/CRAN.package.exactextractr, https://isciences.gitlab.io/exactextractr/, https://github.com/isciences/exactextractr (2023).
58. Instituto Geográfico Nacional. Base de Datos de Divisiones Administrativas de España: Recintos municipales de España. Preprint at https://centrodedescargas.cnig.es/CentroDescargas/busquedaSerie.do?codSerie=LILIM (2024).
59. QGIS Development Team. QGIS Geographic Information System. Preprint at (2024).
60. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. dplyr: A Grammar of Data Manipulation. Preprint at https://dplyr.tidyverse.org (2023).
61. Pebesma, E. Simple Features for R: Standardized Support for Spatial Vector Data. *R J* **10**, 439–446 (2018).
62. Dunnington, D., Vanderhaeghe, F., Caha, J. & Muenchow, J. R package qgisprocess: use QGIS processing algorithms. *Version 0.2.0* Preprint at https://doi.org/10.5281/zenodo.10418745 (2023).
63. Bartolomé-rodríguez, M. I., Rubio-varas, M. & Sesma-martín, D. Water for Whom? Unravelling the Allocation of Water Storage Capacity between Irrigation and Electricity Uses in Spain during the 20th Century. *Hist Agrar* **94**, 165–201 (2024).
64. INE. Introducción, Tomo III. Volúmenes provinciales. Censo de 1960 = Introduction, Volume III. Provincial volumes. 1960 Census. Fondo documental. Instituto Nacional de Estadística (INE). https://www.ine.es/inebaseweb/pdfDispacher.do;jsessionid=CF3C7F2070AA86B3ABC52A83BB5BFBEB.inebaseweb02?td=126937&ext=.pdf (2024).
65. Rodríguez-López, G., Serrano González, A., Martín-Retortillo, M. & Cazcarro, I. HISTORECO: Historical Spanish Transition Database on Climate, Geography, and Economics of the 20th-21st Century. *figshare. Dataset v2*. https://doi.org/10.6084/m9.figshare.27262032.v2 (2024).
66. Meier, J., Zabel, F. & Mauser, W. Global Irrigated Areas. Preprint at https://doi.org/10.1594/PANGAEA.884744 (2018).
67. FAO. AQUASTAT - FAO's Global Information System on Water and Agriculture -> Spain. Last accessed 3-3-2025. https://www.fao.org/aquastat/en/geospatial-information/global-maps-irrigated-areas/irrigation-by-country/country/ESP (2025).
68. IGN. Sistema de Información de Ocupación del Suelo de España (SIOSE) [Cartografía Digital]. 1:25.000. Preprint at https://www.siose.es/presentacion (2011).

69. Barciela, C., Giráldez, J. & López, I. Sector agrario y pesca. in *Estadísticas históricas de España: siglos* XIX - *XX* (eds. Carreras, A. & Tafunell, X. (coordinadores)) vol. 3, 257–356 (2006).
70. Ministerio de Agricultura, P. y A. *Anuario de Estadística*. vol. 1 (Madrid, 2000).
71. Squintu, A. A., van der Schrier, G., Brugnara, Y. & Klein Tank, A. Homogenization of daily ECA&D temperature series. *International journal of climatology* **39**, 1243–1261 (2018).

Acknowledgements

Authors acknowledge the financial support provided by Fundación Ramón Areces (CISP20A6661), of the Spanish Ministry of Science and Innovation research projects (PID2021-123220NB-I00, PID2022-138886NB-I00, and PID2022-140010OB-I00) and of the Government of Aragon–European Fund for Economic and Regional Development (research projects S55_23R and S40_23R). Also, authors want to acknowledge the comments received by the Editor and anonymous reviewers, as well as from participants at the XLVII Meeting of Regional Studies (RER-International Conference on Regional Science), IV International Conference/XVIII Congreso de Historia Agraria/XI Encontro Rural RePort, I Seminar for New Academics in Regional Science (SNAR), X Workshop de jóvenes investigadores en Economía y Empresa and Taller de Historia Económica of the Universidad de la República (Uruguay).

Author contributions

Guillermo Rodríguez-López: Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – revised draft. Ana Serrano: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – revised draft. Miguel Martín-Retortillo: Conceptualization, Data curation, Investigation, Methodology, Writing – original draft, Writing – revised draft. Ignacio Cazcarro: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Writing – original draft, Writing – revised draft.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05055-z.

Correspondence and requests for materials should be addressed to I.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.