DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 52, ARTICLE 1, PAGES 1–24
PUBLISHED 3 JANUARY 2025
https://www.demographic-research.org/Volumes/Vol52/1
DOI: 10.4054/DemRes.2025.52.1

Data Description

A comprehensive database of estimates and forecasts of Spanish sex–age death rates by climate area, income level, and habitat size (2010–2050)

Celia Sifre-Armengol

Jose M. Pavía

Josep Lledó

Contents

Demographic Research: Volume 52, Article 1
Data Description

A comprehensive database of estimates and forecasts of Spanish sex–age death rates by climate area, income level, and habitat size (2010–2050)

Celia Sifre-Armengol[1]

Jose M. Pavía[2]

Josep Lledó[3]

Abstract

BACKGROUND
Analysing mortality is relevant for decision-making. Life tables have traditionally been based on age and sex, assuming homogeneous mortality rates within these groups. This omits other factors that could affect mortality risks. Advances in information technology and improved access to official microdata now enable the construction of life tables that incorporate additional variables, offering a more detailed analysis.

OBJECTIVE
This paper aims to expand the classical approach of using age and sex by integrating additional risk factors related to the area of residence. Specifically, the factors of climate, habitat size, and income are considered, using detailed georeferenced population data at the census level. Additionally, we aim to estimate future central death rates using various forecasting models.

METHODS
Utilising almost 2 billion microdata events from the Spanish population between 2010 and 2019, we begin by estimating new life tables that incorporate climate, habitat size, and income as risk factors. Then, after addressing random variations, erratic peaks, and the unexplained observed decline in mortality at extreme older ages, we use a triad of classical longevity models to project future mortality trends. All the generated data are offered in a public repository.

[1] Universitat de València, Valencia, Spain. Email: celia.sifre@uv.es. ORCID: 0009-0001-9588-6929.
[2] Universitat de València, Valencia, Spain. Email: pavia@uv.es. ORCID: 0000-0002-0129-726X.
[3] Universitat de València, Valencia, Spain. Email: josep.lledo@uv.es. ORCID: 0000-0002-7475-8549.

CONTRIBUTION

The database introduced in this paper can be used by social planners, demographers, and insurers, as well as being employed to validate existing findings and explore new research questions, particularly within the demographic and actuarial-economic fields.

1. Introduction

In the 21st century, developed nations are facing longevity challenges, characterised by higher life expectancy and ageing populations. This yields a complex set of consequences for wellbeing systems, with ramifications for the insurance industry and government welfare programmes, including social support services, healthcare infrastructure, and public pension plans (De Waegenaere, Melenberg, and Stevens 2010). Within this context, demographic analyses and population projections emerge as indispensable tools for understanding and reducing the uncertainties of future societal demands. As changes in population size and composition give rise to social, economic, environmental, and political challenges (George et al. 2001), failure to adequately anticipate the structure and magnitude of future population cohorts and their needs may cause added stress on public budgets and available infrastructure as a result of a deficit of resources. Demographic forecasting and projection facilitate the decision-making processes of public and private agents, helping them plan for and adjust to situations such as future medical needs and costs (Miller 2001), as well as prevent rural depopulation and its consequences (ESPON 2017).

To accurately project and forecast longevity and population cohorts across an entire territory and its sub-areas, georeferenced stocks of population must be effectively combined with geo-detailed predictions of death and birth rates and of domestic and foreign migration flows that take into account the particular trends and idiosyncrasies of the different sub-populations. In this study we focus on Spain and concentrate on mortality by estimating (2010–2019) and forecasting (2020–2050) central mortality rates and probabilities of death by age and sex, while also considering other area-related, contextual factors (measured using climate, income, and habitat size variables) that impact on mortality. This paper details the data, methods, and procedures followed to attain the estimates and forecasts and describes the database (available at https://data.mendeley.com/datasets/jbtwjbgx5f/3) with the obtained values. These data may be incorporated, among other processes, in the calculus of insurance premiums and annuities, to obtain georeferenced projections and forecasts of population, or for assessing the COVID-19 impact on mortality in Spain by age group and geographical area.

Central mortality rates, m_x, and probabilities of death, q_x, are the building blocks for constructing life tables. From any of these variables, other biometric variables are derived based on their interrelationships, among them the widely used life expectancy statistics. Traditionally, life tables have been constructed by age and sex, given the significance of these two variables in mortality patterns. This approach assumes homogeneity in mortality rates within the same age–sex group, overlooking other important variables that could impact on mortality risks. The evolution of information technologies and the improvement in the access to official microdata opens up new opportunities to construct and project new life tables based on additional variables of interest, such as income level, climatic conditions, or size of habitat.

Income levels and wealth inequalities have been shown to affect the health and mortality of populations, impacting on their death risks (see, e.g., Babones 2008; Bosworth, Burtless, and Zhang 2016). Air temperature and precipitation have also been linked to death risks, with both cold and heat waves modifying average probabilities of death. Both heat-related mortality in specific cities (Basu and Samet 2002) and a relationship between hot nights and mortality in the south of Europe (Royé et al. 2021) have been reported, while winter has been identified as a particularly risky season for older people (Pavía and Lledó 2022). Size of habitat has also been linked to mortality: Living conditions and access to health and care services are different for populations living in rural areas, with lower habitat sizes, compared to larger cities, and this affects death rates and mortality ratios (Elizalde and Díaz 2016).

It is widely recognised that health and mortality are closely linked to many other socioeconomic variables and lifestyles, not just those mentioned above. While we understand that a vast array of variables can influence these phenomena, obtaining the necessary data can sometimes be challenging. Consequently, in this study we only focus on income, climate, and habitat size, relying on detailed georeferenced population statistics at the census section level and the possibility offered by current statistical systems and tools to couple these risk factors with geographical units. In future research we aim to incorporate additional data to explore the relationship between mortality and other social and economic factors. Hence, this study opens up new opportunities for extensive research in the demographic and actuarial–economic fields.

From a methodological perspective, we first construct new life tables, segmented by age, gender, and the three factors mentioned above (income level, climate, and habitat size) for the Spanish population from 2010 to 2019. We choose the highest level of territorial disaggregation available in Spain, the census section, and adapt the estimators developed in Pavía and Lledó (2022). In total, we obtain 32,400 crude/raw estimates of central mortality rates, covering the age range 0 to (around) 107, after processing around 2 billion (2,000 million) microdata events. We then project/forecast central death rates using various forecasting models (Lee and Carter 1992; Renshaw and Haberman 2006;

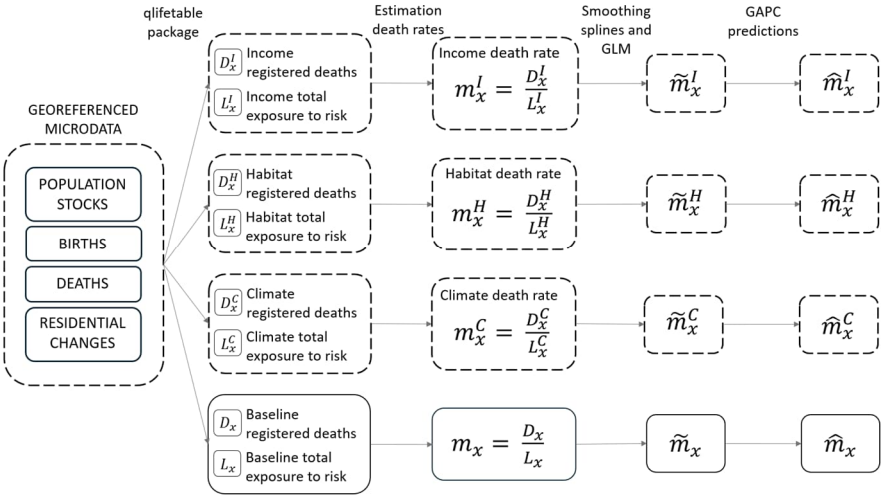
Plat 2009). Before projecting the raw rates, we smooth them to mitigate issues with outliers and random fluctuations and also use a model to correct the estimated trend for older ages, since the latter is more affected by raw data quality issues and typically exhibits greater variability in estimates due to the smaller size of the surviving cohorts. We leverage the model to extend estimates up to the age of 120, yielding a file with a total of 36,300 final estimates of central death rates. The second file contains, in total, 337,590 projected/forecasted death rates from 2020 to 2050 classified by age, gender, and the three abovementioned risk factors and forecasting model.

The rest of the paper is structured as follows. Section 2 details the data and methodological procedures used to calculate the life tables and their projections, taking into account not only the traditional variables of age and sex but also income level, climatological conditions, and habitat size of the area of residence. This is achieved using spatial tools and geolocation to segment the Spanish population. Section 3 provides a summary of the two files, describing all the relevant statistics and outcomes obtained in this study. Section 4 presents and explains the results of various validation checks performed on the processed data. The final section briefly explores potential applications of the data and explains some of the limitations of the study. All results, including estimates and forecasts, are publicly available, stored in a non-commercial repository, at https://data.mendeley.com/datasets/jbtwjbgx5f/3.

2. Data and methods

The IT revolution offers an unprecedented opportunity to collect, store, and transmit detailed data, particularly in the demographic and actuarial fields. Data, however, are not enough. Raw data need to be processed to gain meaningful insights, extract valuable knowledge, and assist decision-making. Estimating and projecting mortality rates over time requires processing and transforming microdata into summary statistics that capture risk exposure times and specific events, such as deaths, across different levels of risk factors. This involves several steps, including data cleaning, categorisation, and statistical analysis, to ensure accurate and reliable projections. Figure 1 illustrates the full process schematically.

Figure 1: A schematic summary of the workflow for creating mortality rate projections by risk factor



2.1 Data

All demographic georeferenced microdata used in this research were provided by the Spanish National Institute of Statistics (INE) as part of the project BE092-093-2021, 'Efectos Sociales y Económicos de las Fluctuaciones Intra-anuales en los Eventos Demográficos.' The data were provided by INE after submitting a research plan with administrative approval and making an advance payment. Specifically, we obtained the following datasets: (1) individual records of deaths, including the exact date of birth and death and the census section of residence of each deceased; (2) individual records of births by date of birth and municipality; (3) internal and external residential variations by municipality, including the date of birth and residential variation; and (4) population stocks as registered on 1 January each year, containing for each resident his/her date of birth and census section of residence. The datasets used in this study are sex-disaggregated and were provided by INE under a confidentiality agreement, in full compliance with Spain's data protection law, which prohibits the disclosure of individual records or cross-referencing personal data. The publication of aggregated data and derived results, such as those we have produced, is permitted.

Each demographic microdatum is processed a total of four times: once when estimating age–sex death rates conditioned on each risk factor (income level, habitat size,

and climate zone) and again when estimating base or baseline death rates, which does not account for any of the risk factors. Each of the risk factors is divided into different levels, allowing the population to be segmented at higher levels of homogeneity. The segmentations, in addition to considering gender and age, allow the population to be grouped based on specific levels of each variable.

2.2 Segmentation by risk factors

Stratification by income divides the population into four groups based on quartiles of per capita income measured at the census section of residence: low, low-medium, medium-high, and high income levels. The variable is provided by INE, free of charge. Each individual is assigned to an income level according to her/his census section of residence, so this variable can be considered as capturing 'contextual wealth effects' of the area of residence rather than the effect of individual income on mortality risks. Indeed, since recent research suggests that Spanish census sections are becoming more internally homogeneous due to residential segregation processes, exacerbating economic and social disparities between areas (Lledó and Pavía 2024), this partition also offers the advantage of capturing contextual area-related factors, such as the socioeconomic conditions of the neighbourhoods, which also influence death probabilities.

Census section per capita income has been categorised into only four levels to prevent the small group sizes that could result from considering additional levels, which would increase variability. It is important to note that the analyses are conducted for each individual age and sex without aggregating age groups, and that income groups are determined on a yearly basis, from 2010 to 2019. Although the classifications remain very stable over time, this means that a specific census section (and its residents) could belong to different risk groups in different years. These issues will be further discussed in Section 5, where we present the limitations of our study.

The climate areas are determined using a set of climate-related variables, which include temperature, precipitation, wind velocity, and atmospheric pressure, collected daily (from 1 January 2010 to 31 December 2021) at 272 meteorological stations. These data were obtained from the Spanish State Meteorological Agency (AEMET) and were extrapolated to each census section using Kriging (Matheron 1963). Cluster analysis was employed to analyse all the climatological data, identify the optimal number of groups, and create the clusters. In particular, census sections were grouped using the k-means algorithm. The final climate division categorises Spain into 6 spatially climatological zones/areas, which we call the Mediterranean climate area (including the Spanish coastal areas of the Mediterranean Sea), Subtropical climate area (the Canary Islands), Central-continental climate area (the very central part of Spain surrounding Madrid), North-

continental climate area (areas nearby and in Aragon and Castile-Leon), South-continental climate area (areas nearby and within Andalusia, excluding its coastal areas, as well as Extremadura and Castile La Mancha), and Oceanic climate area (Atlantic and Cantabrian coastal areas). Each person is assigned to a climatological area following his/her census section of residence. This process was conducted for each year, comparing the obtained clusters, which remain unchanged, thereby ensuring a consistent classification over time. This means that although temperatures may rise, the invariable aggregation of areas with a similar climate guarantees that this increase will similarly impact all its sub-regions, so the classification is not affected.

It is important to note that our approach differs from the well-known global climate classification system provided by the Köppen-Geiger maps (Peel, Finlayson, and McMahon 2007), based on 0.5° grids, and their recent adaptation into 1-kilometre grids (Beck et al. 2023). Unlike these global systems, which are designed to categorise climate zones based on broad environmental criteria, our method focuses on a more tailored division suited to the specific characteristics and needs of our area of analysis, Spain.

The population is also divided by habitat size into 4 levels based on the registered population in each municipality on the first day of each year. Municipalities are grouped as follows: those with fewer than 10,000 inhabitants, those with more than 10,000 but fewer than 50,000 inhabitants, those with more than 50,000 but fewer than 250,000 residents, and those municipalities with more than 250,000 residents. Each demographic event is assigned to a habitat size in line with the size of the municipality where it is recorded.

2.3 Estimates of raw death rates

Once demographic events are divided into groups, according to the levels of the factor under consideration (where it should be noted that residential variations between groups are observed as external movements), the next step is to estimate (crude) central death rates by age and sex. The death rate at a specific age x, m_x, is estimated as the quotient between the number of deaths recorded at age x, D_x, and the total exposure to death at age x, L_x. Our study has an annual periodicity, with crude death rates estimated for each age x (with ages ranging from 0 to 107), sex s (men and women), and year t (where t ranges from 2010 to 2019). Predictions extend this timeframe to the year 2050. The estimations of total exposure are based on population stocks, number of deaths, and residential changes or migration flows (Pavía and Lledó 2022). For $x > 0$, the total exposure at risk for group f_i is estimated as follows:

$$L^{f_i}_{x,s,t} = \sum_{d=1}^{T} \sum_{j=1}^{P^{f_i}_{x,d,s,t}} \frac{(T-d-\zeta_j)}{T} + \sum_{d\leq\tau}^{T} \sum_{j=1}^{I^{f_i}_{x+d,\tau,s,t}} \frac{(T-\tau-\zeta_j)}{T} + \sum_{d>\tau}^{T} \sum_{j=1}^{I^{f_i}_{x+d,\tau,s,t}} \frac{(T-d-\zeta_j)}{T} -$$
$$\sum_{d\leq\tau}^{T} \sum_{j=1}^{E^{f_i}_{x+d,\tau,s,t}} \frac{(T-\tau-\zeta_j)}{T} - \sum_{d>\tau}^{T} \sum_{j=1}^{E^{f_i}_{x+d,\tau,s,t}} \frac{(T-d+\zeta_j)}{T} - \sum_{d\leq\tau}^{T} \sum_{j=1}^{D^{f_i}_{x+d,\tau,s,t}} \frac{(T-\tau-\zeta_j)}{T} -$$
$$\sum_{d>\tau}^{T} \sum_{j=1}^{D^{f_i}_{x+d,\tau,s,t}} \frac{(T-d+\zeta_j)}{T} + \sum_{d=1}^{T} \sum_{j=1}^{P^{f_i}_{x-1,d,s,t}} \frac{(d+\zeta_j)}{T} + \sum_{d>\tau}^{T} \sum_{j=1}^{I^{f_i}_{x-1+d,\tau,s,t}} \frac{(d-\tau+\zeta_j)}{T} -$$
$$\sum_{d>\tau}^{T} \sum_{j=1}^{E^{f_i}_{x-1+d,\tau,s,t}} \sum_{d>\tau}^{T} \frac{(d-\tau+\zeta_j)}{T} - \sum_{d>\tau}^{T} \sum_{j=1}^{D^{f_i}_{x+d,\tau,s,t}} \frac{(T-d+\zeta_j)}{T}, \quad (1)$$

and when $x = 0$, also considering births, through:

$$L^{f_i}_{0,s,t} = \sum_{d=1}^{T} \sum_{j=1}^{P^{f_i}_{0,d,s,t}} \frac{(T-d-\zeta_j)}{T} + \sum_{\tau=1}^{T} \sum_{j=1}^{B^{f_i}_{\tau,s,t}} \frac{(T-\tau-\zeta_j)}{T} + \sum_{d\leq\tau}^{T} \sum_{j=1}^{I^{f_i}_{d,\tau,s,t}} \frac{(T-\tau-\zeta_j)}{T} +$$
$$\sum_{d>\tau}^{T} \sum_{j=1}^{I^{f_i}_{d,\tau,s,t}} \frac{(T-d-\zeta_j)}{T} - \sum_{d\leq\tau}^{T} \sum_{j=1}^{E^{f_i}_{d,\tau,s,t}} \frac{(T-\tau-\zeta_j)}{T} - \sum_{d>\tau}^{T} \sum_{j=1}^{E^{f_i}_{d,\tau,s,t}} \frac{(T-d+\zeta_j)}{T} -$$
$$\sum_{d\leq\tau}^{T} \sum_{j=1}^{D^{f_i}_{d,\tau,s,t}} \frac{(T-\tau-\zeta_j)}{T} - \sum_{d>\tau}^{T} \sum_{j=1}^{D^{f_i}_{d,\tau,s,t}} \frac{(T-d+\zeta_j)}{T}, \quad (2)$$

where f_i denotes the level i of risk factor f, $P^{f_i}_{x,d,s,t}$ stands for the number of individuals (population stock) with exact age $x + d$ at the beginning of the year, $I^{f_i}_{x+d,\tau,s,t}$, $E^{f_i}_{x+d,\tau,s,t}$ and $D^{f_i}_{x+d,\tau,s,t}$ represent the number of immigrants, emigrants, and deaths, respectively, recorded with exact age $x + d$ on the calendar day τ with sex s in year t, $B^{f_i}_{\tau,s,t}$ designates the number of births registered on calendar day τ, T is the total number of days in the calendar year, and ζ_j is a random number from a uniform distribution between 0 and 1, introduced to avoid systematic bias, that simulates the exact time within a day when a demographic event (birth, emigration, or death) occurs.

Central mortality rates are computed as quotients between the number of deaths recorded at age x in year t for each level i of risk factor f, $D^{f_i}_{x,s,t}$, and $L^{f_i}_{x,s,t}$, the number of person years from which $D^{f_i}_{x,s,t}$ occurred, as follows:

$$\widehat{m}^{f_i}_{x,s,t} = \frac{D^{f_i}_{x,s,t}}{L^{f_i}_{x,s,t}}. \quad (3)$$

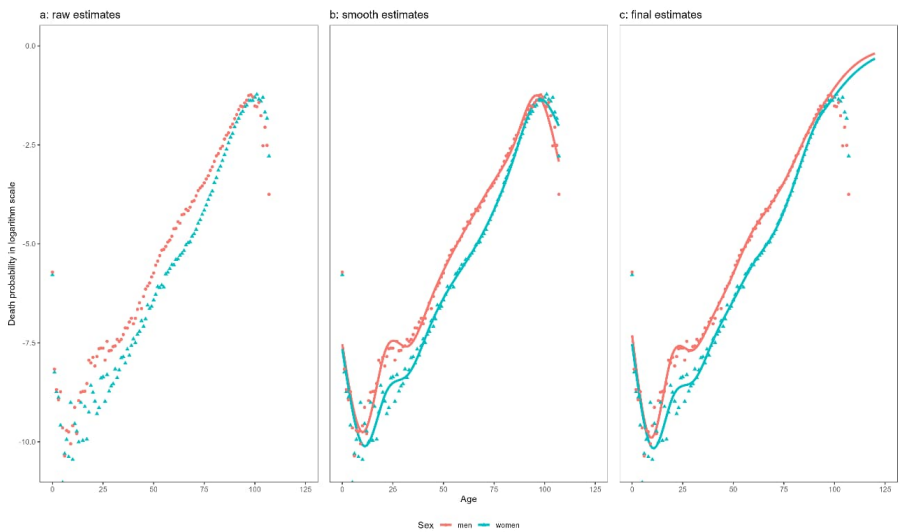
The estimation of the mortality rates for all the combinations of year, age, sex, and level of risk factor requires the handling of a huge quantity of microdata, a total of 2 billion microdata events. To manage this, we have developed scripts ad hoc in the free statistical software R (R Core Team 2024) with the help of packages *dplyr* (Wickham et

al. 2023) for performing data wrangling, and *qlifetable* (Pavía and Lledó 2024) for calculating times of exposure and counting the number of deaths.

2.4 Smoothing death rates

Given the continuous nature of human ageing and the gradual improvements in healthcare, nutrition, and living conditions that typically occur over extended periods, mortality risks are expected to evolve smoothly as a function of both age and time. However, raw estimates of central death rates are rarely that smooth, often showing erratic peaks and random fluctuations. An example of this can be seen in the series of raw estimates displayed in Figure 2a, where crude death probabilities by sex (with men as red points and women as blue triangles) corresponding to the year 2019 are displayed in the log scale for the low-income group. If these irregular fluctuations in the mortality trends are not properly addressed, they will eventually affect future predictions.

Figure 2: Estimates of raw and smoothed death probabilities by sex (women: blue triangles; men: red points) for the low income level during 2019



Note: Raw estimates (panel a) were obtained after log-transforming the death probabilities implied by the estimated death rates attained using Equation (3). Intermediate smooth estimates (panel b) were obtained by fitting raw estimates using cubic natural splines. Final estimates (panel c) were obtained by fitting a GLM model for extreme ages and after imposing a gradual transition between smooth and model estimates. All the probabilities are presented in log-scale, with the R package *ggplot2* (Wickham 2016) used for visualization. *Source*: Compiled by the authors.

Therefore, it is essential to correct anomalies and random variations, eliminate unusual values, and achieve a smooth trend that more accurately reflects the underlying risk process. To manage this issue, demographers and actuaries routinely apply smoothing techniques (see, e.g., D'Amato, Piscopo, and Russolillo 2011): we apply natural cubic splines (Reinsch 1967) as a smoothing technique. The aim of this technique is to minimise the sum of squared residuals while penalising excessive curvature to prevent overfitting. An example of smoothness of crude series of death probabilities can be seen in Figure 2b, which shows the smoothed death probabilities corresponding to the raw estimates drawn in Figure 2a.

2.5 Mortality at the oldest ages

Once smoothing splines are applied to fix random fluctuations and outliers, we then consider the estimates corresponding to the smallest cohorts – those with a small number of person-years exposed to risk of dying. In our estimates we detect an illogical decreasing mortality trend for the oldest extreme ages, likely caused by an accumulation over time of small inaccuracies due to unidentified deaths not being recorded in the population registers (Cairns et al. 2016). These inaccuracies inflate the number of survivors at older ages, particularly noticeable when the actual number of survivors is small. To rectify this trend and also extend estimates of death probabilities up to age 120, we model the estimated trend in death probabilities within each sex and group for the oldest ages. As life expectancy has been rising every year, to properly forecast death rates over the years to come, it seems appropriate to extend the age range of death-rate estimates beyond the oldest age available in our database, which is currently 107 years.

When modelling estimated death probabilities, we intentionally exclude ages where illogical estimates are observed. Specifically, taking as global reference the minimum age by sex across groups when the crude estimated mortality trends start to decline, we consider the previous ten ages for modelling. This results in modelling the range of ages 87 to 97 for males and 90 to 100 for females.

For modelling, we leverage the fact that the biometric variable $D_{x,s,t}^{f_i}$ – which measures the number of deaths at age x and year t for the risk factor f and level factor i for sex s – theoretically follows a binomial distribution with size being (approximately) the semi-sum of the total person-years exposed to risk with ages $x-1$ and x, $\frac{1}{2}(L_{x-1,s,t}^{f_i} + L_{x,s,t}^{f_i})$, and probability parameter being the death probability, $q_{x,s,t}^{f_i}$, and specify a generalised linear model (GLM) with time as predictor and $D_{x,s,t}^{f_i}$ as a binomial response variable. This model is used to obtain (initial) estimates of death probabilities from age x_0 to 120 (with x_0 being equal to 87 for males and 90 for females), which are combined

with the smoothed estimates. To guarantee a smooth transition between smoothed and model estimates, we calculate, for the range of ages $x = x_0, x_0 + 1, \ldots, x_0 + 10$, final estimates as weighted averages of smoothed and model estimates with respective weights equal to $1 - (x - x_0)/10$ and $(x - x_0)/10$. As an example of final estimates of death probabilities, Figure 2c shows the final death probabilities corresponding to the raw estimates drawn in Figure 2a.

2.6 Forecasts of death rates

The last part of the process (see Figure 1) consists of projecting death rates. This topic has received significant attention in the literature over recent years due to the growing interest in longevity and its associated challenges, and the potential offered by models to predict future mortality trends and patterns. Many mortality models observe the rates as a function of age, period, and cohort. Accordingly, Hunt and Blake (2020) developed a taxonomy of models with these dimensions in mind. Nevertheless, all the models can be thought of as particular instances of generalised age–period–cohort models (GAPC) (Villegas, Millossovich, and Kaishev 2018), whose general predictor structure is:

$$\eta_{x,t} = \alpha_x + \sum_{i=1}^{N} \beta_x^{(i)} \kappa_t^{(i)} + \beta_x^{(0)} \gamma_{t-x}, \tag{4}$$

where $\eta_{x,t}$ is the transformation of the death rate at age x and time period t; α_x denotes the general mortality shape over a specific age x and across all ages; $\beta_x^{(i)} \kappa_t^{(i)}$ is the product of the time trend κ_t with β_x denoting the mortality changes across ages; and $\beta_x^{(0)} \gamma_{t-x}$ is the product between the cohort term, γ_{t-x}, and an age function $\beta_x^{(0)}$.

The GAPC models selected to forecast in our application are the Lee–Carter model (Lee and Carter 1992), probably the most well-known mortality forecasting model; the age–period–cohort (APC) model, which is a substructure of the model proposed by Renshaw and Haberman (Renshaw and Haberman 2006); and the Plat model (Plat 2009), which combines two age–period terms with a number of Lee-Carter characteristics. Once the predictions up to year 2050 for each model are estimated, we then apply splines to obtain a more realistic and smoother trend. We consider having several death rate predictions to be a richness in data as this allows model uncertainties to be incorporated into the analyses. This approach provides more flexible and robust forecasts, enabling better planning for future scenarios and enhanced decision-making.

3. Estimated and forecasted death rates

The previous section details the sequential process followed to estimate and project central mortality rates by age and sex up to 2050, conditioned on the three risk factors under consideration (income, habitat size, and climate area), using georeferenced microdata from the population of Spain. This section describes the two open-format files (available at https://data.mendeley.com/datasets/jbtwjbgx5f/3) generated after applying the process illustrated in Figure 1. The files are organised by variables, in a long–wide format.

The first file, called *Estimates of death rates, Spain 2010–2019, by risk factor.csv*, offers the results of converting nearly two billion microdata events into estimates of central mortality rates for each risk factor, categorised according to various variables. Table 1 provides an extract from the file, which has a total of 36,300 rows (central mortality rate estimates) of which 32,400 correspond to raw central mortality rate estimates. The first two variables (columns) in the file indicate the time dimensions of the estimates: age (from 0 to 120 years) and $year$ (from 2010 to 2019). Following the variable for sex, there are two variables that determine the risk profile: $risk\_factor$ and $factor\_level$. The $risk\_factor$ variable indicates the risk factor (income, habitat size, and climate), while the $factor\_level$ variable designates the corresponding level of the factor, according to the following keys:

- Climate. C1: Central-continental, C2: Canary Islands, C3: Mediterranean, C4: North-continental, C5: South-continental and C6: Oceanic.
- Habitat. H1: fewer than 10,000 inhabitants, H2: more than 10,000 but fewer than 50,000 inhabitants, H3: more than 50,000 but fewer than 250,000 residents, and H4: more than 250,000 residents.
- Income. I1: low, I2: low-medium, I3: medium-high, and I4: high.

The baseline estimates are identified in the $risk\_factor$ variable using the same name, with no level in the $factor\_level$.

The columns identified with the biometric variables Lx, Dx, mx_raw correspond to the results from applying formulas (1), (2), and (3), respectively. The variable mx_smooth contains the values obtained after applying the smoothing techniques and correcting for inconsistencies at extreme ages. The values for the columns of raw and smoothed probabilities (qx_raw and qx_smooth), which indicate the probability that an individual aged x will not reach age $x + 1$, q_x, are calculated using the relation between q_x and m_x under the assumption of uniform intra-age distribution of deaths:

$$q_x = \frac{m_x}{1+0.5 \cdot m_x}.$$

Table 1: Extract of the file corresponding to estimates of death rates

age	Year	sex	risk_factor	factor_level	Lx	Dx	mx_raw	qx_raw	qx_smooth	mx_smooth
0	2010	men	climate	C1	44,820.40	175	0.003904	0.003897	0.000540	0.000540
0	2010	men	climate	C2	8,556.78	32	0.003740	0.003733	0.000381	0.000381
…	…	…	…	…	…	…	…	…	…	…
50	2015	women	habitat	H1	72,245.40	125	0.001730	0.001729	0.001488	0.001489
50	2015	women	habitat	H2	97,014.65	143	0.001474	0.001473	0.001501	0.001502
…	…	…	…	…	…	…	…	…	…	…
65	2019	women	income	I3	62,690.12	369	0.005886	0.005869	0.005127	0.005141
65	2019	women	income	I4	74,788.67	321	0.004292	0.004283	0.004652	0.004663
…	…	…	…	…	…	…	…	…	…	…
45	2015	men	baseline		383,399.13	650	0.001695	0.001694	0.001715	0.001716
45	2016	men	baseline		385,144.65	631	0.001638	0.001637	0.001648	0.001650
…	…	…	…	…	…	…	…	…	…	…

Note: The file, which contains 36,300 rows, includes raw and (final) smoothed estimates of central mortality rates and probabilities of death for the years 2010 to 2019, categorised by age, sex, and risk factor, with risk factors divided into several levels: climatic condition (C1, C2, C3, C4, C5, C6), habitat size (H1, H2, H3, H4), and income level (I1, I2, I3, I4). The table also presents risk exposures, calculated using formula (1), number of deaths, and estimates without dividing the population by risk factors.

The second file, called *Forecasts of death rates, Spain 2020–2050, by risk factor.csv*, includes the projections of the death rates from 2020 to 2050. In total, it contains 337,590 projected central mortality rates, segmented by risk factors and forecasting model. Table 2 presents a sample of rows from this file. This file shares some variables with the previous file and includes the variable f_model which informs about the model used for forecasting: LC, APC, or PLAT.

Table 2: Extract of the file corresponding to forecasts of death rates

age	year	sex	risk_factor	factor_level	f_model	mx	qx	qx_smooth	mx_smooth
0	2020	men	climate	C1	LC	0.000504	0.000504	0.000502	0.000502
0	2020	men	climate	C2	LC	0.000538	0.000538	0.000536	0.000536
…	…	…	…	…	…	…	…	…	…
50	2030	women	habitat	H1	APC	0.000951	0.000951	0.000952	0.000952
50	2030	women	habitat	H2	APC	0.001008	0.001008	0.001009	0.001009
…	…	…	…	…	…	…	…	…	…
65	2026	women	income	I3	PLAT	0.005036	0.005024	0.005025	0.005037
65	2026	women	income	I4	PLAT	0.004288	0.004279	0.004278	0.004287
…	…	…	…	…	…	…	…	…	…
45	2032	women	baseline		LC	0.000836	0.000836	0.000835	0.000836
45	2032	women	baseline		APC	0.000661	0.000661	0.000657	0.000657
…	…	…	…	…	…	…	…	…	…

Note: The file, which contains 337,590 rows, includes initial and smoothed predictions/projections of central mortality rates and probabilities of deaths for the years 2020 to 2050, categorised by age, sex, risk factor, and forecasting model (LC, APC, and PLAT). Each risk factor is divided into several levels: climatic condition (C1, C2, C3, C4, C5, C6), habitat size (H1, H2, H3, H4), and income level (I1, I2, I3, I4). The table also presents baseline forecasts, obtained without dividing the population by risk factors.

4. Data validation

In this research we have processed and summarised an extensive database consisting of over 1,997 million microdata events. The process involved multiple stages, each requiring rigorous data validation controls to guarantee accuracy and reliability. In this section we detail the main validation tests performed. Initial checks involved verifying the correctness of recorded dates in each microdata entry. Birth dates recorded in both population registers and birth datasets must not be later than the date of data extraction. For population movements, including deaths, emigration, and immigration, the date of birth must not occur after the date of death or the date of the movement. Also, the artificially inflated record of immigrants being given the first day of the year as their birth date (Lledó, Pavía, and Simó-Noguera 2024) is taken into account.

The levels for each risk factor are assigned to each individual based on the geolocation attached to the microdata. Income levels and climatological areas are determined by the census section, while habitat size is assigned based on the municipality of residence. Using geolocated microdata, we generate the numerators, $D_{x,s,t}^{f_i}$, and denominators, $L_{x,s,t}^{f_i}$, needed to calculate the central mortality rates across groups. Two validation controls are performed during this process. First, we compare the total number of deaths by age, sex, and risk factor to guarantee they match the baseline data. Second, we verify that the time of risk exposure, categorised by age, sex, and risk factor, aligns with the baseline risk exposures. We find no discrepancies for counts of number of deaths and extremely small discrepancies for time of exposure. Some discrepancies for the latter variable are expected even with perfect programming and data. Table 3 displays for each risk factor and year the relative differences between the aggregates of time exposed by level and age for each risk factor and the corresponding aggregates without dividing the population for level of risk factor, with discrepancies calculated as in Equation (5):

$$2\,\frac{\left[\left(\sum_{i=1}^{n_f} L_{x,s,t}^{f_i}\right) - L_{x,s,t}^{baseline}\right]}{\left(\sum_{i=1}^{n_f} L_{x,s,t}^{f_i}\right) + L_{x,s,t}^{baseline}}, \tag{5}$$

where f denotes the risk factor (income, habitat, or climate) with i being one of its levels, $i \in \{1, \cdots, n_f\}$; n_f the number of levels; x age, with x being a natural number between 0 and 107; s sex; and t the time period, where t is a year between 2010 and 2019.

Table 3: Relative differences between the time exposed to each risk factor and the baseline

	Habitat size		Income		Climate	
Year	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**
2010	0.00113	0.00206	0.00109	0.00202	0.00081	0.00169
2011	0.00135	0.00215	0.00116	0.00201	0.00064	0.00147
2012	0.00050	0.00146	0.00041	0.00134	0.00003	0.00092
2013	–0.00051	0.00034	–0.00064	0.00020	–0.00119	–0.00035
2014	–0.00035	0.00032	–0.00035	0.00029	–0.00062	–0.00001
2015	0.00096	0.00137	0.00082	0.00123	0.00028	0.00068
2016	0.00170	0.00223	0.00160	0.00209	0.00113	0.00162
2017	0.00230	0.00286	0.00224	0.00278	0.00191	0.00241
2018	0.00445	0.00451	0.00439	0.00446	0.00417	0.00420
2019	0.00661	0.00628	0.00646	0.00613	0.00595	0.00562

The minor differences observed in Table 3 could be attributed to the random component assigned to the exact time of occurrence of each demographic event (birth, death, and movements) within a given day and to the fact that the effective number of residential variations is different, conditioned on each risk factor. For instance, a residential variation between different census sections within the same climatological area is omitted when the risk factor is climate, but is taken into account when the risk factor under consideration is income and it entails a change between income levels.

In addition to conducting internal validations, we have also carried out external validations. Specifically, we have compared the observed and predicted death probabilities from our study with those provided by other organisations, as summarised in Tables 4 and 5. Table 4 presents the absolute differences between the death probabilities available for Spain in the HMD database (Human Mortality Database 2024) and our baseline estimates. Table 5 summarises the absolute differences between the World Population Prospects (WPP) predictions and our model predictions, also considering our baseline data. Table 4 presents comparisons by year (2010–2019), including the overall average, across sexes, and with individuals grouped into 5 age categories. By contrast, Table 5 aggregates data into 10-year intervals, using slightly different age groups. The maximum age in Table 5 is 99, as WPP life tables end at that age.

Table 4: Summary of discrepancies in absolute values with HMD-estimated death probabilities

	Men					Women				
Year	**0**	**[1, 85]**	**(85, 90]**	**(90, 100]**	**>100**	**0**	**[1, 85]**	**(85, 90]**	**(90, 100]**	**>100**
2010	0.00263	0.00018	0.00467	0.03050	0.01953	0.00243	0.00013	0.00175	0.03084	0.06305
2011	0.00255	0.00020	0.00549	0.03505	0.03390	0.00239	0.00013	0.00070	0.02676	0.03838
2012	0.00259	0.00019	0.00528	0.03815	0.03278	0.00227	0.00011	0.00128	0.03233	0.06168
2013	0.00222	0.00017	0.00504	0.03968	0.06054	0.00200	0.00012	0.00118	0.02977	0.06694
2014	0.00239	0.00020	0.00249	0.03135	0.02981	0.00218	0.00010	0.00115	0.02715	0.06624
2015	0.00232	0.00021	0.00401	0.03011	0.02352	0.00207	0.00008	0.00117	0.02915	0.06174
2016	0.00246	0.00023	0.00320	0.02740	0.02043	0.00184	0.00011	0.00137	0.02621	0.06220
2017	0.00233	0.00024	0.00230	0.02200	0.01850	0.00203	0.00010	0.00138	0.02515	0.05401
2018	0.00213	0.00021	0.00172	0.01619	0.0240	0.00206	0.00014	0.00143	0.02342	0.04864
2019	0.00228	0.00023	0.00225	0.01589	0.01894	0.00186	0.00010	0.00123	0.02151	0.05118
Average	0.00239	0.00021	0.00364	0.02863	0.02819	0 .00211	0.00011	0.00126	0.02723	0.05741

In Table 4, the largest differences are observed in the youngest age group (age 0) and the oldest age groups (90–100 and over 100 years). For age 0, the discrepancy arises from the different methods used to derive, from estimated death rates, death probabilities. HMD uses an approximation (Wilmoth et al. 2021), while our study uses microdata to calculate the average age at death for new-borns, a_0, specific to the Spanish population. For ages above 90, the differences result from distinct methodologies and model-fitting age ranges. HMD smooths death rates using a logistic curve from ages 80 to 110, assuming a Poisson distribution for deaths (Wilmoth et al. 2021), while our approach extends probabilities to age 120 with a generalised linear model with binomial response, fitted to the last 10 observed ages before death probabilities decline. In the core age range [1, 85], where no modelling was applied, the differences are negligible.

Similar to the findings for estimated death probabilities (see Table 4), the largest discrepancies in predictions (see Table 5) occur at older ages, with these differences increasing, as expected, over time. This variation arises from the differing methodologies applied to observed data for older age groups and the predictive model employed by WPP (United Nations, Department of Economic and Social Affairs, Population Division 2024), which does not align with any of our forecasting models. The WPP forecasts are generated using an extended Lee-Carter model (Li, Lee, and Gerland 2013). Notably, for the oldest ages, the smallest differences between WPP values and our forecasts are observed when we apply the Lee-Carter model. Again, the central age ranges show very small absolute differences.

Table 5: Summary of discrepancies in absolute values with WPP-predicted death probabilities

		Men					Women				
Model	**Year**	**0**	**[1, 65]**	**(65, 75]**	**(75, 85]**	**(85, 99]**	**0**	**[1, 65]**	**(65, 75]**	**(75, 85]**	**(85, 99]**
APC	2020–30	0.00193	0.00035	0.00130	0.00505	0.03780	0.00150	0.00011	0.00134	0.00202	0.02566
	2031–40	0.00142	0.00059	0.00202	0.00505	0.05873	0.00093	0.00011	0.00238	0.00471	0.03457
	2041–50	0.00115	0.00064	0.00453	0.00887	0.07272	0.00063	0.00008	0.00178	0.00806	0.03323
	Average	0.00150	0.00053	0.00262	0.00632	0.05642	0.00102	0.00010	0.00183	0.00493	0.03115
LC	2020–30	0.00182	0.00026	0.00239	0.00611	0.01382	0.00155	0.00015	0.00130	0.00497	0.01718
	2031–40	0.00123	0.00048	0.00501	0.01228	0.01097	0.00096	0.00030	0.00288	0.01028	0.01742
	2041–50	0.00089	0.00064	0.00706	0.01717	0.02431	0.00064	0.00043	0.00417	0.01477	0.02250
	Average	0.00131	0.00046	0.00482	0.01185	0.01637	0.00105	0.00029	0.00278	0.01001	0.01903
PLAT	2020–30	0.00194	0.00027	0.00138	0.00330	0.03280	0.00154	0.00013	0.00111	0.00243	0.02731
	2031–40	0.00145	0.00037	0.00123	0.00232	0.04234	0.00106	0.00012	0.00183	0.00382	0.03444
	2041–50	0.00119	0.00030	0.00137	0.00177	0.04270	0.00081	0.00010	0.00134	0.00748	0.03012
	Average	0.00153	0.00031	0.00133	0.00246	0.03928	0.00114	0.00012	0.00143	0.00458	0.03062

5. Data potentialities and limitations

Forecasted life tables are of value for many agents, including insurers, demographers, and social planners. For insurers, life tables are crucial for evaluating risks and determining life insurance premiums. Demographers use series of life tables to analyse population mortality trends, while social planners rely on them to anticipate future social needs. Examining a series of life tables can provide insight into mortality patterns over time and aid financial planning and public policy development.

In this study we have created two groups of life tables, one based on estimated, realised death rates, and the other with predicted rates. Additionally, we have developed life tables not only for the general population but also segmented by various risk factors such as income, climate, and habitat size. This stratification allows for the aggregation of data into more homogeneous groups, reducing heterogeneity when applied to individuals. Understanding these data is relevant, as it can directly improve the functioning of key economic sectors, including the insurance industry and the management of pension funds (Külekci and Selcuk-Kestel 2021).

Indeed, the potential applications of these two databases extend across multiple fields. For instance, they could be used to properly quantify expected increases in longevity risks, which have a profound impact on the healthcare system. The future health needs of an ageing population play a crucial role in healthcare planning and resource allocation. Anticipating these needs can significantly enhance the effectiveness of public health management. Our database can provide Spanish health managers with access to

future projections of mortality rates segmented by various risk factors, enabling more accurate forecasts for both the overall population and specific sub-groups. This allows for more effective resource allocation and infrastructure planning in geographical areas where future needs are expected to be greatest. Furthermore, when combined with data on migration patterns and birth rates, this information can provide valuable insights for forecasting future demands in educational infrastructure, childcare facilities, and transportation systems.

The generated datasets are also of value for the insurance industry, particularly for implementing better pricing schemes. Following the Test-Achats ruling, which prohibits the use of sex as a pricing factor in the European Union (EU), life insurance companies in the EU are, in practice, limited to using age as the primary pricing criterion. However, the assumption that all individuals of the same age share similar death risks a priori (i.e., without an initial health examination) is debatable, if not outright erroneous. Using our estimates, actuaries and risk managers could incorporate three, easily observable, additional risk factors – income level, habitat size, and climatic area of residence – into their life insurance pricing processes. This would enable better portfolio segmentation and more precise alignment of premiums with individual risk profiles, resulting in more accurate and equitable insurance premiums. In this context, the area of study initiated by Lledó and Pavía (2024) is noteworthy. They suggest modelling the differences between baseline and income-based death rates to develop a series of risk coefficients that capture the varying risks associated with different income levels (contextual wealth effects). This set of coefficients can be directly applied to any life table, whether for the general or insured population, to improve the calculation of premiums and reserves.

Our results, however, are not without limitations. As the unusual estimated mortality trends at the extreme oldest ages show, errors in official records are possible. Despite our thorough checks and validation processes, some errors may still remain undetected, potentially affecting the findings. While data accuracy is crucial in empirical research, issues with data quality are a recurring challenge in demographic studies (e.g., Cairns et al. 2009), though fortunately they typically have a very limited impact.

A limitation of our results is that mortality trends can be significantly affected by social, health, or economic crises, as well as by major medical advances, which can sometimes lead to structural changes (Preston 1977; Van Berkum, Antonio, and Vellekoop 2016). For example, future severe economic downturns, civil unrest, wars, or pandemics could transform our predictions into underestimates. A recent instance is the COVID-19 pandemic, which caused a temporary spike in death rates, particularly among individuals aged 50 and older (Lledó, Pavía, and Sánchez 2023). Indeed, to mitigate the impact of this unusual event, we deliberately exclude the years 2020 to 2021 from our estimates. This prevents contamination of future projections with the temporal anomalies produced by the COVID-19 pandemic. By contrast, significant improvements in the

public healthcare system or in medical advances could reduce future death rates, converting our forecasted death rates into overestimates. Furthermore, our climate classifications alone do not account for extreme weather events which can have a significant impact on health and increase mortality risk, especially among the more vulnerable populations. However, it is reasonable to expect that certain extreme conditions may be more likely to occur in specific climates. For example, as recent experience shows, severe heat waves are more likely to occur in regions with warmer climates (ECMWF 2024), such as the South-continental climate area, composed of areas near and within Andalusia (excluding its coastal areas) as well as Extremadura and Castile La Mancha.

When classifying mortality risks, an additional factor to consider is how risks are managed when either an individual or a census section transitions to a new mortality risk group. Specifically, the questions arise about which life table should be applied to the residents of a census section that moves, for example, from a medium-low income level to a medium-high income level, or which life table should be used for a person migrating between climate groups. Although census sections and habitat sizes show considerable stability regarding the risk factors analysed in this study, we adopt a similar approach to that followed in demography concerning foreign immigrants, where it is (implicitly) assumed that new residents face the same mortality risks as the native population. In this case, we consider that the person or group immediately adopts the risk profile of their new group, without accounting for past data, including their previous risk levels and history. This approach overlooks the potential long-term effects of prior exposure to different risk factors.

Another limitation of our study is that it does not address the construction of life tables that intersect multiple risk factors. This presents a potential avenue for future research, which could involve estimating life tables for combinations of mortality risk factors or developing an algorithm to calculate new coefficients that account for these interactions. In some regions of Spain these risk factors are highly correlated, and different risk factors often overlap within the same areas. Finally, it is important to note that income levels and habitat sizes have been categorised into only four groups to avoid dividing the population into segments that are too small, which could hinder the ability to obtain reliable results due to limited data. A potential solution to increase the number of groups and generate more adjusted estimates could be to use 5-year age groupings, as this would help enlarge sample sizes and would probably not affect accuracy. Increasing the number of groups could also enable the use of modelling in a second step, to exploit the underlying continuous nature of these factors. This is certainly a topic that warrants further research in the future.

6. Data statement

The raw data used in this research include (1) birth statistics, (2) stocks of population, (3) microdata of deaths, (4) residential variations microdata, (5) per capita income statistics and (6) climatological data. Demographic microdata was acquired from INE (the Spanish official statistics office) in compliance with their approved procedures, following a research-focused request and payment in advance (<https://ine.es/infoine/>). Per capita income statistics are available at the link: <https://links.uv.es/eoHKMd0>. Climatological data are available at AEMET <https://www.aemet.es/es/portada>. The files with the results described in this research are available at the link https://data.mendeley.com/datasets/jbtwjbgx5f/3. The degree of accuracy/reliability of the results derived from the exploitation of the data provided by INE is the sole responsibility of the authors.

7. Acknowledgments

Funding: This study was supported by Generalitat Valenciana (grants HIECPU/2023/2, Conselleria de Hacienda, Economía y Administración Pública, and CIGE/2023/7 and CIACO/2023/031, Conselleria de Educación, Cultura, Universidades y Empleo, Fundación Mapfre (grant *Modelización espacial e intra-anual de la mortalidad en España. Una herramienta automática para el cálculo de productos de vida*) and Ministerio de Ciencia e Innovación (grant PID2021-128228NB-I00).

Authors' contribution: Conceptualization: JMP, JL. Data curation: CSA, JL. Formal analysis: CSA, JMP, JL. Funding acquisition: JMP, JL. Investigation: CSA, JMP, JL. Methodology: CSA, JMP, JL. Project administration: JMP. Software: CSA, JL. Supervision: JMP, JL. Validation: CSA, JMP. Visualization: CSA, JL. Writing – original draft: CSA, JMP, JL. Writing – review and editing: CSA, JMP, JL.

The authors wish to thank Marie Hodkinson for revising the paper's English, INE for providing the demographic raw data analysed in this research, and an associate editor and two reviewers for valuable comments and suggestions.

References

Babones, S. (2008). Income inequality and population health: Correlation and causality. *Social Science and Medicine* 66(7): 1614–1626. doi:10.1016/j.socscimed.2007.12.012.

Basu, R. and Samet, J. (2002). Relation between elevated ambient temperature and mortality: A review of the epidemiologic evidence. *Epidemiologic Reviews* 24(2): 190–202. doi:10.1093/epirev/mxf007.

Beck, H.E., McVicar, T.R., Vergopolan, N., and et al. (2023). High-resolution (1 km) Köppen–Geiger maps for 1901–2099 based on constrained CMIP6 projections. *Scientific Data* 10(1): 724. doi:10.1038/s41597-023-02549-6.

Bosworth, B., Burtless, G., and Zhang, K. (2016). Later retirement, inequality in old age, and the growing gap in longevity between rich and poor. *Economic Studies at Brookings* 87: 1–166.

Cairns, A., Blake, D., Dowd, K., Coughlan, G., Epstein, D., Ong, A., and Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal* 13(1): 1–35. doi:10.1080/10920277.2009.10597538.

Cairns, A.J.G., Blake, D., Dowd, K., and Kessler, A.R. (2016). Phantoms never die: Living with unreliable population data. *Journal of the Royal Statistical Society Series A: Statistics in Society* 179(4): 975–1005. doi:10.1111/rssa.12159.

D'Amato, V., Piscopo, G., and Russolillo, M. (2011). The mortality of the Italian population: Smoothing techniques on the Lee–Carter model. *The Annals of Applied Statistics* 5(2A): 705–724. doi:10.1214/10-AOAS394.

De Waegenaere, A., Melenberg, B., and Stevens, R. (2010). Longevity risk. *De Economist* 158: 151–192. doi:10.1007/s10645-010-9143-4.

ECMWF (2024). Heatwaves – A brief introduction [electronic resource]. https://climate.copernicus.eu/heatwaves-brief-introduction.

Elizalde, B. and Díaz, V. (2016). Aging in rural areas of Spain: The influence of demography on care strategies. *The History of the Family* 21(2): 214–230. doi:10.1080/1081602X.2016.1157828.

ESPON (2017). *Shrinking rural regions in Europe*. Luxembourg: ESPON EGTC.

George, M.V., Smith, S.K., Swanson, D.A., and Tayman, J. (2001). Population projections. In: Siegel, J. and Swanson, D.A. (eds.). *The methods and materials of demography*. San Diego: Elsevier Academic Press: 561–601.

Human Mortality Database (2024). University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org and www.humanmortality.de (accessed on 01/10/24).

Hunt, A. and Blake, D. (2020). On the structure and classification of mortality. *North American Actuarial Journal* 25(sup1): S215–S234. doi:10.1080/10920277.2019.1649156.

Külekci, B. and Selcuk-Kestel, A. (2021). Assessment of longevity risk: Credibility approach. *Journal of Applied Statistics* 48(13–15): 2695–2713. doi:10.1080/02664763.2021.1922613.

Lee, R. and Carter, L. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association* 87(419): 659–671. doi:10.1080/01621459.1992.10475265.

Li, N., Lee, R., and Gerland, P. (2013). Extending the Lee–Carter method to model the rotation of age patterns of mortality decline for long-term projections. *Demography* 50(6): 2037–2051. doi:10.1007/s13524-013-0232-2.

Lledó, J. and Pavía, J.M. (2024). Integrating wealth-indexes into life insurance and financial products [unpublished manuscript]. Valencia: Universitat de València.

Lledó, J., Pavía, J.M., and Sánchez, J. (2023). An alternative approach to manage mortality catastrophe risks under Solvency II. *Risk Management* 25: 16. doi:10.1057/s41283-023-00120-6.

Lledó, J., Pavía, J.M., and Simó-Noguera, C. (2024). Anomalous distributions of birthdates across days of the month. An analysis using Spanish statistical records. *Population Studies* forthcoming. doi:10.1080/00324728.2024.2393622.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology* 58(8): 1246–1266. doi:10.2113/gsecongeo.58.8.1246.

Miller, T. (2001). Increasing longevity and medicare expenditures. *Demography* 38(2): 215–226. doi:10.1353/dem.2001.0018.

Pavía, J.M. and Lledó, J. (2022). Estimation of the combined effects of ageing and seasonality on mortality risk: An application to Spain. *Journal of the Royal Statistical Society Series A: Statistics in Society* 185(2): 471–497. doi:10.1111/rssa.12769.

Pavía, J.M. and Lledó, J. (forthcoming). qlifetable: An R package for constructing quarterly life tables. *PLoS One* forthcoming. doi:10.1371/journal.pone.0315937.

Peel, M.C., Finlayson, B.L., and McMahon, T.A. (2007). Updated world map of the Köppen–Geiger climate classification. *Hydrology and Earth System Sciences* 11(5): 1633–1644. doi:10.5194/hess-11-1633-2007.

Plat, R. (2009). On stochastic mortality modeling. *Insurance: Mathematics and Economics* 45(3): 393–404. doi:10.1016/j.insmatheco.2009.08.006.

Preston, S.H. (1977). Mortality trends. *Sociology* 3(1): 163–178. doi:10.1146/annurev.so.03.080177.001115.

R Core Team (2024). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Reinsch, C.H. (1967). Smoothing by spline functions. *Numerische Mathematik* 10(3): 177–183. doi:10.1007/BF02162161.

Renshaw, A. and Haberman, S. (2006). A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics* 38(3): 556–570. doi:10.1016/j.insmatheco.2005.12.001.

Royé, D., Sera, F., Tobías, A., Lowe, R., Gasparrini, A., Pascal, M., de'Donato, F., Nunes, B., and Teixeira, J.P. (2021). Effects of hot nights on mortality in Southern Europe. *Epidemiology* 32(4): 487–498. doi:10.1097/EDE.0000000000001359.

United Nations, Department of Economic and Social Affairs, Population Division (2024). *World population prospects 2024: Methodology of the United Nations population estimated and projections*. UN DESA/POP/2024/DC/NO. 10, July 2024 [Advance unedited version]. New York: UN.

Van Berkum, F., Antonio, K., and Vellekoop, M. (2016). The impact of multiple structural changes on mortality predictions. *Scandinavian Actuarial Journal* 2016(7): 581–603. doi:10.1080/03461238.2014.987807.

Villegas, A., Millossovich, P., and Kaishev, V. (2018). StMoMo: An R package for stochastic mortality. *Journal of Statistical Software* 84(3): 1–38. doi:10.18637/jss.v084.i03.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. https://ggplot2.tidyverse.org. doi:10.1007/978-3-319-24277-4.

Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. (2023). *dplyr: A grammar of data manipulation*. https://dplyr.tidyverse.org.

Wilmoth, J.R., Andreev, K.F., Jdavov, D.A., Glei, D.A., and Riffe, T. (2021). *Methods protocol for the human mortality database*. Human Mortality Database. University of California Berkeley and Max Planck Institute of Demographic Research [online]. Available at https://mortality.org/.