Multipopulation Mortality Analisys: bring out the unobservable with Latent Cluster

Ana Maria Debon Auceio[1], Steven Haberman[2], and Gabriella Piscopo[3]

[1]Facultad de Administracion y Direccion de Empresas,Universitat Politecnica de Valencia, ana.maria.debon@ext.uv.es
[2] Bayes Business School, Faculty of Actuarial Science and Insurance, City University of London, s.haberman@city.ac.uk
[3]Department of Economic and Statistical Science, University of Naples Federico II, 80138 Naples, gabriella.piscopo@unina.it

August 28, 2023

Abstract

Mortality patterns experienced in closely related populations show similarities in some aspects and differences in others. Indeed, if a decline in mortality rates among low-mortality countries is observed, it is possible that populations experience different trends through which this decline occurs. Observing mortality rates for ages and over specific time windows, it is evident that the different interactions between the variables age and time influence longevity trends. Therefore, to grasp the complexity of the phenomenon, the similarities or differences in mortality need to be analyzed by considering three dimensions: age, year, and country, simultaneously. With this aim in mind, we propose applying a multidimensional latent clustering approach to multipopulation mortality data in this paper. We investigate some similarities between the mortality experience of different countries, searching for latent structure across these groups. Starting from the observation units represented by single countries, we nest them in higher-level units of clusters. We apply the proposed model to the mortality rates of 20 developed countries using data from 1965 to 2019 from the Human Mortality Database. We present detailed results for the lower mortality cluster, which collects ages from 50 to 60 among all countries of the selected dataset and highlights different mortality trends between the countries.

Keywords: Latent Clustering, Multipopulation mortality data.

1 Introduction

In recent years, the demographic literature on multipopulation modelling of mortality rates that identifies clusters of countries showing similar mortality trends has blossomed. Some

of these models are designed starting from synthetic summary measures of longevity, such as life expectancy [Bohk-Ewald et al., 2017, Amin and Steinmetz, 2019, Levantesi et al., 2022]) or life span inequality [Edwards and Tuljapurkar, 2005, Vaupel et al., 2011, Debón et al., 2017]. Other contributions focus on clustering models of specific components of mortality, such as the age-specific death rates [Léger and Mazzuco, 2021], or the mortality improvement rates [Djeundje et al., 2022]. Piscopo and Resta [2014] introduce the use of Self Organizing Maps (SOMs) in multidimensional mortality analysis. Piscopo and Resta [2017] apply spectral biclustering to mortality datasets to capture the period, the age, and the cohort effects. Dong et al. [2020] propose multi-population mortality forecasting using tensor decomposition. Another strand of the literature collects contributions concerning the clustering of populations by specific causes of death [Grigoriev and Pechholdová, 2017, Nigri et al., 2022]. Cardillo et al. [2022] apply a tensor-based approach to the mortality by cause of death.

While the demographic literature has focused on identifying common mortality trends across countries and highlighting differences by looking for economic and social explanations, the actuarial literature has explored models for the fitting and forecasting of mortality rates by exploiting multi-population datasets. Starting from the Lee and Carter [1992] model, which is considered a milestone to forecast mortality for a single country, many extensions in the multipopulation setting have been proposed [Tuljapurkar et al., 2000, D'Amato et al., 2014])to obtain coherent forecasts [Li and Lee, 2005, Hatzopoulos and Haberman, 2013, D'Amato et al., 2019, Wu and Wang, 2019]. Thus, Lam and Wang [2022] model joint mortality and forecast multiple subpopulations through multivariate functional principal component analysis techniques. Tsai and Cheng [2021] implement statistical clustering methods into mortality models in order to improve their forecasting performance. Hatzopoulos and Haberman [2013] propose a fuzzy c-means cluster analysis based on the main time trends to produce coherent mortality forecasts.

In the papers cited so far, the clustering of countries is made on the basis of mortality trends, which take into account the evolution of mortality over time considering the influence that all ages, or subsets of them, have on the chosen measures of mortality, allowing for how they react overall to improvements in mortality. In this work, however, we propose a different approach to mortality clustering and offer complementary information to that provided by the existing literature. We propose a more specific level of clustering, trying to identify groups of countries that show similarities in mortality at the same ages and in the same years. In fact, it may happen that some countries, despite showing similar mortality trends, differ in the behavior of individual ages in specific years. In order to bring out these not directly observable characteristics, we introduce Latent Class Clustering (LCC) techniques into the analysis of multipopulation mortality trends.

LCC is a mixture model for classifying individuals based on their responses to multiple items. When existing subgroups in the data represent different populations, it is possible to analyse the latent class structure across these groups. In multiple-group LCC models, observation units (in our case, a single country) are nested within a higher-level unit (in our case, the cluster of countries). The belief at the basis of LCC methodologies is that there are one or more latent variables that explain the clustering. In the multipopulation mortality setting, we observe groups (each country is a group of individuals with different ages in different years). However, there could be some latent variables that we do not observe directly (for example, socio-economic or political factors) that explain the coun-

tries' clustering. The LCC permits the identification of groups of countries that show similar mortality for the same ages and in the same years, providing a three-dimensional grouping: by age, year, and country. LCC results provide different information compared to two-dimensional clustering: while with the latter approach, the groups of countries show the same general trend in mortality over some time, the former approach captures similar characteristics between countries that have the same mortality at certain ages and in specific years.
In recent years, LCC techniques have been used in the epidemiological and clinical literature [Larsen et al., 2017, Santaolalla et al., 2019, Li et al., 2020, Luo et al., 2021] and in the demographic literature [Larsen et al., 2017] to study mortality in relation to specific causes of death, but its exploitation in the context of multipopulation models of mortality trends appears to be new. The application of latent clustering in the medical sciences has involved patients with a specific disease, with patients being categorized according to specific and observable risk factors. Despite this, within each group of patients, the course of the disease for some has been different than for others and the reason has been attributed to the presence of latent factors that are not directly observable - this has allowed the further clustering of the patients in order to explain the differences in the reaction to the treatments. In this paper, we propose to adopt the same concept in the context of the analysis of multipopulation mortality: we examine the mortality rates in different population groups corresponding to different countries and classified also by age and time variables. However, there are some specific features in the data that can only be captured by nested groups and nesting could be explained by the presence of latent variables not directly observable or not available to the researcher. This might involve, for example, economic and social factors, or a more or less austere political environment.
The remainder of this paper is organized as follows. Section 2 is devoted to describing the basic concepts of the LCC method. In Section 3 the technique previously described is applied in the specific context of a multi-population analysis. Section 4 presents an empirical application of LCC to mortality rates of a group of 20 developed countries from the Human Mortality Database. Finally, concluding remarks are offered in Section 5.

2 Methodology: LCC

LCC is a statistical methodology for capturing similarities among observable data when other unobservable categorical variables explain the segmentation of the dataset into several latent classes or clusters. Therefore, LCC permits identifying latent subpopulations within a dataset and analysing the responses to the observed variables, which are called *indicators*. Data are attributed to the same cluster if they show similar patterns of variations to the available indicators; these similarities are measured in scores calculated through probability distributions, whose unknown parameters have to be estimated. In this sense, LCC is used to capture latent heterogeneity in samples; for a detailed discussion, see Hagenaars and McCutcheon [2002]. Unlike classical cluster analysis techniques, LCC is a model-based clustering approach, in the sense that the analysis is based on the hypothesis of the existence of a mixture of underlying probability distributions from which the data are generated. The parameter estimation problem of these distributions is solved through the maximum likelihood method. It follows that LCC is a probabilistic clustering approach: each object is attributed to one cluster with a certain degree of uncertainty.

This aspect makes LCC similar to fuzzy clustering, but in the latter case, the grades of membership attributed to each object have to be estimated while, in the former case, each grade is calculated after having estimated the model parameters. Consequently, unlike fuzzy clustering, LCC allows the classification of other data belonging to the population not present in the original dataset. In the following, we describe the methodological framework for LCC.
Let z_i be all explanatory variables called *predictors* for the single case i of the selected dataset and let y_i be the dependent variable or *response* or *indicator* corresponding to the case i. The model assume that the dependent variable depends not only on the predictors but also on one or more latent variables. Let x be a single latent variable with K categories, called *Latent Classes or Clusters*. The LCC defines a general mixed probability structure that describes the relationship between predictor, latent, and response variables as follows:

$$f(\mathbf{y_i}|\mathbf{z_i}) = \sum_{x=1}^{K} P(x|\mathbf{z_i}) f(\mathbf{y_i}|x, \mathbf{z_i}) \tag{1}$$

that can be rewritten as

$$f(\mathbf{y_i}|\Theta) = \sum_{x=1}^{K} \pi_x f_x(\mathbf{y_i}|\Theta_{x,\mathbf{z_i}}) \tag{2}$$

where $\pi_x = P(x|\mathbf{z_i})$ is the prior probability to belong to cluster x given the observed explanatory variables, Θ is the set of parameters of the explanatory variables, $\Theta_{x,\mathbf{z_i}}$ is that of the mixture densities of $\mathbf{y_i}$ given x and $\mathbf{z_i}$.
The assumption behind the LCC framework is to describe a model for $f(\mathbf{y_i}|\mathbf{z_i})$. According to Eq.(2), the latent variable may be influenced by the explanatory variables, and the indicators may be influenced by both latent and observed variables.
A particular distribution is chosen depending on the scale and the type of indicators. When the variables are categorical, a multinomial distribution is assumed; when the variables are continuous, the multivariate normal distribution is selected; and when the variables are discrete, Poisson or binomial distributions are considered. The distribution of latent variable x given the observed variables is assumed to come from a joint multinomial distribution. We refer to Vermunt and Magidson [2016b] for a detailed description of the formulations of all joint distributions and their parametrization according to a classic GLM model. The wide choice of models for the joint distributions of indicators, latent and observable variables allows for the management of variables of different types and scales in a multidimensional space without having to proceed with data scaling or without having to abandon dealing at the same time with continuous variables or discrete ones, ordinal or cardinal [Vermunt and Magidson, 2016a]. This wide choice undoubtedly represents one of the great advantages of LCC compared to other cluster methodologies.

Just to give an example, when the indicators are modelled with a Normal distribution within latent classes with parameters μ_x and Σ_x, the clustering algorithm estimates separately a set of parameters for each latent class, also allowing us to identify classes that differ with respect to their means or variances so that the clusters might be homogeneous with respect to the responses to the explanatory variables.
An extension of Eq.(2) is used when the indicators are continuous, nominal and ordinal

variables with different scales:

$$f(\mathbf{y_i}|\Theta) = \sum_{x=1}^{K} \pi_x \prod_{j=1}^{J} f_x(y_{ij}|\Theta_{jx}) \tag{3}$$

where j is the single indicator and J the total number of indicators.
Once the whole probability structure has been defined, the parameters must be estimated. The two main methods used are the Maximum Likelihood (ML) and the Maximum Posterior Method (MAP). The MAP is used to estimate the parameters and maximizes the log-posterior distribution given by the sum of the log-likelihood function and the logs of the prior distributions. For more details, please refer to Vermunt and Magidson [2016a]. Finally, once the parameters have been estimated, each object has to be allocated to the cluster with the higher posterior class membership probability:

$$\pi_{x|\mathbf{y_i}} = \frac{\pi_x \prod_{j=1}^{J} f_x(y_{ij}|\Theta_{jx})}{\sum_{x=1}^{K} \pi_x \prod_{j=1}^{J} f_x(y_{ij}|\Theta_{jx})} \tag{4}$$

The most widely used classification method is the modal allocation, according to which each data is assigned to the class with the highest scores, i.e., the highest posterior probability. One of the advantages of LCC models is the ability to obtain an equation, called a scoring equation, for calculating these posterior membership probabilities directly from the observed variables.
The advantages of LCC over other clustering algorithms are many. Among these, one of the main advantages comes from the fact that it is a statistical model-based clustering model. Indeed if, on the one hand, the statistical hypotheses on mixed probability distributions linking the observable variables to the dependent and latent variables make the analyst's choice subjective, on the other hand, they make it possible to identify objective statistical criteria for determining the number of clusters or for the segmentation of data among the various groups. Further, LCC permits the fitting of probabilistic models to the data, in contrast to distance-based clustering methods, like the k-means, which segments observations based on a dissimilarity criterion. Moreover, the flexibility of the LCC permits incorporation in the analysis of many real-world circumstances, like unequal covariance matrices between clusters, unequal numbers of observations in clusters, or correlation between variables inside clusters. Regarding the number of selected clusters, although a sphere of subjectivity is present in all clustering methods, for model-based approaches, quantitative criteria like Bayesian Information Criterion (BIC) or the Integrated Completed Likelihood (ICL) criterion are available. LCC, being a probabilistic model, allows for statistical procedures for determining the number of clusters and provides results which are stated in terms of probabilities and are more interpretable. For a more detailed comparison of clustering methods, we refer to Xu [2013] and Eshghi et al. [2022].

Nowadays, the availability of software packages to implement LCC makes the tool attractive and easy to use [Haughton et al., 2009], so that its applications in various research fields have proliferated, ranging from medicine to economics and the social sciences [Kaplan, 2004].

3 The LCC of a mortality dataset

The use of LCC appears appropriate in the context of mortality for the following reasons that have guided our choice. First, the phenomenon of longevity trends is known to be very complex, and mortality is known to be linked to multiple demographic and socio-economic factors that are not always directly observable. Datasets in which mortality rates are linked to different causes of death are not always available and comparable. Awareness that there are latent variables in addition to observable ones such as age or time that can influence trends in mortality can help improve our understanding of the phenomenon. Moreover, in the mortality dataset, the analyst has to deal with different types of variables, discrete variables such as age and year, continuous variables like the mortality rate, and often also qualitative variables, like gender and country. As explained in Section 2, LCC can deal with different types of data and create clusters working on combinations of categorical and numeric data.
In contrast, most cluster algorithms can only deal with numeric variables. Another advantage over the other algorithms is that LCC does not require any data scaling procedure, and mortality datasets often present data with different measurement scales. Furthermore, in the context of mortality, LCC is flexible and allows us to incorporate different phenomena (e.g., predictor variables, covariates, direct or indirect effect inside clusters, unequal covariance matrices...) depending on the objectives of the analysis; these are elements which are not readily addressed with the other cluster algorithms. In the following, we put the model described in Section 2 in the context of multi-population mortality data.
Let $m^i_{x,t}$ be the central mortality rate for an individual aged a at time t belonging to country c with

$$a = \{a_1, a_2, ..., a_\omega\}$$
$$t = \{t_1, t_2, ..., t_T\}$$
$$c = \{c_1, c_2, ..., c_N\}.$$

In our case, we have one indicator m and three explanatory variables, of which age and time are ordinal, and mortality rates are continuous, so we have to deal with a mixed model with variables having different scales. In this model

$$\mathbf{z_i} = \{a, t, c\}.$$

The mortality rates are grouped in N population. With the LCC model, we look for a latent structure among these groups and suppose there is a single latent variable x .
Starting from the observation units represented by a single country, we nest them in higher-level units of clusters, considering both the variables age and time simultaneously. The LCC is implemented through a two-step procedure. In the first step, we estimate the parameters of the mixed probability structures through the MAP as described in Section 2:

$$f(\mathbf{m_i}|\mathbf{z_i}) = \sum_{x=1}^{K} \pi_x f_x(\mathbf{m_i}|\Theta_{x,\mathbf{z_i}}) \tag{5}$$

In the second step, we allocate each object to the cluster with the higher posterior class membership probability:

$$\pi_{x|\mathbf{m_i}} = \frac{\pi_x \prod_{j=1}^{J} f_x(m_{ij}|\Theta_{jx})}{\sum_{x=1}^{K} \pi_x \prod_{j=1}^{J} f_x(m_{ij}|\Theta_{jx})} \tag{6}$$

where j is the single indicator.

4 Numerical Application

In this section, we show the implementation of the LCC procedure to a mortality dataset, the details of which are described in Section 4.1. The clustering object variable is the mortality rate, the indicator variables are age and year, the grouping categorical variable is country. Due to the specificity of the clustering algorithm, as noted above, no preliminary data scaling is required (Vermunt and Magidson [2002]).

4.1 Data

In this paper, we implement the two-step cluster algorithm, as described in the previous section, on mortality data available from the Human Mortality Database [2022]. We focus the analysis on recent mortality trends, selecting data from 1965 to 2019. Having chosen the time interval of the analysis, we exclude Germany from consideration because data for the combined country are available only starting from 1990. Greece, which presents data from 1971, is included because the number of missing years compared to other countries is relatively low. Twenty countries with large populations and developed economies are identified in this analysis. Moreover, we select mortality rates for single ages from 50 to 95 and by individual calendar year for the male population, noting that data at more advanced ages are of reduced quality as the cases are less numerous and errors in the ages recorded may be present. The complete dataset is listed in Table 1 and consists of 198168 entries.

4.2 Results

The LCC algorithm described in Section 3 is implemented with the software XLSTAT provided by Addinsoft [2022]. The scoring Eq. 5 and the posterior probabilities in Eq. 6 are calculated under a multinomial logit model. Parameter estimation is carried out through a two-step iterative algorithm, which begins using the Expectation Maximization (EM) algorithm until either the maximum number of EM iterations ($EMIt$) or the EM convergence tolerance criterion (EMT) is reached. Then, a Newton Raphson (NR) algorithm is used until the maximum number of NR iterations ($NRIt$), or the overall convergence tolerance criterion (NRT) is reached. For more details see Vermunt and Magidson [2016a].
We set $EMIt = 250$, $NRIt = 50$, $EMT = NRT = 0.01$.
The first step of any clustering procedure is choosing the number of clusters. Although the literature proposes many identification and classification criteria, a subjective evaluation

Country	**HMD label**
1 Australia	AUS
2 Austria	AUT
3 Belgium	BEL
4 Canada	CAN
5 Switzerland	CHE
6 Denmark	DNK
7 Spain	ESP
8 Finland	FIN
9 France	FRANCNP
10 United Kingdom	GBRNP
11 Grece	GRC
12 Ireland	IRL
13 Italy	ITA
14 Japan	JPN
15 Netherlands	NLD
16 Norway	NOR
17 Portugal	PRT
18 Sweden	SWE
19 Taiwan	TWN
20 United States of America	USA

Table 1: Selected countries in the dataset

N.Cl.	**LL**	**BIC**	**AIC**	**Class.Error**	**Entropy R^2**	**ICL-BIC**
2	-291474.46	584094.83	583160.91	0,018	0.939	588302.88
3	-271367.17	543934.31	542956.34	0.034	0.927	551836.96
4	-259950.04	521154.10	520132.07	0.047	0.918	532411.61
5	-252910.16	507128.39	506183.32	0.062	0.908	521796.82
6	-248216.91	497795.96	496811.82	0.076	0.898	515834.28
7	-245031.72	491479.62	490456.43	0.091	0.887	513095.26
8	-242951.89	487374.04	486311.79	0.100	0.883	511209.39
9	-240714.34	482952.98	481851.68	0.101	0.888	507095.96
10	-238517.08	478612.52	477472.17	0.101	0.892	503052.71

Table 2: The Information and Classification criteria

is likely to affect the final choice. Therefore, "clustering is in the eye of the beholder" (Estivill-Castro [2002]), and the analyst must evaluate the interpretability of the clusters and the population's size. Table 2 shows the value of three information statistics (the Log-Likelihood (LL), the BIC, and the AIC) and three classification statistics (the Classification Error, the Entropy R^2 and the ICL-BIC) for LCC models with the number of clusters $k = 1, 2, \ldots, 10$. As the number of clusters increases, the identification and classification statistic values improve, except for the classification error, which, as expected, is smallest in the case of a single cluster.
We set $k = 10$: this is a high enough number of clusters to differentiate the numerous data points being worked on and is a value at which further increments do not lead to significant increases in the values of the baseline statistics. Figure 1 shows the clustered data under the three variables age, time, and country. Figures 2 and 3 represent the same clustering in bidimensional Trellis plots.

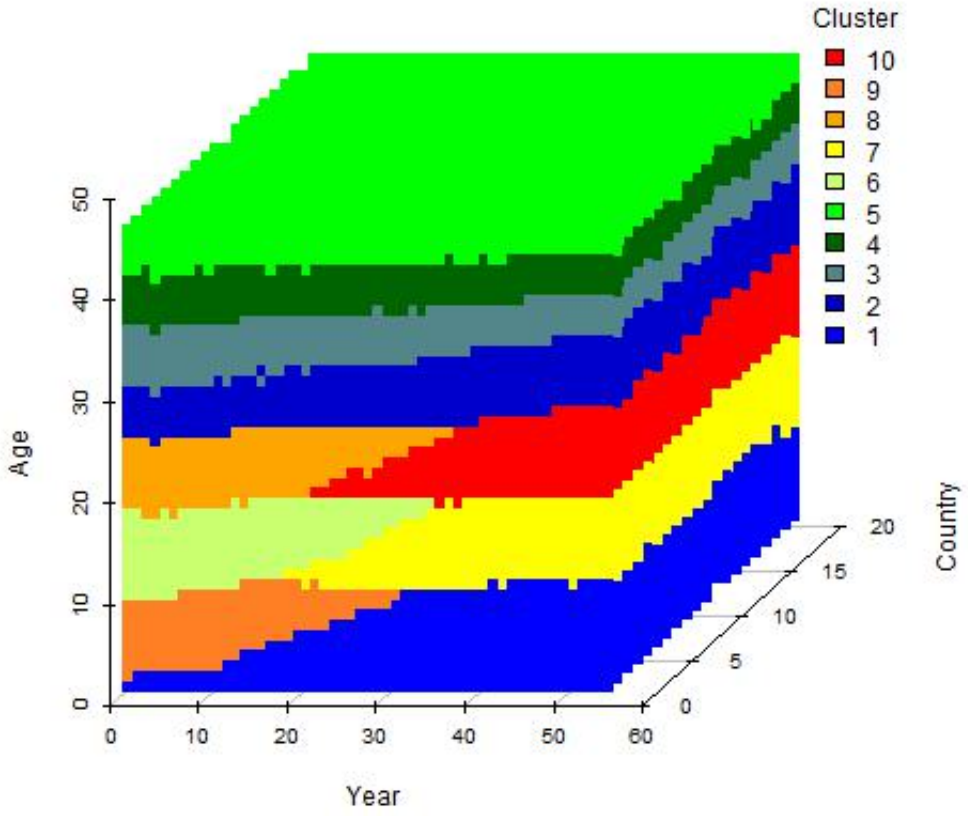

Figure 1: The Clusterized Mortality Rates

Trellis plots represent the interaction of each pair of variables in multivariate data. Two variables are chosen to represent a common set of axes (repeated in each panel), against which all combinations of two categorical variables called "conditioning variables" are plotted. In Figure 2, on the y-axis, there are the ages; on the x-axis the countries and, from one plot to another, the years vary. Figure 2 shows how different age groups for all countries are classified into different clusters from one plot to another as the years vary. For example, the first plot at the top left represents the group of individuals aged between 50 and 60 during the first 9 years of the dataset. The color blue represents the first cluster. As highlighted in the next section 4.3, for Finland and USA, during this period, the individuals of the group do not fall into the low mortality cluster but they will enter it only later; for this reason, the bar corresponding to these two counties does not show the color blue, but the orange one. In Figure 3, on the y-axis, there are the years; on the x-axis, the countries, and, from one plot to another, the ages vary. The first plot on the top left represents the ages between 50 and 59: this group of individuals falls into the first cluster (the blue one) for all years and almost all countries. Here, two orange bars are highlighted, correspodning to Finland and USA, indicating that this group of individuals belongs to a different cluster for the first few years. As highlighted in Subsection 4.3, the

50-59 group in these two countries falls into the cluster with low mortality with a time delay.

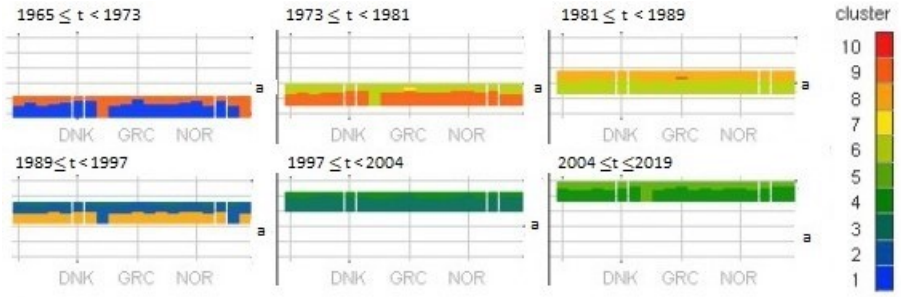

Figure 2: Trellis Plot by Age and Countries

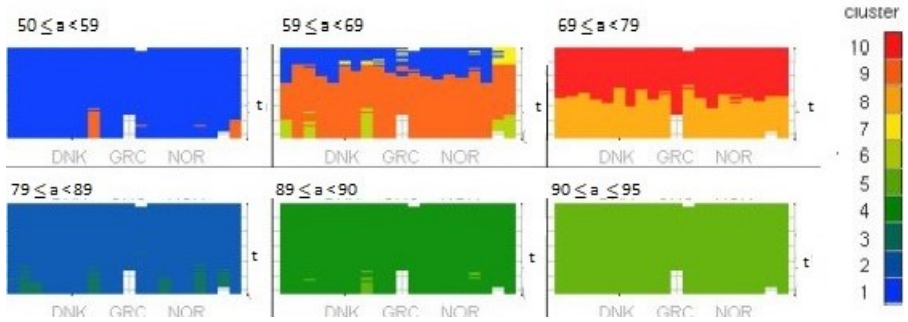

Figure 3: Trellis Plot by Year and Countries

4.3 Discussion

In this section, we focus the discussion on cluster 1, which is the cluster with the lowest mortality experience, and so it collects ages between 50 and 60. The cluster composition is detailed in Table 3 for each individual value of the country, age, and year variables for which the mortality rates fall within cluster 1. As can be seen from the results, going from 1965 to 2019, the cluster collects progressively more advanced ages reflecting the general progressive improvement of living conditions: while in the first years of the dataset, the individuals with the lowest mortality are aged 50-51. Then, as the years pass, more advanced ages are included in the cluster, so that, in the most recent years, individuals aged between 50 and 60 are collected. Almost all counties are present in this cluster, with individuals aged between 50 and 60 and years from 1965 to 2019, but with some differences, reflecting different mortality trends between countries. In some countries, older people enter the low mortality cluster earlier, for others they enter later. In the following paragraphs, we describe some relevant pieces of evidence.

In most countries, we observe that in the years between 1965 and the early 1970s, the ages falling into the cluster with the lowest mortality are between 50 and 52, while only in the mid-1980s are ages up to 55 also included. More or less, starting from the 2000s, we are witnessing a forward shift of the ages present in the cluster, up to including individuals aged 58-60. Compared to this general trend, we highlight some essential differences recorded between the two North American countries in the dataset, the USA, and Canada,

in line with the findings found in the demographic literature for individual countries.
Although over the last century, life expectancy has improved in most high-income countries, in recent decades, the United States and Denmark have experienced life expectancy stagnation (Christensen et al. [2010],Bureau [2011]). As can be seen from Table 3, as far as Denmark is concerned, until the end of the 1980s, only individuals aged between 50 and 53 were included in the low mortality cluster, with an evident lag compared to other countries, and individuals aged 55 were included in the cluster 10 years later. The phenomenon is even more pronounced in the USA, where individuals aged 50 fall into the cluster with a six years' delay (only starting from 1975) and those aged 55 only in the 90s. However, starting in the 2000s, while increases in life expectancy in Denmark have resumed, the same phenomenon has not occurred in the USA. In particular, from 1995 and then more strongly after the 2000s, Denmark has experienced high rates of mortality improvement for all ages [Andreev, 2002, Jarner et al., 2008]. The stagnation and the subsequent recovery may affect mortality projections in Denmark: Djeundje et al. [2022] have calibrated mortality improvement rates for the period 1965-2010 and then forecasted them for 2011-2017 and demonstrated that the observed mortality improvements for Denmark are higher than projected and by a greater margin than in other countries; a possible explanation being that the period of stagnation affected the calibration phase.
Another emblematic case among the Scandinavian countries is that of Finland, which despite its wealth, is characterized by high suicide rates [Holopainen et al., 2014, Partonen et al., 2022, Statistics Finland, 2021]). As can be seen from Table 3, also for Finland, individuals are included in the low mortality cluster with an evident time lag compared to other countries.
As far as the USA is concerned, the literature is full of contributions that highlight the mortality gap compared to other high-income countries [Berkman et al., 2011, Wilmoth et al., 2011] and identify some of the causes (Nigri et al. [2022]), including the poor efficiency of the health system and lifestyles and obesity linked to socio-economic differences [Preston and Ho, 2009]. In contrast, the opposite situation occurs in neighboring Canada and emerges strongly [Milligan and Schirle, 2021]. Compared to the USA, individuals in Canada of all ages in the range 50-59 fall into a low mortality cluster 10 years earlier, and Canada is the only country in the dataset that has individuals aged up to 59 in the cluster for the last 20 years, placing it among the first countries for favourable longevity trends. Figure 4 shows the low mortality clustering for USA and Canada.
At the end of this section, it must be emphasized that any reference to possible explanatory causes of mortality (like the suicide rate in some Scandinavian countries and obesity in the USA) has been made exclusively to show how the results obtained from the application of the LCC are consistent with those of the cited literature. However, analysis of individual causes of death is beyond the scope of this work due to the essence of latent clustering: suggesting that there are factors that can explain mortality but are latent variables unavailable to the analyst.

Cluster 1					
AUS		**AUT**		**BEL**	
Age	**Year**	**Age**	**Year**	**Age**	**Year**
50-51	1965-75	50	1968,71	50	1965-67
50-52	1975-77	50-51	1967,70,72,75-76,78	50-51	1966,68-73

50-53	1978-80	50-52	1965-66,69,73-74,77,79-82,84	50-52	1974-77,1979
50-54	1981-83	50-53	1983,85-87,89	50-53	1978,80-83
50-55	1984-87	50-54	1988	50-54	1984-86,88
50-56	1988-90	50-55	1990-94	50-55	1987-89
50-57	1991-94	50-56	1995-96,99	50-56	1990-94,97-99
50-58	1995	50-57	1997-98,2000-2003	50-57	1995-96,2000,02-03
50-59	1996-2005,2007-15	50-58	2004-08	50-58	2001,2004-2008
50-60	2006,2008-14,2016-19	50-59	2009-19	50-59	2009-19
CAN		**CHE**		**DNK**	
Age	**Year**	**Age**	**Year**	**Age**	**Year**
50-51	1965-68,70,72-73	50-52	1965-70	50-52	1967,72,75-76
50-52	1969,71,74-78	50-53	1971-72,75-76	50-53	65-66,68-71,73-74,77-82,84-88
50-53	1979-80	50-54	1973-74,1977-82	50-54	1983-89-91
50-54	1981-83	50-55	1983-1988	50-55	1992-93
50-55	1984-88	50-56	1989,91-92	50-56	1994-97,99
50-56	1989-90,92	50-57	1990-93	50-57	1998,2000-03
50-57	1991,93-94	50-58	1994-95	50-58	2004-07,09-10
50-58	1995-99	50-59	1997-2006,10,12	50-59	2008,12,19
50-59	2000-19	50-60	2007-19		
ESP		**FIN**		**FRATNP**	
Age	**Year**	**Age**	**Year**	**Age**	**Year**
50-52	1966-70	50	1976,78	50	1965-69,71
50-53	1965,71-76	50-51	1979,82	50-51	1970-81
50-54	1977-81	50-52	1980,83-85	50-52	1982-86
50-55	1982-86,88,89	50-53	1986-88-89	50-53	1987-88
50-56	1987,90-92	50-54	1987,90-91	50-54	1989-90
50-57	1993-97	50-55	1993,95-96	50-55	1991-94
50-58	1998-2003	50-56	1992,94,98,99,01	50-56	1995-98
50-59	2004-2019	50-57	1997,02,03-05,07	50-57	1997-03
		50-58	2000,06,09-11	50-58	2004-07,09,11-12
		50-59	2012-2019	50-59	2008,10,13-19
GBRNP		**GR**		**IRL**	
Age	**Year**	**Age**	**Year**	**Age**	**Year**
50-51	1965-69	50-55	1981,83	50-51	65-67,69,73-74,76,78,80
50-52	1970-79	50-56	1982,84-85,87,88	50-52	68,70-72,73-74,76,78,80
50-53	1980-82	50-57	1986,89-94	50-53	1975,83-84,86,87
50-54	1983-87	50-58	1995-00,04,05,07,12,15,18	50-54	1985
50-55	1988-90	50,59	2001-03,08-11,13-14,16,17,19	50-55	1988,90,91
50-56	1991-94			50-56	1989-92,95,98
50-57	1995-98			50-57	1996-97,99-00
50-58	1999-02			50-58	2001-02,04
50-59	2003-19			50-59	2003,05-14,16
				50-60	2015,17
ITA		**JPN**		**NLD**	
Age	**Year**	**Age**	**Year**	**Age**	**Year**
50-51	1965,69	50-51	1965	50-52	1965-66,69-72
50-52	1966-68,70-78	50-52	1966-69	50-53	1967-68,73,75-76
50-53	1979-83	50-53	1970-72	50-54	1974,77-80,82-83
50-54	1984-86	50-54	1973-76	50-55	1981,85-87
50-55	1987-89	50-55	1977-81,83	50-56	1988-93
50-56	1990-93	50-56	1982,84-89	50-57	1994-97
50-57	1994-96	50-57	1990-94	50-58	1998-01
50-58	1997-99	50-58	1995-97	50-59	2002-18
50-59	2000-13,15,17,18	50-59	1998-14,16,18	50-60	2019
50-60	2014,16,19	50-60	2015,17,19		

NOR		PRT		SWE	
Age	**Year**	**Age**	**Year**	**Age**	**Year**
50-53	1965-75,81	50-51	1965-71,73-77	50-53	1965,67
50-54	1976-80,82-83,87	50-52	1972,78-79,81	50-54	1966,68-79
50-55	1984-85	50-53	1980,82-85,87-88	50-55	1980-82,84-85
50-56	1986,88-92,95	50-54	1986,89,91-93	50-56	1983,88-88,91
50-57	1993	50-55	1990,94	50-57	1989-90,92
50-58	1994,96-98	50-56	1995-99	50-58	1993-95
50-59	1999-2009,11-14	50-57	2000-02	50-59	1996-06,08,10,12
50-60	2010,15-19	50-58	2003,05,06,17	50-60	2007,09,11,13-19
		50-59	2004,07-16,18-19		
TWN		**USA**			
Age	**Year**	**Age**	**Year**		
50-51	1970-73	50	1971-73		
50-52	1974-77	50-51	1974-77		
50-53	1978-84	50-52	1978-81		
50-54	1985-94,96	50-53	1982-85		
50-55	1995,97-98,2000	50-54	1986-90		
50-56	1999,01-02,04-05	50-55	1991-95		
50-57	2003,06-08	50-56	1996-2000		
50-58	2009-2013,15,16,18,19	50-57	2001-06		
50-59	2014,17	50-58	2007-19		

Table 3: Ages, Years and Countries in cluster 1

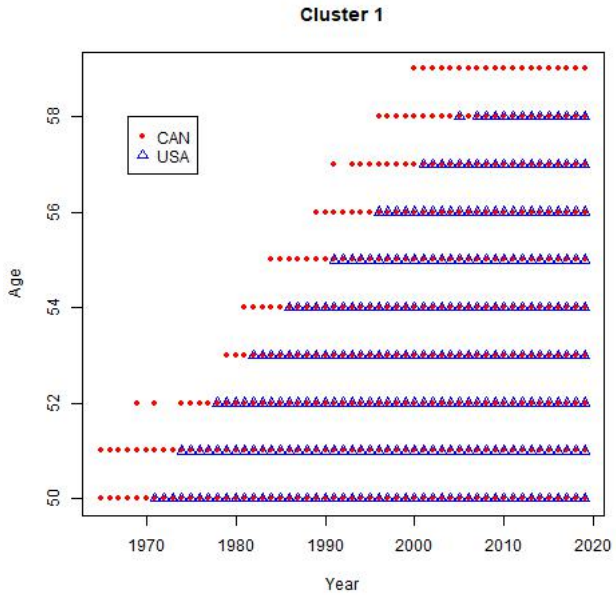

Figure 4: The Low Mortality Cluster in USA and Canada

The ability of the methodology used to discriminate between ages and countries over the years can also be highlighted for other age groups beyond that of the main focus of the work. The plot in the upper centre of Figure 3 relates to the second age group that welcomes individuals between the ages of 60 and 69. In this plot, we still find some parts in blue that relate to sixty-year-olds entering the cluster in recent years with low

mortality. However, the blue bars in the plot are not all of the same length, demonstrating the fact that there are differences between countries. For example, corresponding to the first country in the dataset, Austria, the blue bar is the longest because age 60 enters the low mortality cluster first (as shown in Table 3 above). In line with the dataset's first years, other colors emerge corresponding to other clusters for which higher mortality levels are recorded for this same age group. The plot at the top right relates to the 70-79 age group. In this plot, there are two clusters, the orange one and the red one; the former has a relatively higher mortality level than the latter. Also, in this case, it is highlighted how, over the years, individuals in this age group have experienced a reduction in mortality, passing from the orange cluster to the red one, but in different years in the various countries. However, Figure 3 shows that the discriminating capacity of the methodology applied to the dataset considered is higher for the younger age groups and lower for the more advanced ones. In particular, the bottom and central right plots relating to the older age groups are mostly characterized by a single color; thus, individuals, for all the countries considered and mostly in all years, fall into the same cluster. This evidence shows that the proposed clustering methodology applied to the dataset considered is useful in clustering mainly middle-aged individuals rather than particularly elderly ones. Thus, it could serve as a complementary tool to other clustering methods presented in the literature and also be useful in the case of ad hoc analyses, like, for example, the actuarial applications that aim to design targeted welfare policies for assistance to the population group in the early years of retirement, such as those relating to privatized health and social care.
At the end of this section it should be specified that the first cluster includes individuals with lower mortality does not mean that clusters are ordered in increasing mortality levels. In the LCC, the clusters are sorted according to the proportional class assignment, i.e., to each cluster is given a weight equal to the posterior membership probability of the entire cluster. The fact that the first cluster coincides with the one with the lowest mortality means that this class includes cases with the highest posterior probability as a whole. Regarding the colors of the clusters, as we move from blue to red we are moving from cluster with the highest posterior probability to the one with the lowest probability.

5 Conclusions

In this paper, we have proposed applying the Latent Cluster method to mortality data in a multipopulation setting in order to cluster groups of the populations not only on the basis of observable variables available in mortality databases but also on the basis of latent factors that are not directly observable. The advantage of the proposed model is that it allows a greater degree of depth in the analysis of the differences or similarities of the mortality trends that have occurred in different countries. In fact, through latent clustering, it is possible to cluster data under the triple dimensions of age, time, and country, working simultaneously on variables of a quantitative and qualitative nature and different measurement scales. Compared to the two-dimensional cluster models proposed in the demographic literature, the Latent Clustering method allows us to have a greater depth of detail in the analysis, managing to identify not only groups of countries which have shown similar longevity trends, but countries which have shown, in the same years and for the same ages, similar mortality trends. In this paper, we have focused on

high-income countries and a lower mortality cluster of individuals aged 50-60. Further research could be aimed at studying mortality clusters of more advanced ages.
The scope of the application of this new way of looking at mortality data could be multiple and diverse, and could encompass both governmental interest in the identification of appropriate welfare and assistance policies for the various population sub-groups and private sector interest in the identification and pricing of ad hoc insurance and pension products. Currently, most longevity indicators are summary indicators that summarize all factors affecting longevity. However, more detailed information, such as social indicators that can explain trends in mortality, would play a crucial role in designing appropriate policies for governments to achieve their welfare goals. Unfortunately, their availability can only sometimes be guaranteed. In these cases, an analysis by latent factors can improve the degree of understanding of the phenomena that impact on mortality experience, even though detailed datasets for specific social and economic variables or for specific causes of death are not often available.

6 Declaration

The authors confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome. They confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. They further confirm that the order of authors listed in the manuscript has been approved by all of them.

References

Addinsoft. Xlstat. `https://www.xlstat.com`, 2022.

Raid W Amin and Julia Steinmetz. Spatial clusters of life expectancy and association with cardiovascular disease mortality and cancer mortality in the contiguous United States: 1980-2014. *Geospatial Health*, 14(1), 2019.

K Andreev. Evolution of the Danish population from 1835 to 2000. *Monographs on Population Aging*, 9, 2002.

Lisa Berkman, Ichiro Kawachi, and Anton Kunst. Do Americans have higher mortality than Europeans at all levels of the education distribution?: a comparison of the United States and 14 European countries. *International differences in mortality at older ages: Dimensions and sources*, page 313, 2011.

Christina Bohk-Ewald, Marcus Ebeling, and Roland Rau. Lifespan disparity as an additional indicator for evaluating mortality forecasts. *Demography*, 54(4):1559–1577, 2017.

Population Reference Bureau. Trends in life expectancy in the United States, Denmark, and the Netherlands: Rapid increase, stagnation, and resumption. *Today's research on aging*, August 2011(22):1–5, 2011.

Giovanni Cardillo, Paolo Giordani, Susanna Levantesi, and Andrea Nigri. A tensor-based approach to cause-of-death mortality modeling. *Annals of Operations Research*, 11 2022. doi: 10.1007/s10479-022-05042-2.

Kaare Christensen, Michael Davidsen, Knud Juel, Laust Hvas Mortensen, Roland Rau, and James W Vaupel. The divergent life-expectancy trends in Denmark and Sweden and some potential explanations. In *International differences in mortality at older ages: Dimensions and sources*, pages 385–407. The National Academies Press, Washington, DC, 2010.

Valeria D'Amato, Steven Haberman, Gabriella Piscopo, Maria Russolillo, and Lorenzo Trapani. Detecting common longevity trends by a multiple population approach. *North American Actuarial Journal*, 18(1):139–149, 2014.

Valeria D'Amato, Steven Haberman, and Gabriella Piscopo. The dependency premium based on a multifactor model for dependent mortality data. *Communications in Statistics-Theory and Methods*, 48(1):50–61, 2019.

Ana Debón, L Chaves, Steven Haberman, and F Villa. Characterization of between-group inequality of longevity in European Union countries. *Insurance: Mathematics and Economics*, 75:151–165, 2017.

Viani B Djeundje, Steven Haberman, Madhavi Bajekal, and Joseph Lu. The slowdown in mortality improvement rates 2011–2017: a multi-country analysis. *European Actuarial Journal*, pages 1–40, 2022.

Yumo Dong, Fei Huang, Honglin Yu, and Steven Haberman. Multi-population mortality forecasting using tensor decomposition. *Scandinavian Actuarial Journal*, 2020(8):754–775, 2020.

Ryan D Edwards and Shripad Tuljapurkar. Inequality in life spans and a new perspective on mortality convergence across industrialized countries. *Population and Development Review*, 31(4):645–674, 2005.

Abdolreza Eshghi, Dominique Haughton, and Pascal Legrand. Identifying groups: A comparison of methodologies. *Journal of Data Science*, 9(2):271–291, 2022. ISSN 1680-743X. doi: 10.6339/JDS.201104_09(2).0009.

Vladimir Estivill-Castro. Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter*, 4(1):65–75, 2002.

Pavel Grigoriev and Markéta Pechholdová. Health convergence between East and West Germany as reflected in long-term cause-specific mortality trends: to what extent was it due to reunification? *European journal of population*, 33(5):701–731, 2017.

Jacques A Hagenaars and Allan L McCutcheon. *Applied latent class analysis*. Cambridge University Press, 2002.

Petros Hatzopoulos and Steven Haberman. Common mortality modeling and coherent forecasts. an empirical analysis of worldwide mortality data. *Insurance: Mathematics and Economics*, 52(2):320–337, 2013.

Dominique Haughton, Pascal Legrand, and Sam Woolford. Review of three latent class cluster analysis packages: Latent Gold, poLCA, and MCLUST. *The American Statistician*, 63(1):81–91, 2009.

Jari Holopainen, Samuli Helama, and Timo Partonen. Finnish suicide mortality from 1950 to 2009 in an European comparison. *Duodecim; Laaketieteellinen Aikakauskirja*, 130(15):1536–1544, 2014.

Human Mortality Database. *University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Avaliable at www.mortality.org or www.humanmortality.de (Data downloaded in September 2022)*, 2022.

Søren Fiig Jarner, Esben Masotti Kryger, and Chresten Dengsøe. The evolution of death rates and life expectancy in Denmark. *Scandinavian Actuarial Journal*, 2008(2-3): 147–173, 2008.

David Kaplan. *The Sage handbook of quantitative methodology for the social sciences*. sage, 2004.

Ka Kin Lam and Bo Wang. Multipopulation mortality modelling and forecasting: the weighted multivariate functional principal component approaches. *Journal of Applied Statistics*, pages 1–22, 2022.

Finn Breinholt Larsen, Marie Hauge Pedersen, Karina Friis, Charlotte Glümer, and Mathias Lasgaard. A latent class analysis of multimorbidity and the relationship to socio-demographic factors and health-related quality of life. a national population-based study of 162,283 Danish adults. *PloS one*, 12(1):e0169426, 2017.

Ronald D Lee and Lawrence R Carter. Modeling and forecasting US mortality. *Journal of the American statistical association*, 87(419):659–671, 1992.

Ainhoa-Elena Léger and Stefano Mazzuco. What can we learn from the functional clustering of mortality data? an application to the human mortality database. *European Journal of Population*, 37(4):769–798, 2021.

Susanna Levantesi, Andrea Nigri, and Gabriella Piscopo. Clustering-based simultaneous forecasting of life expectancy time series through long-short term memory neural networks. *International Journal of Approximate Reasoning*, 140:282–297, 2022.

Nan Li and Ronald Lee. Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, 42(3):575–594, 2005.

Yisha Li, Margaret Ragland, Erin Austin, Kendra Young, Katherine Pratte, John E Hokanson, Terri H Beaty, Elizabeth A Regan, Stephen I Rennard, Christina Wern, et al. Co-morbidity patterns identified using latent class analysis of medications predict all-cause mortality independent of other known risk factors: The copdgene® study. *Clinical Epidemiology*, 12:1171, 2020.

Juhua Luo, Paul Dinh, Michael Hendryx, Wenjun Li, Jennifer Robinson, and Karen L Margolis. Risk patterns and mortality in postmenopausal women using latent class analysis. *American Journal of Preventive Medicine*, 61(5):e225–e233, 2021.

Kevin Milligan and Tammy Schirle. The evolution of longevity: Evidence from Canada. *Canadian Journal of Economics/Revue canadienne d'économique*, 54(1):164–192, 2021.

Andrea Nigri, Susanna Levantesi, and Gabriella Piscopo. Causes-of-death specific estimates from synthetic health measure: A methodological framework. *Social Indicators Research*, pages 1–22, 2022.

Timo Partonen, Olli Kiviruusu, Marjut Grainger, Jaana Suvisaari, Aki Eklin, Antti Virtanen, and Riitta Kauppila. Suicides from 2016 to 2020 in finland and the effect of the covid-19 pandemic. *The British Journal of Psychiatry*, 220(1):38–40, 2022.

Gabriella Piscopo and Marina Resta. Multi-country mortality analysis using self organizing maps. In *Recent Advances of Neural Network Models and Applications*, pages 233–240. Springer, 2014.

Gabriella Piscopo and Marina Resta. Applying spectral biclustering to mortality data. *Risks*, 5(2):24, 2017.

Samuel H Preston and Jessica Y Ho. Low life expectancy in the United States: Is the health care system at fault? Technical report, National Bureau of Economic Research, 2009.

Aida Santaolalla, Hans Garmo, Anita Grigoriadis, Sundeep Ghuman, Niklas Hammar, Ingmar Jungner, Göran Walldius, Mats Lambe, Lars Holmberg, and Mieke Van Hemelrijck. Metabolic profiles to predict long-term cancer and mortality: the use of latent class analysis. *BMC molecular and cell biology*, 20(1):1–15, 2019.

Statistics Finland. Official statistics of Finland (OSF): causes of death in 2019 [e-publication]. 2021.

Cary Chi-Liang Tsai and Echo Sihan Cheng. Incorporating statistical clustering methods into mortality models to improve forecasting performances. *Insurance: Mathematics and Economics*, 99:42–62, 2021.

Shripad Tuljapurkar, Nan Li, and Carl Boe. A universal pattern of mortality decline in the G7 countries. *Nature*, 405(6788):789–792, 2000.

James W Vaupel, Zhen Zhang, and Alyson A van Raalte. Life expectancy and disparity: an international comparison of life table data. *BMJ open*, 1(1):e000128, 2011.

Jeroen K Vermunt and Jay Magidson. Latent class cluster analysis. *Applied latent class analysis*, 11(89-106):60, 2002.

Jeroen K Vermunt and Jay Magidson. Latent gold 5.0 upgrade manual. *Belmont, MA: Statistical Innovations Inc*, 2016a.

Jeroen K Vermunt and Jay Magidson. Technical guide for latent gold 5.1: Basic, advanced, and syntax. *Belmont, MA: Statistical Innovations Inc*, 2016b.

John R Wilmoth, Carl Boe, Magali Barbieri, EM Crimmins, SH Preston, and B Cohen. Geographic differences in life expectancy at age 50 in the united states compared with other high-income countries. *International differences in mortality at older ages: Dimensions and sources*, pages 333–366, 2011.

Ruhao Wu and Bo Wang. Coherent mortality forecasting by the weighted multilevel functional principal component approach. *Journal of Applied Statistics*, 46(10):1774–1791, 2019.

Beijie Xu. Clustering educational digital library usage data: Comparisons of latent class analysis and k-means algorithms. In *Educational Data Mining*, 2013.