SAPIENZA
UNIVERSITÀ DI ROMA

Advances in Model Based Clustering for the Social Sciences

Department of Statistical Sciences

XXXV Ph.D. program in Methodological Statistics

Candidate

Emiliano Seri

ID number 1530531

Thesis Advisor

Prof. Roberto Rocci

Co-Advisor

Prof. Thomas Brendan Murphy

Thesis defended on 30/05/2023
in front of a Board of Examiners composed by:

Prof. Marco Alfò (chairman)

Prof. Enea Buongiorno

Prof. Roberto Di Marti

Prof. Arthur White

Advances in Model Based Clustering for the Social Sciences
Ph.D. thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: emiliano.seri@uniroma1.it

Abstract

This dissertation attempts to gather the main research topics I engaged during my PhD, in collaboration with several national and international researchers. The primary focus of this work is to highlight the power of model based clustering for identifying latent structures in complex data and its usefulness in the social sciences. This methods have become increasingly popular in social science research as they allow for more accurate and nuanced understanding of complex data structures. In the thesis are presented 3 papers that contribute to the development and application of model-based clustering in social science research, covering a range of scenario. The thesis pays particular attention to the practical applications of the treated methods, providing insights that can improve our understanding of complex social phenomena.

The first chapter of this dissertation introduces the usefulness of clustering model to deal with the complexity of society, and aware of some of the main issues when analysing socio-economic data. Following this conceptual introduction, the second chapter delves more into the technical aspects of model based clustering and estimation. These first two chapters pave the road for the three developments presented thereafter. The third chapter includes the application of a Mixture of Matrix-Normals classification model to the Migrant Integration Policy Index (MIPEX), that measures and evaluates countries policies toward migrants' integration over time. The used model is suitable for longitudinal data and allows for the identification of clusters of countries with similar patterns of migrant integration policies over time. The work is published in Alaimo et al. [2021a]. The fourth chapter uses MIPEX data too, but for a single year, and a finite mixtures of multivariate Gaussian is applied to identify groups of countries with a similar level of integration. Then, the relative proportion of immigrants held in prison among clusters is estimated, exploiting Fisher's noncentral hypergeometric model. The aim of this work is test the existence of an association between countries' level of integration of immigrants and the proportion of immigrants in prison. The work is currently in referral process. The fifth chapter introduce the work developed during my visiting research period at University of Lyon, Lyon 2. It specify the Bayesian partial membership model for soft clustering of multivariate data, namely when units have fractional membership to multiple groups. The model is specified for count data, and it is applied on the data of the bike sharing company of Washington DC and on the data of Serie A football players. The last chapter summarizes the main points of the dissertation, underlining the most relevant findings, the contributions, and stressing out how clustering models altogether yield a cohesive treatment of socio-economic data.

Contents

Chapter 1

Introduction

"if you haven't measured something, you really don't know very much about it." – *Karl Pearson*

My interest in statistics, and the reason why everyone should have a base knowledge in this discipline, stems from the fact that it is the basis of empirical knowledge of the world around us. We all have opinions about what surrounds us, but for these to be worthy to be taken seriously, they must be supported by empirical evidence. *Social statistics put statistical methodologies at the service of the social phenomena that surround us.* One of the main problems in this field is that socio-economic phenomena are made up from a network of elements, which interact both with one another and with the environment, so their complex and multidimensional nature makes them difficult to understand, analyse and represent. This led my interest to focus on the use of clustering models. Cluster analysis is the huge set of methods dedicated to finding groups in a set of objects characterized by certain measurements. This task has a very wide range of applications such for instance: biology, textual analysis, economy, or sociology. In this thesis the focus is dedicated to the use of clustering to analyse the complexity of society. When dealing with socio-economic phenomena, it is common for statistical units to be nations, regions, companies, or large economic constructs. Therefore, as each unit is an extremely complex construct, they cannot be treated as interchangeable. It follows that comparing units with all others should lead to misleading results. The purpose of clustering in social sciences, is to identify groups of units with similar behaviour considering all the aspects of the analysed phenomenon. Namely, finding homogeneous groups among the units can substantially improve the quality of the research. Let's think, for instance, of the Human Development Index [Undp, 1997] developed by United Nations Development Programme (UNDP), which attempt to measure the level of human development

of 191 countries, taking into account 3 dimensions: health, education and income. In this case, finding groups of countries which behave similarly improve the understanding of the phenomena and identify patterns in its complex data structures that may not be easily observable using traditional statistical methods. The clusters could allow for the improvement of the quality of the analysis, the interpretation of the results, their ease of reading, and consequently the diffusion of the results. Clustering consists of a huge and very heterogeneous number of methods. In this thesis, the focus is on those based on models. They offers the advantage of clearly stating the assumptions behind the clustering algorithm, and they allow cluster analysis to benefit from the inferential framework of statistics to address some of the practical questions arising when performing classification: determining the number of clusters, detecting and treating outliers, assessing uncertainty in the clustering [Bouveyron et al., 2019a]. The intent of this dissertation is to introduce the reader to the general ideas, motivation, advantages and potential limits of clustering models for the social sciences, providing a general description of those and presenting my main works on the subject.

1.1 Content of the thesis

To provide a description of model based clustering, in Chapter 2 is presented its general framework. From finite mixture models and how they can be applied to clustering, to how to deal with count data. My main works on the subject follow in the next three chapters. These follow a pattern starting with application of a clustering model for longitudinal data, followed by an application of a mixture of Gaussians clustering model for a single year, and finally a soft clustering model for count data. All works are applied to data of social interest. Chapter 3 presents an application of a Mixture of Matrix-Normals classification model to the Migrant Integration Policy Index. The model is suitable for longitudinal data, allowing for the identification of clusters of countries with similar patterns of migrant integration policies over time. The analysis identify 5 clusters of countries, with the aim to facilitate the evaluation and the comparison of the countries within each cluster and between different clusters over time. This work now published in Alaimo et al. [2021a]. In the work in Chapter 4, now under review, it is questioned whether there exists an association between countries' level of integration of immigrants and the proportion of immigrants in prison. To test such association, finite mixtures of multivariate Gaussian is used to identify three groups of countries with a similar level of integration toward migrants. Then, the relative proportion of immigrants held in prison among

clusters is estimated, exploiting Fisher's noncentral hypergeometric model. Chapter 5 introduce the work developed during my visiting research period at University of Lyon, Lyon 2 between November 2021 and June 2022. The research specify the Bayesian partial membership model for soft clustering of multivariate data, for the case when count data are considered. The model is applied on the data of the bike sharing company of Washington DC, with the aim to improve the allocations of the bikes. It is also applied to the data of the 192 Serie A football players, that played more than 1350 minutes during the 2022/2023 football season, with the aim of highlighting the attitude on the playing field of each player according to their statistics. The results of this application are also compared to those achieved with similar models. A general discussion is given in Chapter 6, highlighting the important findings and the contributions. Nevertheless, this dissertation still leaves room for many questions and open problems, hence some thoughts regarding promising directions for future research will be discussed.

The work developed during the first period of my PhD Alaimo and Seri [2023] is presented in the Appendix 6, as it is not strictly related to the topic of the thesis. In it some of the most known problems in composite indicators construction are pointed out, using the example of the Human Development Index. To address those issues, two indicators aggregation methods are proposed.

Chapter 2

Model Based Clustering

2.1 Finite mixture models

Mixture models born in the late nineteenth century from the biometrician, statistician and eugenicist Karl Pearson and the evolutionary biologist Walter Weldon, as an intuitively simple and practical tool for enriching the collection of probability distributions available for modelling data [Green, 2019]. The latter speculated in 1893 that the asymmetry he observed in a histogram of forehead to body length ratios in female shore crab populations could indicate evolutionary divergence. Pearson [1894] fitted a univariate mixture of two normals to Weldon's data by a method of moments, choosing the five parameters of the mixture so that the empirical moments matched those of the model. Since then, finite mixture models have been successfully applied to many fields, and underpin a variety of techniques in major areas of statistic, including cluster and latent class analyses, discriminant analysis, image analysis and survival analysis, in addition to their more direct role in data analysis and inference of providing descriptive models for distributions. Its flexibility makes a mixture model able to model quite complex distributions through an appropriate choice of its components to represent accurately the local areas of support of the true distribution. It can thus handle situations where a single parametric family is unable to provide a satisfactory model for local variations in the observed data [Peel and MacLahlan, 2000].

2.1.1 The basic formulation

Finite mixture densities are described in detail in Everitt and Hand [1981], Titterington et al. [1985], McLachlan and Basford [1988], Peel and MacLahlan [2000], Frühwirth-Schnatter

[2006]. The basic finite mixture model assumes that data are drawn from a density modelled as a convex combination of components each of specified parametric form [Green, 2019].

Let's consider a J-dimensional random vector $\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_J]^T$, with $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_J]^T$ representing one particular outcome of $\mathbf{Y}$. It is said that $\mathbf{Y}$ arises from a G-component parametric finite mixture distribution if, for all $\mathbf{y} \subset \mathbf{Y}$, its density can be written:

$$f(\mathbf{y}|\theta) = \sum_{g=1}^{G} \tau_g f_g(\mathbf{y}|\theta_g) \tag{2.1}$$

Where $\tau_g > 0$ such that $\sum_{g=1}^{G} \tau_g = 1$, is the gth mixing proportion, $f_g(y|\theta_g)$ is the gth component density, and $\theta = (\tau_1, \ldots, \tau_G, \theta_1, \ldots, \theta_G)$ is the vector of parameters. The component densities $f_1(\mathbf{y}|\theta_1), f_2(\mathbf{y}|\theta_2), \ldots, f_G(\mathbf{y}|\theta_G)$ are often taken to be of the same type. In this formulation of the mixture model, the number of components G is considered fixed, but of course in many applications, the value of G is unknown and has to be inferred from the available data, along with the mixing proportions and the parameters in the specified forms for the component densities.

Often, in classification applications, f_g is considered to arise from a multivariate Gaussian mixture model. A Gaussian mixture model has density:

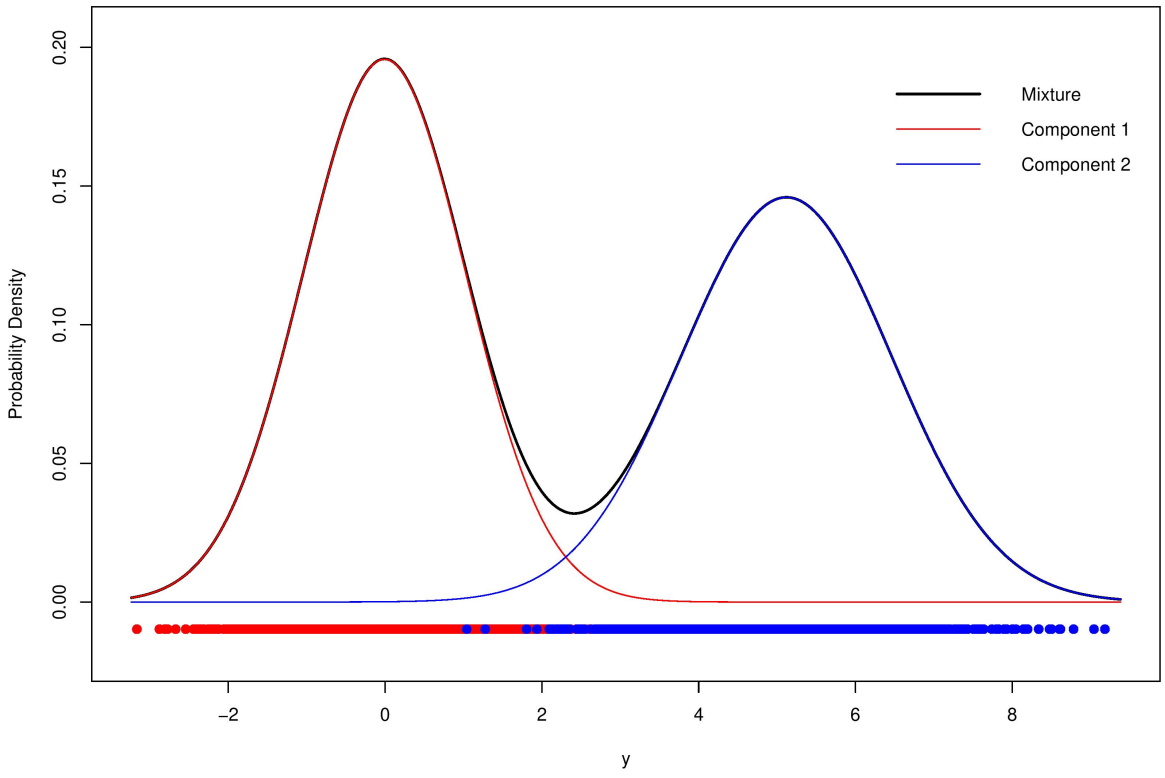
$$f(\mathbf{y}|\theta) = \sum_{g=1}^{G} \tau_g \phi(\mathbf{y}|\mu_g, \mathbf{\Sigma}_g) \tag{2.2}$$

Where

$$\phi(\mathbf{y}|\mu_g, \mathbf{\Sigma}_g) = \frac{1}{\sqrt{(2\pi)^J |\mathbf{\Sigma}_g|}} \exp\left\{ -\frac{1}{2} (\mathbf{y} - \mu_g)^T \mathbf{\Sigma}_g^{-1} (\mathbf{y} - \mu_g) \right\} \tag{2.3}$$

is the density of a random variable $\mathbf{Y}$ from a multivariate Gaussian distribution with mean μ_g and covariance matrix $\mathbf{\Sigma}_g$. In the univariate case, when $\mathbf{y}$ is one-dimensional, $f_g(\mathbf{y}|\theta_g)$ is a $N(\mu_g, \sigma_g^2)$ density function, and $\theta_g = (\mu_g, \sigma_g)$, consisting of the mean and standard deviation for the gth mixture component.

Figure 2.1. Probability density function for a one-dimensional univariate finite normal mixture with two components. The dots show a sample of size 2000 simulated from the density, with the colors indicating the mixture component from which they were generated.



Similar to the one in Bouveyron et al. [2019b], Figure 2.1 shows an example of the density function for a univariate finite normal mixture model with two mixture components, together with a sample simulated from it. Even in the represented case, where the two mixture components are well separated, it can be seen that 2 blue points are to the left of many red points. So even in this fairly clear situation there is uncertainty about which components the points in the middle belong to, if they were not conveniently colored. Assessing this kind of uncertainty is one of the goals of model-based clustering. Model based clustering refer to the cluster analyses methods, based on finite mixture models [Banfield and Raftery, 1993].

2.2 Model based clustering

Finite mixture models can describe populations for which the assumption of a finite number of subpopulations is valid. It follows that finite mixtures provide suitable models for cluster

analysis under the assumption that each group of observations in a data set suspected to contain clusters, comes from a population with a different probability distribution. The latter may belong to the same family, but differ in the values they have for the parameters of the distribution. So, by using finite mixture densities as models for cluster analysis, the clustering problem becomes that of estimating the parameters of the assumed mixture and then using the estimated parameters to calculate the posterior probabilities of cluster membership [Everitt et al., 2011]. In this chapter, a multivariate Gaussian mixture model is used for illustration.

Let denote the component membership of observation i, so that $z_{ig} = 1$ if observation i belongs to component g and $z_{ig} = 0$ otherwise. Consider a clustering scenario, where n j-dimensional data vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are observed and all are unlabelled or treated as unlabelled. Then the Gaussian model-based clustering likelihood can be written

$$L(\theta) = \prod_{i=1}^{n} \sum_{g=1}^{G} \tau_g \phi(\mathbf{y}_i | \mu_g, \mathbf{\Sigma}_g) \tag{2.4}$$

Note that z_i is considered a realization of the component-label vector $\mathbf{Z}_i$, which is a random variable that follows a multinomial distribution with one draw on G categories with probabilities given by $\tau_1, \ldots, \tau_G$. In fact, $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are assumed independent and identically distributed according to a multinomial distribution with one draw on G categories with probabilities $\tau_1, \ldots, \tau_G$. The mixing proportion τ_g can be interpreted as the *a priori* probability that an observation y_i belongs to component g. The corresponding *a posteriori* probability is

$$P\left[Z_{ig} = 1 | \mathbf{y}_i\right] = \frac{\tau_g \phi(\mathbf{y}_i | \mu_g, \mathbf{\Sigma}_g)}{\sum_{h=1}^{G} \tau_h \phi(\mathbf{y}_i | \mu_h, \mathbf{\Sigma}_h)} \tag{2.5}$$

Having estimated the parameters of the mixture distribution, observations can be associated with particular clusters on the basis of the maximum value of the estimated posterior probability 2.5.

2.2.1 Maximum likelihood estimation

Over the years, a variety of approaches have been used to estimate mixture distributions. They include graphical methods, method of moments, minimum-distance methods, maximum likelihood, and Bayesian approaches. But as explained in Titterington et al. [1985], the dominant approach to inference about unknowns in mixture models is by maximum likelihood, based on maximization of Equation 2.1, usually achieved through the Expectation-Maximization (EM)

algorithm [Dempster et al., 1977]. The EM algorithm, greatly stimulated interest in the use of finite mixture distributions to model heterogeneous data. This is because the fitting of mixture models by maximum likelihood is a classic example of a problem that is simplified considerably by the EM's conceptual unification of maximum likelihood (ML) estimation from data that can be viewed as being incomplete Peel and MacLahlan [2000].

2.2.2 EM algorithm

Consider j-dimensional observed data vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$. Let z_{ig} denote component membership, where

$$z_{ig} = \begin{cases} 1 & \text{if } \mathbf{y}_i \text{ belongs to group } g \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

$\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})$ is unobserved. Parameter estimation can be carried out using an EM algorithm, where the complete-data comprise the observed $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and the labels $\mathbf{z}_1, \ldots, \mathbf{z}_n$. If the $(\mathbf{y}_i, \mathbf{z}_i)$ are independent and identically distributed (iid) according to the probability distribution $\phi(\mathbf{y}_i|\mu_g, \mathbf{\Sigma}_g)$, then the complete-data likelihood is given by

$$L_c(\theta) = \prod_{i=1}^{n} \prod_{g=1}^{G} [\tau_g \phi(\mathbf{y}_i|\mu_g, \mathbf{\Sigma}_g)]^{z_{ig}} \tag{2.7}$$

where θ denotes the model parameters, i.e., with $\tau = (\tau_1, \ldots, \tau_G)$. The observed data likelihood, $L_o(\theta)$, can be obtained by integrating the unobserved data $\mathbf{z}$ out of the complete-data likelihood. The natural logarithm of 2.7 gives the complete-data log-likelihood

$$l_c(\theta) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \left[\log \tau_g + \log \phi(\mathbf{y}_i|\mu_g, \mathbf{\Sigma}_g)\right] \tag{2.8}$$

The EM algorithm alternates between two steps. The first is an "E-step", or expectation step, in which the conditional expectation of the complete data log-likelihood given the observed data and the current parameter estimates is computed. The second is an "M-step", or maximization step, in which parameters that maximize the expected log-likelihood from the E-step are determined. In the E-step, the expected value of the complete-data log-likelihood is updated. This amounts to replacing the z_{ig} in 2.8 by their expected values

$$\hat{z}_{ig}^{(q)} = \frac{\hat{\tau}_g^{(q-1)}\phi(\mathbf{y}_i|\hat{\mu}_g^{(q-1)}, \hat{\mathbf{\Sigma}}_g^{(q-1)})}{\sum_{h=1}^{G}\hat{\tau}_h^{(q-1)}\phi(\mathbf{y}_i|\hat{\mu}_h^{(q-1)}, \hat{\mathbf{\Sigma}}_h^{(q-1)})} \tag{2.9}$$

for $i = 1, \ldots, n$ and $g = 1, \ldots, G$, where $\hat{\tau}_g^{(q)}$ is the value of τ after the qth EM iteration. Note that, in the E-step, the conditioning is on the current parameter estimates, hence the use of hats on the parameters in 2.9. It follows that the expected value of the complete-data log-likelihood is

$$\begin{aligned} Q(\theta) &= \sum_{i=1}^{n}\sum_{g=1}^{G}\hat{z}_{ig}^{(q)}\left[\log\tau_g - \frac{j}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}_g| - \frac{1}{2}\mathrm{tr}\left\{(\mathbf{y}-\mu_g)(\mathbf{y}-\mu_g)^T\mathbf{\Sigma}_g^{-1}\right\}\right] \\ &= \sum_{g=1}^{G} n_g\log\tau_g - \frac{nj}{2}\log 2\pi - \sum_{g=1}^{G}\frac{n_g}{2}\log|\mathbf{\Sigma}_g| - \sum_{g=1}^{G}\frac{n_g}{2}\mathrm{tr}\{\mathbf{S}_g\mathbf{\Sigma}_g^{-1}\} \end{aligned} \tag{2.10}$$

where $n_g = \sum_{i=1}^{n}\hat{z}_{ig}^{(q)}$ and

$$\mathbf{S}_g = \frac{1}{n_g}\sum_{i=1}^{n}\hat{z}_{ig}^{(q)}(\mathbf{y}-\mu_g)(\mathbf{y}-\mu_g)^T \tag{2.11}$$

In the M-step, the model parameters are updated. Then, maximizing $Q(\theta)$ with respect to τ_g, μ_g, and $\mathbf{\Sigma}_g$ yields the updates

$$\hat{\tau}_g^{(q)} = \frac{\hat{n}_g^{(q-1)}}{n}, \quad \hat{\mu}_g^{(q)} = \frac{\sum_{i=1}^{n}\hat{z}_{ig}^{(q-1)}\mathbf{y}_i}{\hat{n}_g^{(q-1)}}, \quad \hat{n}_g^{(q-1)} = \sum_{i=1}^{n}\hat{z}_{ig}^{(q-1)}$$

Computation of the covariance estimate $\hat{\mathbf{\Sigma}}_g^{(q)}$ depends on its parameterization. For further details on parameterizations of the covariance matrix, see Bouveyron et al. [2019b]. Following, it is considered the more general model, where no constraints are placed on the volumes, shapes and orientations of the mixture covariance matrices

$$\hat{\mathbf{\Sigma}}_g^{(q)} = \frac{1}{\hat{n}_g^{(q-1)}}\sum_{i=1}^{n}\hat{z}_{ig}^{(q-1)}(\mathbf{y}-\hat{\mu}_g^{(q)})(\mathbf{y}-\hat{\mu}_g^{(q)})^T$$

The EM algorithm for Gaussian model-based clustering alternates between the E and M steps until convergence. Convergence of the EM algorithm to a local maximum of the log-likelihood function can be assessed in several ways. In essence, these consist of seeing whether the algorithm has been moving slowly in the latest iterations. One possible criterion is that the log-likelihood has changed very little between the last two iterations; a typical threshold is a change of less

than 10^{-5}.

A typical issue, is that as the likelihood for a mixture model is in general not convex, so there can be local maxima. Methods for dealing with this problem can be classified as constraint methods, Bayesian methods, penalty methods, and others. Most of them are reviewed in Bouveyron et al. [2019b]. Furthermore, the converged estimate can depend on the initial value chosen. Many methods for the choice of the starting values for the EM algorithm in multivariate Gaussian mixture models, are explained and compared in Biernacki et al. [2003], Maitra [2009], Melnykov and Melnykov [2012].

An application of clustering via finite mixtures of Gaussians is in Chapter 4.

2.2.3 Bayesian analysis of mixtures

The growth of computing power and the development of the Markov chain Monte Carlo (MCMC) sampling method for estimating the parameters of Bayesian models, attracted increasing interest to Bayesian statistics amongst statisticians, also in the area of model based clustering. As explained in Everitt et al. [2011], there are two main reasons why a Bayesian approach to fitting finite mixture models is worth considering. First, Bayesian modelling allows parameter estimation for models where the likelihood method fails because of singularities in the likelihood surface; here Bayesian modelling is employed primarily for pragmatic reasons. The second reason for the increasing interest in Bayesian mixture modelling is rather philosophical. The Bayesian approach allows probabilistic statements to be made directly about the unknown parameters, and findings from previous research or from expert opinion can be incorporated with the prior distribution. Richardson and Green [1997] argue that the Bayesian paradigm is preferable on grounds of convenience, accuracy and flexibility to the use of analytic approximation if the number of components is unknown. A detailed account of Bayesian methods for finite mixtures is given in Rousseau et al. [2019]. Bayesian finite mixture modelling is, however, not without its problems, of which the most important are: can be computationally demanding, the choice of a prior distribution and the component label-switching problem during MCMC sampling.

The choice of the prior distribution In Bayesian inference the prior distribution $P(\theta)$, describes the information about the parameter θ before the data are seen. After observing the data, the prior distribution is updated, using the likelihood of the data given θ, to the posterior distribution $P(\theta|\mathbf{y})$, which provides the basis for statistical inferences [Everitt et al., 2011]:

$$P(\theta|\mathbf{y}) \propto L(\mathbf{y}|\theta)P(\theta) \tag{2.12}$$

The choice of the prior, which should reflect the available knowledge before the data are seen, is important since it influences the posterior inference. In cluster analysis the number of clusters and the parameters of the cluster model are usually unknown. In this case the prior should have little influence on inference, and the data should mainly determine the posterior distribution through the likelihood [Everitt et al., 2011].

The label switching problem A common problem with the Bayesian approach to model based clustering, is that of label switching during MCMC sampling, which arises because of the symmetry in the likelihood of the model parameters. For Bayesian mixtures the likelihood function and the resulting posterior distribution is invariant under permutations with respect to component labels. The labels of the components during one run of an MCMC sampler may be switched on different iterations, and the lack of identifiability gives rise to problems when making inferences about the individual components. To obtain a meaningful interpretation of the components it is necessary to account for label switching so that components are in the same order at each iteration [Everitt et al., 2011]. One way to deal with the label-switching problem is to impose identifiable constraints on a particular set of model parameters. Another way is to impose a reordering constraint after the simulations have been done. Some algorithm for this purpose are Stephens [2000], Marin et al. [2005], Papastamoulis and Iliopoulos [2010], or probabilistic Sperrin et al. [2010].

2.3 Mixtures of Poisson distributions

The Gaussian distribution is the reference distribution for model based clustering with continuous data. However, it is designed for continuous-valued data, not discrete data, so it is not well adapted for dealing with categorical or count data. For count data the group conditional distributions are typically assumed to be Poisson [Agresti, 2002].

The probability distribution function of a Poisson distribution $\boldsymbol{Y}$ with parameter λ is:

$$Pois(y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} \tag{2.13}$$

where y is a non-negative integer. It satisfies $E(\boldsymbol{Y}) = var(\boldsymbol{Y}) = \lambda$. It should be noted that λ is

not necessarily a continuous random variable. It can be discrete or it can take a finite number of values. The latter case gives rise to finite Poisson mixtures. It is known that the role of the simple Poisson distribution is prominent among the discrete probability distributions. The Poisson distribution is usually used to model situations where only randomness is present. In practice this probabilistic mechanism cannot explain data variation and the need for alternative more complex models is obvious [Dellaportas et al., 2011]. Poisson mixture models are then interesting candidates. In the multivariate setting, the most used multivariate Poisson distribution assumes that the variables are independent. Alternative multivariate Poisson distributions exist [Karlis, 2003] but are difficult to analyze, especially for high-dimensional data. For this reason, the variables are generally assumed to be independent conditionally on the class the observation belongs to [Bouveyron et al., 2019c]. Also, $y_{i|g}$, the conditional distribution of the variable y_i given that observation i belongs to class g, is assumed to follow a Poisson distribution. This is denoted by $y_{i|g} \sim \prod_{j=1}^{J} Pois(y_{ij}|\mu_{ijg})$. Following Bouveyron et al. [2019c] notation, this leads to the following Poisson mixture model:

$$f(y_i|\mu_i, \tau) = \sum_{g=1}^{G} \tau_g \prod_{j=1}^{J} Pois(y_{ij}|\mu_{ijg}) \tag{2.14}$$

where $\sum_{g=1}^{G} \tau_g = 1$ and $\tau_g \geq 0$ for $g = 1, \ldots, G$. The unconditional mean and variance of $\boldsymbol{Y}_{ij}$, respectively, are:

$$E(\boldsymbol{Y}_{ij}) = \sum_{g=1}^{G} \tau_g \mu_{ijg}$$

and

$$Var(\boldsymbol{Y}_{ij}) = \sum_{g=1}^{G} \tau_g \mu_{ijg} + \sum_{g=1}^{G} \tau_g \mu_{ijg}^2 - \left(\sum_{g=1}^{G} \tau_g \mu_{ijg} \right)^2$$

As in Bouveyron et al. [2019c], following Rau et al. [2015], it is considered the following parameterization for the mean μ_{ijg}:

$$\mu_{ijg} = \omega_i \lambda_{jg} \tag{2.15}$$

where ω_i is the intensity level for the observation i and $\lambda_g = (\lambda_{1g}, \ldots, \lambda_{Jg})$ corresponds to the clustering parameters that define the profiles of the observations in cluster g across all variables. The y_{ij} are assumed to be independent given the z_i (latent class model). Each y_{ij} will be distributed according to the Poisson distribution with mean $\omega_i \lambda_{jg}$. Thus the Poisson parameter is expressed as the product of two quantities, namely ω_i, the effect of the row i, and λ_{jg}, the

effect of the component g on the variable j. The probability distribution $f_g(y_i;\theta)$ is then a product of Poisson distributions

$$f_g(y_i;\theta) = \prod_j \frac{e^{-\omega_i\lambda_{jg}}(\omega_i\lambda_{jg})^{y_{ij}}}{y_{ij}!}$$

where $\theta = (\tau_1 \ldots, \tau_g, \omega_1, \ldots, \omega_n, \lambda_{11}, \ldots, \lambda_{Jg})$. in order to ensure the identifiability of the finite mixture, the parameters ω_i and λ_{jg} in Equation 2.15 should be constrained as such $\sum_j \lambda_{jg} = 1$ for all $g = 1, \ldots, G$. The interpretation of this constraint is that the parameters λ_{jg} represent the percentage of total counts per observation that are attributed to each variable. A thorough review of the existing literature on Poisson mixtures is given in Karlis and Xekalaki [2005].

The parameters τ, ω_i and λ_{jg} of the resulting mixture of Poisson distributions can be estimated either in a classical maximum likelihood approach or in a Bayesian framework, coupled with MCMC techniques. The standard Bayesian formulation of the finite mixture problem with a known number of components and its implementation via Markov Chain Monte Carlo (MCMC) is given by Diebolt and Robert [1994]. Other developments of Bayesian estimation via MCMC for finite Poisson mixtures are in Dellaportas et al. [2011], Viallefont et al. [2002].

Chapter 3

A comparison of migrant integration policies via Mixture of Matrix-Normals

Abstract

In recent decades, there has been a growing interest in comparative studies about migrant integration, assimilation and the evaluation of policies implemented for these purposes. Over the years, the Migrant Integration Policy Index (MIPEX) has become a reference on these topics. This index measures and evaluates the policies of migrants' integration in 52 countries over time. However, the comparison of very different countries can be difficult and, if not well conducted, can lead to misleading interpretations and evaluations of the results. The aim of this paper is to improve this comparison and facilitate the reading of the considered phenomenon, by applying a Mixture of Matrix-Normals classification model for longitudinal data. Focusing on data for 7 MIPEX dimensions from 2014 to 2019, our analysis identify 5 clusters of countries, facilitating the evaluation and the comparison of the countries within each cluster and between different clusters.

3.1 Introduction

Immigration regulation and immigrants assimilation have been salient political issues in all industrialised countries for many decades, mainly because of their cultural and economic effects [Alesina and Tabellini, 2022]. The growing interest in the study of immigration, starting from citizenship and moving more recently to integration, has led to a variety of attempts to quantify immigration policies. Policy indices have become mandatory in the study of immigrant-related policies implemented by different countries. However, the study of these phenomena from a quantitative point of view is rather recent, due to the previous lack or difficulties to access of data [Bjerre et al., 2015]. Moreover, quantifying migrant integration is a difficult challenge, linked to its complex nature and lack of uniformity in migration policies of many countries, which are based on multiple criteria.

In this work, we focus on the Migrant Integration Policy Index (MIPEX) [Niessen et al., 2007, Solano and Huddleston, 2020], a complex system of 167 policy indicators across 8 domains of citizenship and integration, combined into a single composite indicator in order to evaluate the migrant integration policies of each considered country over the years. MIPEX has quickly become a solid and useful tool for evaluating and comparing what governments are doing to promote the migrants' integration in a cross-country setting. Indeed, it informs and engages key policy actors about how to use indicators to improve integration governance and policy effectiveness, with

the aim to measure policies that promote integration in both socio-economic and civic terms. Although not without its critics, this index has become a reference for comparative studies on migrant integration over the last decade and its data has been widely used in literature [Hadjar and Backes, 2013, Ruedin, 2015, Rayp et al., 2017, Ingleby et al., 2019]. This paper aims to deeply look at how similar, or dissimilar, countries really are and to add new reading perspectives on the MIPEX data, by discovering structures and patterns in the behaviour of the considered countries. The underlying idea is that, given the complex and multidimensional nature of the phenomenon and the differences in socio-economic and civic terms between the examined countries, it can be misleading to compare all of the units with each others. Therefore, the present work aims at improving the analysis, by grouping countries in order to facilitate the comparison and interpretation of the phenomenon. Thus, the research question to which we try to answer:

- *In order to improve the comparison between the countries regarding their migrant integration policies, is it possible to identify homogeneous groups over time among them, i.e. groups of countries which behave similarly across and within time?*

To answer this research question, a Finite Mixture of Matrix-Normals model has been applied to cluster the units, taking into account the longitudinal dimension along 6 years, on the 52 available countries for 7 of the 8 dimensional indicators of the MIPEX. We relied on an unsupervised parametric clustering approach to minimize the risk of arbitrariness[1] in the choices made and to be able to better evaluate the results.

The paper is structured as follows. Section 3.2 describes the immigrants integration framework and some works related to migration indicators. Section 3.3 presents the description of the analysed data and the structure of the MIPEX theoretical framework. In Section 3.4 we present the methodology implemented. Section 3.5 reports data analysis and the results and Section 3.6 concludes.

[1]Subjectivity is an essential element in any measurement process, but its presence does not make the process arbitrary [Alaimo, 2020].

3.2 Theoretical framework and related works

3.2.1 Immigrants integration framework

Immigration can be generally defined as the set of policies that determine who can enter or exit a country under what conditions, as well as how immigrants are considered once they are settled in a country. Many factors contribute to the migratory flows and stocks (forced or voluntary) to destination countries, which have been extensively addressed in the literature [Dustmann and Preston, 2007, Pedersen et al., 2008, Simpson, 2017]. We distinguish short-term migrants (seasonal agricultural workers, students, tourists, or temporary residents) and long-term migrants that include permanent residents, the first step on a path towards the creation of members, namely the citizenship [Goodman, 2019, Solano and Huddleston, 2021]. Migration and migrant integration dynamics influence the number and characteristics of migrants entering a country, as well as the integration outcomes [Helbling and Leblang, 2019, Garcés-Mascareñas and Penninx, 2016, Czaika and De Haas, 2013, Massey et al., 1998]. At the same time, the receiving society defines all the laws and policies that relate to the selection, admission, integration, settlement, and full membership of migrants in a country [Solano and Huddleston, 2021, Bjerre et al., 2015, Hammar, 1990]. Citizenship, migration, and integration policy, albeit in different ways, are distinct policy domains and creates the conditions that support or hinder migrants' inclusion in the destination society. More attention has been paid to integration policies in recent years, so much so that, in modern countries, they have evolved into very complex legal constructs [Zincone et al., 2011], whereas previously the focus was more on immigrant or assimilation policies. Moreover, as reported in Ramakrishnan [2013], in several countries terms like *assimilation*, *adaptation*, *incorporation* and *integration*, often refer to the same concept and some efforts were needed to provide more conceptual clarity, especially in finding unambiguous definitions of fundamental concepts on the matter. Castles and Davidson [2000] highlight that countries have three main policy options with respect to managing social diversity. The first option is *exclusion*. Although this model is not considered legitimate by humanitarian standards and formally not accepted, it should be noted that it is still predominant in large areas of the world. The second option is *assimilation*. According to this policy model, immigrants should be granted full citizenship: the immigrants' distinct culture is seen as in transition and it is expected that they fully adopt the national culture and generally accepted social norms. The third option is *integration*, with respect which policy makers are aware that immigrants do not abandon their distinct culture immediately and, therefore, their cultural identity can be

considered an opportunity. Legal integration, intended as an immigrant's legal status, residence rights, citizenship, and equal access to rights, goods, services, and resources, receives wide expert acceptance as the first step in promoting societal integration. It is considered a key determinant [Penninx and Martiniello, 2004] and can hardly be overestimated as either "a firm base" for societal integration or a "clear signal" committing public authorities to an inclusive agenda [Groenendijk et al., 1998]. These differences are strictly linked to the complex nature of immigration policies, which involve different political, social and economical spheres that are interconnected with each other. As explained in Niessen and Huddleston [2009], integration is developed by policymakers in conjunction with their policies on social inclusion/cohesion, employment, demography, competitiveness. It follows that immigrant integration is only one part of the broader good governance framework. In recent years, various studies have tried to develop this framework and quantitative indices of immigration policies have been proposed. These indices play a central role in the study of immigrant-related policies, starting with citizenship and moving to immigration and integration [Goodman, 2019, 2015, Helbling, 2013]. The next sub-section, although not exhaustively, present some of the most used immigrant-related policy indexes, highlighting how over time they assume greater specificity in relation to integration policies.

3.2.2 Immigration policies indexes: a literature review

The policy indices reflect the tendency in social sciences to reduce the complexity of socio-economic phenomena, allowing comparisons across countries and times [Rainer and Marc, 2011, Skaaning, 2010]. A sample of immigrant-related policy indexes will be presented below, providing information on index content, type, scope, and source. All of the indices reported in this paper make important and innovative contributions to the field of comparative immigration policy research. It is not our goal to discuss whether and which indexes are better than others. Each index has different methodological and conceptual assumptions and answers specific research questions. In the migratory field, the first index was proposed by Waldrauch and Hofinger [1997] in a study on citizenship, examining the Legal Obstacles to Integration (LOI). But indexing did not stop at citizenship. Several studies have documented the expansion of indexing from citizenship to integration, assuming more specificity for immigration policies [Goodman, 2019, 2015, Helbling, 2013]. The first immigrant-related policy indexes proposed, do not differentiate between immigration and integration policy domains. An exception is represented by the index

proposed by Boushey and Luedtke [2011], who first consider the distinction between immigration control and immigrant integration measures. This index provides “conceptual clarification to indexing by distinguishing immigration as control policies [that] deal with keeping out “unwanted immigrants" and integration policy as dictat[ing] the transition and settlement of resident immigrants" [Goodman, 2019, p. 579]. Recently, an interdisciplinary community of scholars has developed multi-dimensional indices capable of differentiating across types of policies, target groups, and instruments [Goodman, 2019, 2010, Koopmans et al., 2012]. We briefly present some of the main ones:

- First released by Banting et al. [2006], the *Multiculturalism Policy Index* (MCP) is a scholarly research project that monitors the evolution of multiculturalism policies in 21 Western democracies. The MCP is designed to provide information about multiculturalism policies in a standardized format that aids comparative research and contributes to the understanding of State-minorities relations. The project provides an index at 3 points in time: 1980, 2000, 2010, and for 3 types of minorities: one index relating to immigrant groups; one relating to historic national minorities; one index relating to indigenous peoples.

- The Migrant Integration Policy Index (MIPEX) [Niessen et al., 2007, Solano and Huddleston, 2020] is a complex system of 167 policy indicators across 8 domains of citizenship and integration combined into a single composite indicator, in order to evaluate the migrant integration policies of each considered country (for details, see Section 3.3).

- Based on the selection of data for 9 countries, between 1999 and 2008, and with the aim of measuring and comparing immigration, asylum, and naturalization policies across countries, the *International Migration Policy and Law Analysis* (IMPALA) database collects comparable data on immigration law and policy across 6 major areas of migration legislation: economic migration, family reunification, humanitarian migration, irregular migration, student migration, and the acquisition and loss of citizenship for migrants resident [Gest et al., 2014, Beine et al., 2016].

- Helbling et al. [2017] presented the *Immigration Policies in Comparison* (IMPIC) project, which proposes a data set that allows to measure immigration regulations.

- *The Canadian Index for Measuring Integration* (CIMI), is an interactive tool that allows for measuring the outcomes of immigrants in Canadian regions. It is a data-driven index that examines 4 dimensions of immigrants’ integration in Canada to assess the gaps between

immigrants and the Canadian-born population. The CIMI identifies factors that underline successful immigrants' integration, assesses changes and trends over time (currently from 1991 to 2020), enables detailed examination of 4 dimensions of integration and provides rankings based on empirical evidence for Canadian geographies.

- The *Immigration Policy Lab* (IPL) [Harder et al., 2018] is a survey-based measure of immigrant integration, to provide scholars with a short instrument that can be implemented across survey modes, with the aim to strike a pragmatic compromise to help generate cumulative knowledge on immigrant integration. The IPL captures 6 dimensions of integration: psychological, economical, political, social, linguistical, and navigational.

With the proliferation of such policy indices, scholars have more refined tools than ever for classifying and comparing policy plans and practices. Immigration and integration policies vary across dimensions, and limiting them to a single dimension reduces the ability to observe variations that could be significant. For this reason, we focused our analysis on MIPEX dimensions instead of the final composite indicator.

3.3 Data

Analyzing a complex phenomenon [Alaimo, 2021a] is often connected to the measuring of some non-directly measurable latent variables [Maggino et al., 2021, Maggino and Alaimo, 2021, 2022]. The measurement process in social sciences is associated with the construction of system of indicators. The indicators within a system are interconnected and new properties typical of the system emerge from these interconnections. As it can be easily understood, these kinds of systems are complex systems [Alaimo, 2021b]. Therefore, a system of indicators allows the measurement of a complex concept that would not otherwise be measurable by taking into account the indicators individually [Alaimo and Maggino, 2020].

The MIPEX is a system of 167 policy indicators[2] and it includes 52 countries and collects data from 2007 to 2019, in order to provide a view of integration policies across a broad range of differing environments. The values of each indicator are chosen by experts from each country, by means of a questionnaire. The MIPEX synthetic indicator is constructed by means of an aggregative-compensative approach [Nardo et al., 2005, OECD, 2008, Alaimo and Maggino, 2020]. The 167 basic indicators are first aggregated in 58 indicators (for more information, please

[2] A policy indicator is a question relating to a specific policy component of one of the 8 policy areas.

consult Solano and Huddleston [2020]), which cover the 8 policy areas designed to benchmark current laws and policies against the highest standards through consultations with top scholars and institutions,[3]. The policy areas of integration covered are the following::

- Labour Market Mobility (X1)
- Family Reunion (X2)
- Education (X3)
- Political Participation (X4)
- Long-term Residence (X5)
- Access to Nationality (X6)
- Anti-discrimination (X7)
- Health[4]

For each area, a synthetic measure (dimensional) is calculated as the arithmetic mean of the elementary indicators[5], i.e. those selected for measuring each policy area. Each dimensional synthetic indicator is bounded between $[0, 100]$: the higher the value, the better the situation in that policy area.

The method and the approach adopted for the construction of the synthetic index have not been without criticism. Even if it is the most widespread among the aggregation methods for composite indicators construction, the arithmetic mean it has been highly criticized. The main advantage of this method is that it is simple, largely known and gives easy-to-understand results. The main drawback is that it is a full compensative method; consequently, low values in some indicators can be compensated by high values in other ones [OECD, 2008]. This assumption is very strong and has a great impact on the results obtained, leading in many cases to an extreme flattening of the differences between the units [Alaimo and Seri, 2021]. Despite its success, the aggregative-compensative approach has been deeply criticized as inappropriate and often inconsistent, from both conceptual and methodological point of view [Freudenber, 2003, Maggino, 2017, Fattore, 2017]. To address and try to overcome the limitations of this approach, in recent years alternative procedures to synthesis have been developed in the literature [for instance, see: Kerber and Brüggemann, 2015, Kerber, 2017, Alaimo et al., 2021b, 2022b]. However, the purpose of this paper is to improve the analysis of the dimensions of MIPEX in its present form, albeit we suggest a critical read of it. The analysis carried out in the present work uses the listed above dimensions (excluding health), of which we are going to give a brief description in

[3]The highest standards are drawn from Council of Europe Conventions, European Union Directives and international conventions (for more information see: `http://mipex.eu/methodology`).

[4]This dimensions was excluded from the analysis, because it presents data only available for years 2014 and 2019.

[5]The elementary indicators are described in [Solano and Huddleston, 2020].

the following sub-sections[6].

3.3.1 Labour Market Mobility

Integration of immigrants into the labor market is a process that happens over time and depends on general policies, context, immigrants' skills and the reason for migration. Labour market mobility policies qualify as only halfway favourable for promoting equal quality employment over the long-term. In most countries, family members and permanent residents can access the labour market and job training, as well as social security and assistance. However, according to Solano and Huddleston [2020], full equality of rights and opportunity in the labour market is still far from being achieved, especially in the public sector.

3.3.2 Family Reunion

Family reunification policies determine if and when separated families can reunite and settle in their new home. According to Solano and Huddleston [2020], policies are more favourable in traditional destination countries, Northern European countries and new countries of labour migration (e.g. Italy, Portugal and Spain). On the other hand, for family reunification some countries require a high fee to pay and little support (e.g. Austria, Denmark, France, Germany, the Netherlands, Switzerland, UK). Increasingly, countries make exceptions for the highly-skilled and the wealthy, but rarely for the most vulnerable (minors and beneficiaries of international protection).

3.3.3 Education

Despite being an increasing priority for integration, education is the greatest weakness in the integration policies of many countries. Most immigrant pupils receive little support in finding the right school or class, or in 'catching up' with their peers. As described in Solano and Huddleston [2020], Australia, Canada and New Zealand have developed strong targeted education policies through multiculturalism, while the US focuses additional support on vulnerable racial and social groups. In contrast, the education systems of Austria, France, Germany and Luxembourg are less responsive to the needs of their relatively large number of immigrant pupils. New destination countries with small immigrant communities offer inconsistent targeted support (e.g. Japan and Central Europe).

[6] A more extensive explanation is given in [Solano and Huddleston, 2020].

3.3.4 Political Participation

In most countries, foreign citizens are not enfranchised or regularly informed, consulted or involved in local civil society and public life. Political participation is one of the weakest areas of integration [Solano and Huddleston, 2020]. Foreign citizens' political opportunities differ enormously from one country to another. For instance, in Australia, New Zealand and Western Europe, they enjoy greater voting rights, stronger consultative bodies, more funding for immigrant organisations and greater support from mainstream organisations. With the exception of Korea, immigrants in Asian countries enjoy almost none of these rights unless they (can) naturalise. Despite European norms and promising regional practices, political participation is still almost absent from integration strategies in Bulgaria, Lithuania, Romania and Slovakia.

3.3.5 Long-term Residence

The security of permanent residence may be a fundamental step on the path to full citizenship and better integration outcomes. Permanent residence is a normal part of the integration process in top-scoring countries in the MIPEX composite indicator, such as Canada, most Latin American countries (Brazil, Chile and Mexico), Nordic countries (Finland and Sweden), and few other European countries (Hungary, Iceland, Slovenia, Ukraine). In contrast, many newcomers are ineligible for permanent residence in China, Denmark, Ireland, Israel, Japan, Switzerland and Turkey. Countries rarely reform their legal routes to permanent residence. The limited major reforms of recent years have been driven by the politicisation of immigration. Brazil, Estonia, Macedonia, Russia, and Turkey have removed previous restrictions, while Austria, Denmark, Korea, Norway, Poland, Ukraine and the US have imposed new ones.

3.3.6 Access to Nationality

Facilitating access to nationality can significantly increase naturalisation rates and boost integration outcomes. Nationality policies are a major area of weakness in most European and non-European countries [Solano and Huddleston, 2020], especially Austria, Bulgaria, the Baltics, Eastern Europe, and India. By contrast, immigrants have favourable opportunities to become citizens in many countries, e.g., Sweden and the traditional destination countries (Canada, New Zealand and US). Since 2014, nationality policies have become more restrictive in Argentina, Denmark, Greece and Italy, while immigrants' access to nationality has improved significantly in Brazil and Luxembourg and, to lesser extent, in China, Greece, Latvia, Moldova, Portugal,

Spain, Switzerland and Turkey.

3.3.7 Anti-discrimination

Anti-discrimination laws are becoming increasingly widespread. Victims of discrimination are often too poorly informed or supported to take the first step in the long path to justice, so most do not report their experience to the authorities. Victims are best informed and supported to seek justice in traditional destination countries (Canada, New Zealand and the US) and some EU Member States (Finland, Portugal and Sweden). Since the adoption of EU law in 2000, anti- discrimination has been the greatest and most consistent area of improvement in integration policy across Europe. Over the past 5 years, 7 countries have made positive reforms to discrimination policy (Croatia, Finland, Iceland, Ireland, Luxemburg, Slovenia and Turkey) and more than half of the MIPEX countries now protect against ethnic, racial, religious and nationality discrimination in all areas of public life [Solano and Huddleston, 2020]. China, India, Japan, Russia and Switzerland are critically behind schedule on these international trends.

3.4 Methodology

The basic finite mixture model assumes that data are drawn from a density modelled as a convex combination of components each of specified parametric form [Green, 2019]. The usage of finite mixture models as clustering procedures comes clear when supposing that the population from which we are sampling is heterogeneous and so there are multiple groups. Model-based clustering refers to the use of statistical models to cluster data, where the (multivariate) observations are assumed to have been generated from a finite mixture of component distributions, each regarded as a cluster, whose specific probability distribution has generated the units belonging to it [Titterington et al., 1985, Hennig et al., 2015]. Model-based clustering offers the advantage of clearly stating the assumptions behind the clustering algorithm, and allows the analysis benefit from the inferential framework of statistics to address some of the practical questions arising when performing clustering: determine the number of clusters, detecting and treating outliers, assessing uncertainty [Bouveyron et al., 2019a]. In our case, we deal with longitudinal data; model-based clustering of such data is far from simple. Indeed, longitudinal data, sometimes referred to as panel data, track the same sample taking measurements at different time occasions. They are very different from time series: in the longitudinal case we observe short sequences of data in correspondence to a large number of individuals or statistical

units, whereas in the time series case we observe long sequences of data referred to one or few statistical units [Bartolucci et al., 2012]. The ideal way to model these data would be to take into account the temporal evolution and models all the responses at the same time. Thus, the analysis will exhibit typical temporal evolution behaviours, which are the objects that researchers in human and social sciences wish to study.

In this paper, we adopt a clustering approach to longitudinal data that consists of arranging the data in a three-way format and modelling them through a matrix-variate mixture model. This approach offers the advantage of accounting for the overall time-behavior, grouping together the units that have a similar pattern across and within time. While not being new [Basford and McLachlan, 1985], matrix-variate distributions have recently gained attention, and Mixtures of Matrix-Normals (MMN) have been developed and applied both in a frequentist framework [Viroli, 2011b] and within a Bayesian one [Viroli, 2011a]. From a frequentist point of view, these models represent a natural extension of the multivariate normal mixtures to account for temporal (or even spatial) dependencies, and have the advantage of being also relatively easy to estimate by means of EM algorithm (a nice short description of the EM application to MMN is provided in Wang and Melnykov [2020]). Very recently, Tomarchio et al. [2022] applied MMN to cluster longitudinal students' career indicators for Italian universities.

3.4.1 Mixture of Matrix-Normals

MMN, as introduced in Viroli [2011b], can be a useful tool to cluster time-dependent data. Suppose we observe N independent and identically distributed random matrices $Y_1, \ldots, Y_N$ of dimension $J \times T$, with J-variate vector observations measured repeatedly over T time points (i.e. $Y \in \mathbb{R}^{J \times T}$), as in a longitudinal study case. Assume that Y follows a matrix-normal distribution, $Y \sim \mathcal{MN}_{(J \times T)}(M, \Phi, \Omega)$, where $M \in \mathbb{R}^{J \times T}$ is the matrix of means, $\Phi \in \mathbb{R}^{T \times T}$ is a covariance matrix containing the variances and covariances between the T occasions or times and $\Omega \in \mathbb{R}^{J \times J}$ is the covariance matrix containing the variance and covariances of the J variables. The matrix-normal probability density function (pdf) is:

$$f(Y \mid M, \Phi, \Omega) = (2\pi)^{-\frac{TJ}{2}} \mid \Phi \mid^{-\frac{J}{2}} \mid \Omega \mid^{-\frac{T}{2}} \exp\left\{-\frac{1}{2}\text{tr}[\Omega^{-1}(Y - M)\Phi^{-1}(Y - M)^{\intercal}]\right\} \tag{3.1}$$

The matrix-normal distribution represents a natural extension of the multivariate normal distribution, since if $Y \sim \mathcal{MN}_{(J \times T)}(M, \Phi, \Omega)$, then $\text{vec}(Y) \sim \mathcal{MVN}_{TJ}(\text{vec}(M), \Phi \otimes \Omega)$, where vec(.) is the vectorization operator and $\otimes$ denotes the Kronecker product. Then, the mean and

the variance of the matrix-normal distribution are:

$$\mathbb{E}(\text{vec}(Y) \mid M, \Phi, \Omega) = \text{vec}(M) \quad , \quad \mathbb{V}(\text{vec}(Y) \mid M, \Phi, \Omega) = \Phi \otimes \Omega.$$

Being a particular specification of the multivariate normal distribution, the matrix-normal distribution shares the same various properties, like for instance, closure under marginalization, conditioning and linear transformations [Gupta and Nagar, 1999]. The separability condition of the covariance matrix has the twofold advantage of allowing the modeling of the temporal pattern of interest directly on the covariance matrix Φ and of representing a more parsimonious solution than that of the unrestricted $\Phi \otimes \Omega$. The pdf of the MMN model is:

$$f(Y \mid \boldsymbol{\pi}, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \phi^{(J \times T)}(Y \mid M_k, \Phi_k, \Omega_k) \tag{3.2}$$

where K is the number of mixture components, $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$ is the vector of mixing proportions, subject to constraint $\sum_{k=1}^K \pi_k = 1$ and $\boldsymbol{\Theta} = \{\Theta_k\}_{k=1}^K$ is the set of component-specific parameters with $\Theta_k = \{M_k, \Phi_k, \Omega_k\}$.

Matrix-variate models suffer from over-parametrization that leads to estimation issues. This issue is addressed in Sarkar et al. [2020] and Zhu et al. [2022], with the aim to explain the data with as few parameters as possible. To do so, the spectral decomposition of the covariance matrix [Banfield and Raftery, 1993, Celeux and Govaert, 1995] is used. The spectral decomposition of the general covariance matrix Ω_k is given by $\Omega_k = \lambda_k \Gamma_k \Delta_k \Gamma_k^\intercal$, where $\lambda_k = \mid \Omega_k \mid^{1/J}$, Γ_k is the matrix consisting of the eigenvectors of Ω_k and Δ_k is the diagonal matrix composed by the eigenvalues. From a geometrical interpretation point of view, λ_k mirrors the volume of the k-th mixture component, Γ_k the orientation and Δ_k the shape. In MMN, there are two covariance matrices, one measuring covariance in time and one among variables. For identifiability issues of the model, the determinant of the time-covariance matrix must be restricted to be $\mid \Phi_k \mid = 1$, hence imposing K restrictions and making $\lambda_k = 1$ for the matrix Φ_k. Moreover, two kinds of mean matrices M are considered: a general (no constraints) and an additive one. An additive matrix M_k has the structure $M_k = \alpha_k \mathbf{1}_T^\intercal + \mathbf{1}_J \beta_k^\intercal$, where $\mathbf{1}_T$ represents a T-dimensional vector of 1s, α_k is the J-dimensional mean vector for the variables (row-wise) and β_k is the T-dimensional mean vector across time (column-wise). This structure gives rise to identifiability issues, which are resolved by imposing K constraints $\beta_{k,T} = 0$. Last, as introduced in Mcnicholas and Murphy [2010], the time-covariance matrix can be further decomposed through the modified Cholesky

decomposition to parameters interpretable in an Auto-Regressive (AR) fashion. Any or all among volume, shape or orientation can be constrained across mixture components. Following the conventional notation in Bouveyron et al. [2019a], for the covariance matrices parameterizations E stands for equal, V denotes variable, I represents identity, configuring different types of constraints that can be imposed. Since Ω_k can be decomposed in 3 submatrices, and Φ_k in 2, we have 14 different possible combination for the former and 8 (including AR) for the latter, giving rise to $14 \times 8 = 112$ different parametrizations. Since the mean matrix M_k can be in turn parametrized with a general or an additive structure, in total we can fit $2 \times 112 = 224$ differently parametrized models.

3.5 Analysis and results

Data used are freely downloadable from the Migrant Integration Policy Index website[7]. For sake of brevity, during the analysis and in all the Tables and Figures, we name the indicators using one-word labels or the codes reported in Section 3.3.

The analysis has been carried out by considering 7 MIPEX dimensions explained in Section 3.3. In this paper, we deal with a three-way "time data array" of the type "units × variables × times" [D'Urso, 2000] that can be algebraically formalised as follows:

$$\mathbf{Y} \equiv \left\{ y_{ijt} : i = 1, \ldots, N; j = 1, \ldots, J; t = 1, \ldots, T \right\} \tag{3.3}$$

where the indices i, j and t stand, respectively, for the units, the quantitative variables and the times. In this paper, $i = 1, 2, \ldots, 52$ indicates the generic country, $j = 1, 2, \ldots, 7$ the generic MIPEX dimensional indicator and $t = 204, 2015, \ldots, 2019$ the generic year; consequently, y_{ijt} represents the determination of the j-th indicator in the i-th country at the t-th year. The first step is to give a geometrical representation of the initial data array $\mathbf{Y}$ to obtain information on the form of the data and the relationships between the basic indicators [Pearson, 1956]. Figure 3.1 outlines that the trajectories of most of the indicators appears quite flat, which means that most of the countries does not change much the values of their indicators (and so the related policies) over time. For instance, Canada, India, Indonesia, Mexico and Romania have no improvement or worsening in any indicator during the considered period; while other countries (for instance, Albania, Austria, Hungary, Italy and Latvia) have just a small change in only

[7] https://www.mipex.eu/download-pdf.

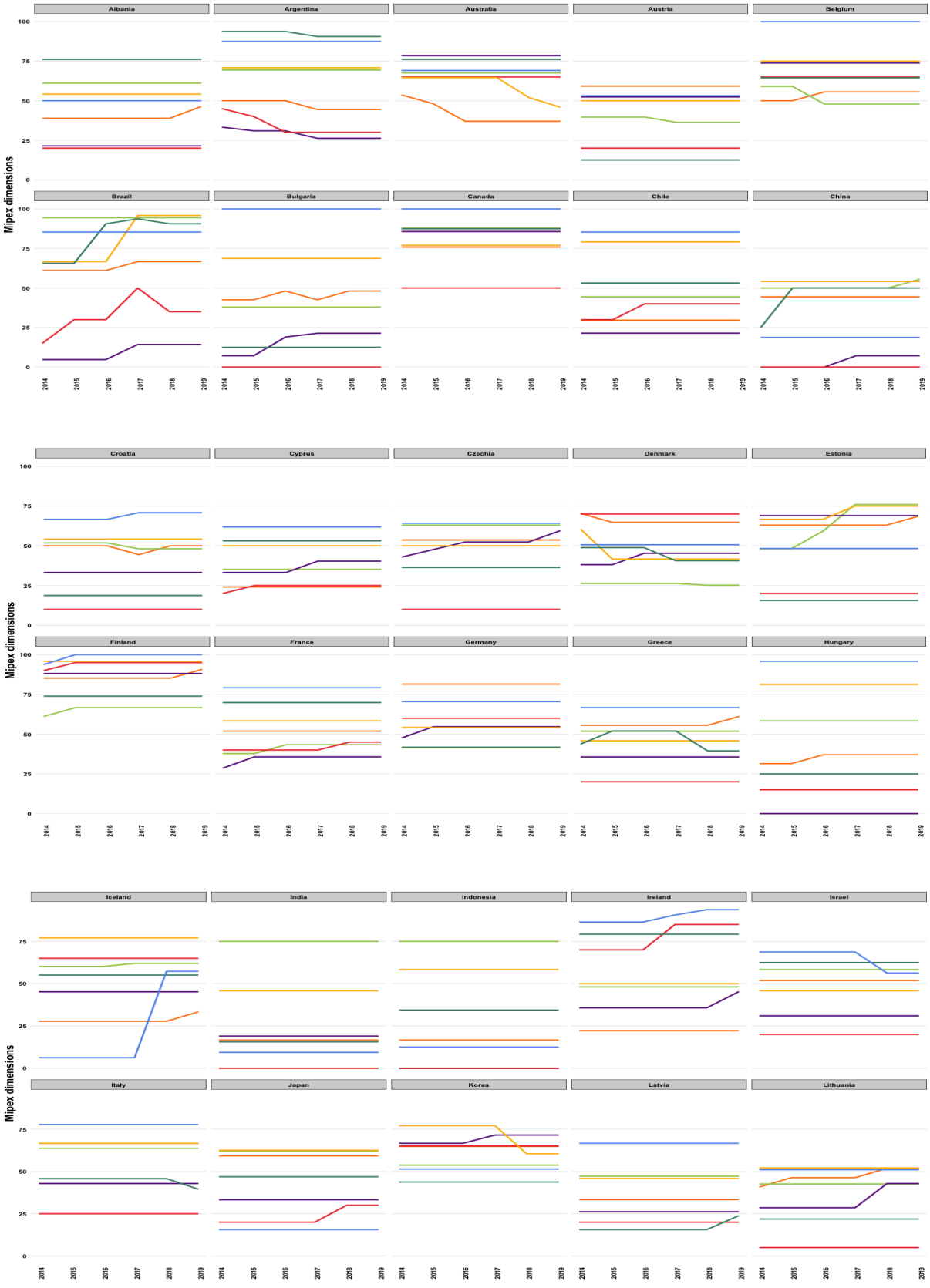
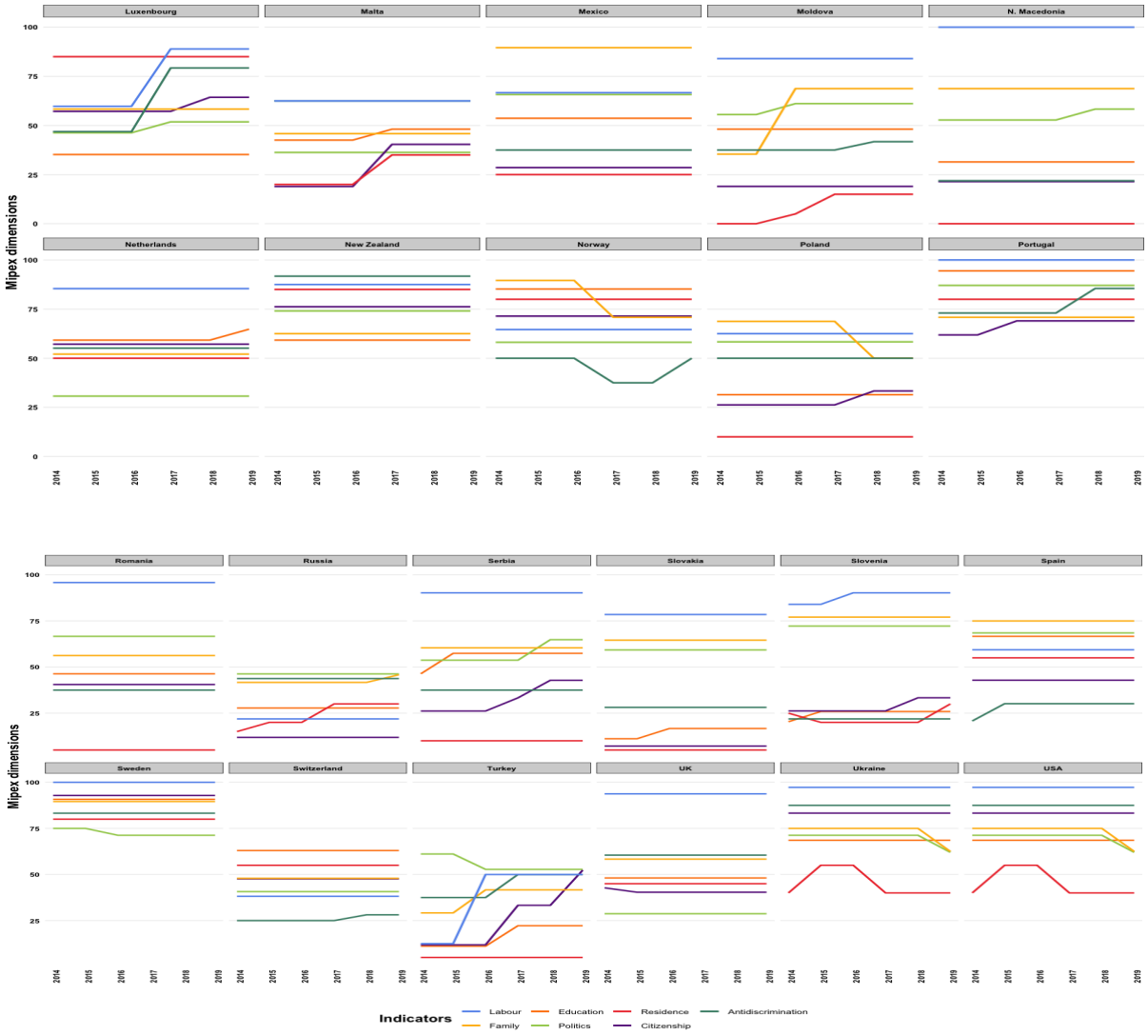

Albania
Argentina
Australia
Austria
Belgium
Brazil
Bulgaria
Canada
Chile
China
Croatia
Cyprus
Czechia
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Iceland
India
Indonesia
Ireland
Israel
Italy
Japan
Korea
Latvia
Lithuania
Mipex dimensions
0
25
50
75
100
2014
2015
2016
2017
2018
2019

Figure 3.1. Country trajectories of the 7 MIPEX dimensions. 52 countries; years 2014 – 2019.



one of the considered years. We can also observe that in most of the countries (for instance, Belgium, Bulgaria, Canada, and so on) the labour dimension is the one that rank higher; at the same time, the residence dimension rank lower. However, this is not true for most of the Asian countries, where the family and politics dimensions tend to rank higher and the labour dimension lower. The MMN will be used to model together the changes between and within time, grouping together the units which behave similarly across and within time.

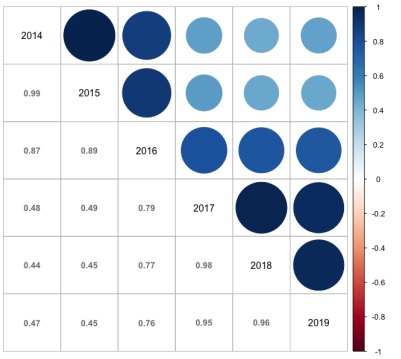
The cluster analysis have been performed with the package MatTransMix [Zhu et al., 2022] of the statistical software R. As usual when performing clustering, the main parameter to set is represented by the number of clusters K. Moreover, it is important that the clusters are interpretable [Fraley and Raftery, 1998, Forgy, 1965]. Since our dataset is composed by 52 units, we carried out the MMN model for K ranging from 1 to 8 and we run the model several times in order to choose the best number of clusters by means of the Bayesian Information Criterion (BIC): the lowest the BIC, the better the model. The selected number of K is 5. The best parametrization of the model, as expressed in Section 3.4.1, is A-VEV-VV[8], which means that the means M_k are better parsimoniously parametrized in additive way, Ω_k with varying volume, equal shape and varying orientation (in a two components case, it would be ellipsoidal with equal shape) and Φ_k has both varying shape and orientation.

Because of the matrices Φ_k and Ω_k, each MMN component models not only the conditional means, but also covariances of the response variables and the covariances among times. This, of course, is visible in the clustering as well, since MMN tends to cluster together not only the units with similar response conditional means, but with conditional covariances among times and variables as well. In this way, each cluster provides a broad profile of units belonging to it. It should be notice that a low correlation in time within cluster means that there have been changes in migration polices in the countries belonging to the cluster; on the other hand, a high correlation in time would signal that little changed. Equally, purified from temporal effect, positive variables correlations mean that the policies' dimensional scores move homogeneously country-wise within cluster. The values of the correlation in time are reported in Figure 3.2, the values of the correlations among variables in Figure 3.3 and the countries that belongs to each cluster in Figure 3.4. The values of the clusters' means over time are reported in Table .1.

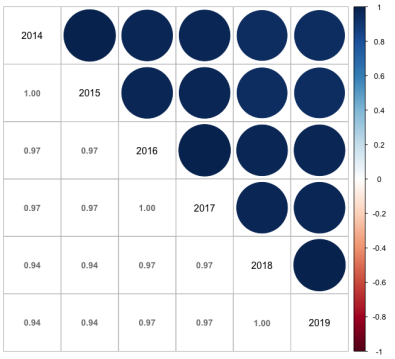
A description and interpretation of the clustering results is as follow:

[8]The total number of estimated parameters is given by $K + (J-1) + KJ(J-1)/2 + KT(T-1)/2 - K + K(J+T-1) = 251$, to be estimated from a total of $J \times T \times N = 7 \times 6 \times 52 = 2184$ observations. For a non parsimoniously parametrized matrix-variate normal mixture the number of parameters would be $K[JT + J(J+1)/2 + T(T+1)/2] - 1 = 454$.

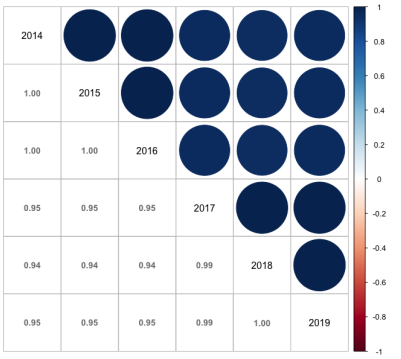
Figure 3.2. MMN clusters' corr-plots in time.



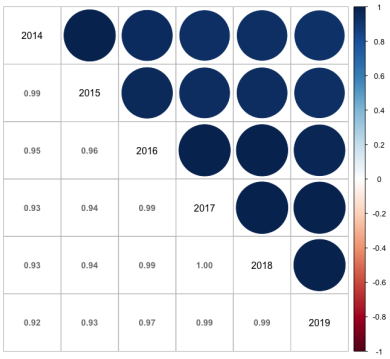
(a) Cluster 1



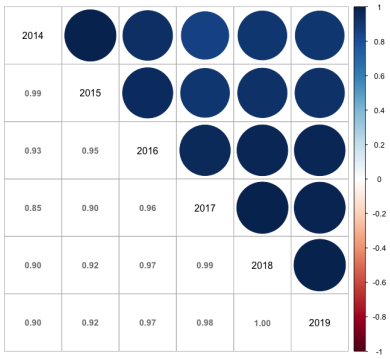
(b) Cluster 2



(c) Cluster 3

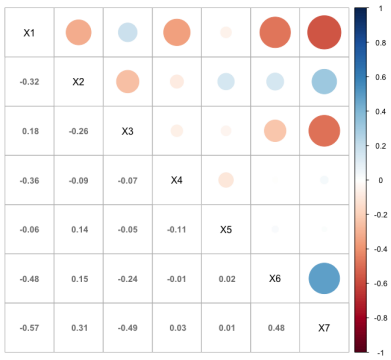


(d) Cluster 4

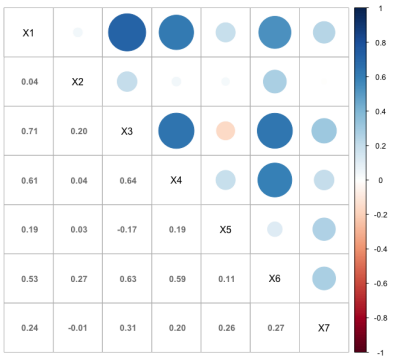


(e) Cluster 5

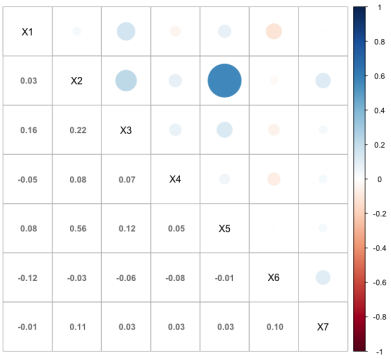
Figure 3.3. MMN clusters' corr-plots among indicators. X1 Labour, X2 Family, X3 Education, X4 Politics, X5 Residence, X6 Citizenship, X7 Anti-discrimination



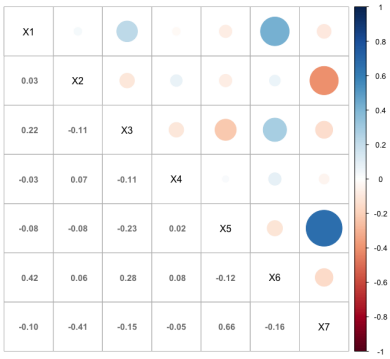
(a) Cluster 1



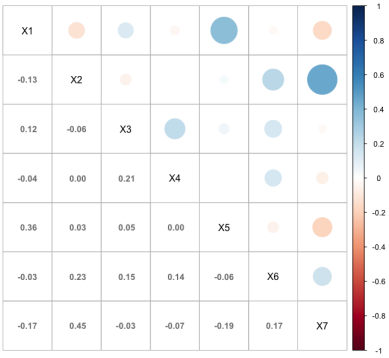
(b) Cluster 2



(c) Cluster 3



(d) Cluster 4



(e) Cluster 5

- **Cluster 1**: Estonia and Slovenia.
  - **Correlation in time**: with respect to the other clusters, Cluster 1 is the one with the lowest correlations within time.
  - **Means**: this is the cluster with the lowest mean values in the Citizenship strand. With respect to the other clusters, it has low values in the Politics indicator but high values for Family, Residence and Anti-discrimination.
  - **Correlation among indicators**: the Labour indicator presents negative correlations with almost all the other indicators except for Family. The correlation is particularly high between the indicators Labour and Anti-discrimination.

  In Cluster 1, we observe relatively low levels of temporal correlation, and this is due to the fact that Estonia has important changes in Family indicator in 2016 and 2017 and Residence in 2017, while Slovenia has important changes in the Anti-discrimination in 2016, in Education in 2018 and Politics in 2019. Cluster 1 is characterized by lower correlations in time between the first 3 years (2014-2016) and the second ones (2017-2019). Moreover, it has negative correlation between Labour Market Mobility and the other dimensions, with the exception of Family Reunion. Countries in this cluster have the lowest score for the Access To Nationality and rank low for Political Participation as well, while ranking high for Family Reunion, Long-term Residence and Anti-discrimination legislation.

- **Cluster 2**: Belgium, Canada, Chile, Hungary, India, Indonesia, Israel, Japan, Mexico, New Zealand, North Macedonia, Poland, Portugal, Romania, Slovakia, Sweden, Switzerland.
  - **Correlation in time**: Cluster 2 presents high correlation values in time.
  - **Means**: with respect to the other clusters, the values of the means of this group are quite low in Politics and Education and high in Family, Residence and Anti-discrimination.
  - **Correlation among indicators**: almost all the indicators of this cluster are positively correlated, with particularly high values between Education and Labour, Politics and Labour, Politics and Education, Citizenship and Education and Citizenship and Politics.

  During the analysed period, countries belonging to this cluster did not change much their policies, and they usually rank high in all the areas. The countries of this group tend

to have good policies for Residence, Family and Anti-discrimination, but rank low for Education and Politics.

- **Cluster 3**: Albania, Austria, China, Croatia, Cyprus, Finland, Germany, Greece, Iceland, Ireland, Italy, Korea, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Russia, Serbia, Spain, Ukraine, UK, USA.
  - **Correlation in time**: Cluster 3 presents the highest correlations in time with respect to the other clusters.
  - **Means**: with respect to the other clusters, this group does not present low mean values for any indicator. It presents medium values in Politics, Labour, Family, Education and Citizenship indicators and quite high values in Residence and Anti-discrimination.
  - **Correlation among indicators**: almost all the correlations values among indicators are low, with exception for Residence and Family.

  The characteristic of Cluster 3 is its high stability in time, that is the tendency to not make huge changes in the legislation, with some remarkable exceptions such as Iceland in Anti-discrimination in 2018 and Citizenship and Anti-discrimination in Luxembourg in 2017. To this cluster, belongs the countries that reformed less their immigration legislation during the study period. They tend to rank average in most of the policies areas, with the exception of Residence and Anti-discrimination laws, where they tend to rank higher. This group could be seen as the “average" cluster, grouping countries which could be located at the middle of the MIPEX overall rank. This does not mean that any country of this cluster do not present high or low values in any indicator, but that overall, among the indicators the tendency is towards the center. However, low correlation among variables signals that countries do not move homogeneously among the policies areas.

- **Cluster 4**: Bulgaria, Czech Republic, France, Turkey.
  - **Correlation in time**: it presents high values but they shades with time.
  - **Means**: with respect to the other clusters, Cluster 4 have the lowest mean values for Politics and quite low values in Education, Citizenship and Labour. It has high mean values in Anti-discrimination.
  - **Correlation among indicators**: it generally presents low correlations with the exception for an high positive value between Anti-discrimination and Residence.

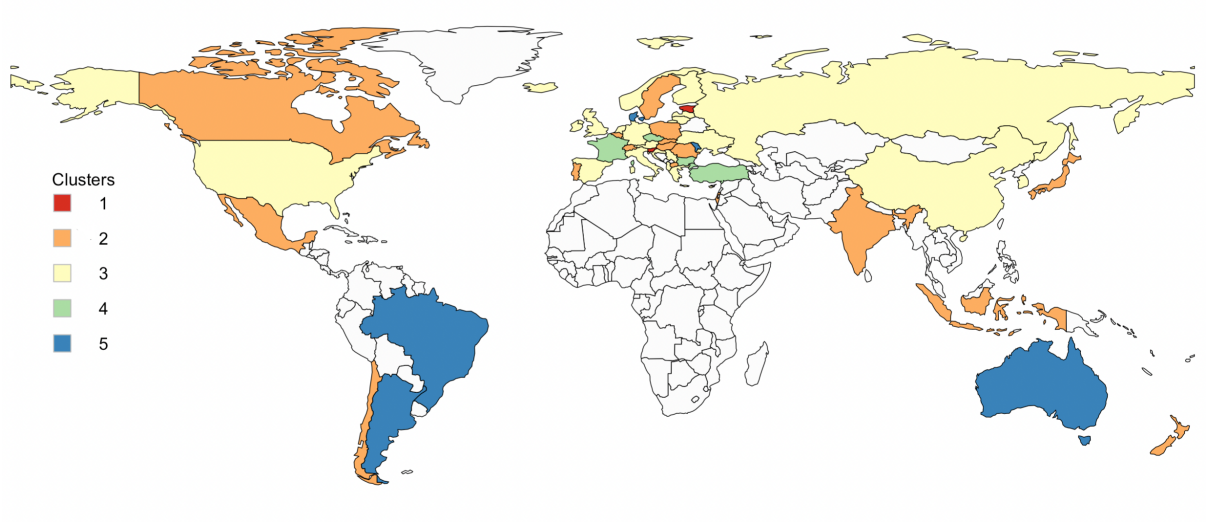
Cluster 4 is mainly characterised by its relatively low values of Politics in every country, including France. Important positive improvements in Education across time for all the countries mostly explaining the time-correlation behaviour. Despite ranking generally high for Anti-discrimination policies, countries within this cluster tend to rank low for policies in Education, Citizenship and Labour, while scoring average for Residence legislation. Yet, low correlation among variables indicates that the countries do not move homogeneously among the dimensions, with the exception of policies regarding Residence and Anti-discrimination, that have high positive correlation. Countries belonging to this cluster have seen their score moderately changing in time, indicating that some changes in the legislation have happened.

- **Cluster 5**: Argentina, Australia, Brazil, Denmark, Moldova.
  - **Correlation in time**: it presents high values but they shade faster.
  - **Means**: with respect to the other clusters, the values of the means of Cluster 5 are quite low in Education and Politics, medium in Labour and high for the other indicators.
  - **Correlation among indicators**: the values of the correlations are generally low.

  Cluster 5 collects countries with smooth evolution, in both positive and negative directions and it generally presents low values in Education (with the exception of Australia). Changes are to be noted in Residence, where all the countries (with the exception of Argentina) see their values change in time (in both directions). Countries belonging to this cluster have high correlation values in time, but they tend to decrease faster with time, meaning that some changes in the policies have been made especially in the last years. Countries of this cluster, are characterized for generally ranking low in policies related to Educational support for foreign pupils and Politics, but high in Family, Residence, Citizenship and Anti-discrimination. However, the low correlation among the dimensions, means that the countries tend not to move homogeneously among them.

Looking at the details of the countries assigned to each cluster, it could be noticed that in the clustering process the algorithm gave more importance to the temporal and variables' dynamics (captured by Φ and Ω) than to their overall scores (captured in M). The clustering privileged the similarity in trajectory rather than in magnitude. This gives us an idea on how

Figure 3.4. MIPEX dimensional indices: MMN clusters' composition. 52 countries; years 2014 – 2019.



the clustering should be read and explains why countries that one could think are quite different in their policies are in the same cluster.

3.6 Conclusions

This paper has explored immigrant regulation and immigrant assimilation policies, analyzing 7 dimensions of the Migrant Integration Policy Index from the year 2014 to 2019. The need for the analysis carried out came from the statement that when comparing very different countries from each other on social and civil issues, the identification of homogeneous groups of units substantially improves the ease of reading and the interpretation of the results. In this paper, we addressed this issue trough the application of an unsupervised clustering approach for longitudinal data namely MMN. The exploration and visualization of the data show that for the 7 MIPEX dimensions analyzed, the considered countries tend to change little over time. This behaviour led us to rely on an approach as MMN, that accounts simultaneously for the within and between time dependency structures. The identification of groups of countries with similar behaviour over time allows the comparison of clusters with each other and the comparison of the countries within each cluster. Moreover, the correlations in time shows the general trend of each indicator over time in each cluster, and the correlations between variables purified from the time effect underline the behaviour of each indicator in relation to the others within each cluster. This analysis allowed the addition of new levels of interpretation of the migration policies and of

several new information about the phenomena. Specifically, the information added helps to better understand which countries have similar legislative attitudes regarding migration policies and which are following similar trends, whether they are virtuous toward integration, static, or toward the marginalization of migrants. For instance, the evidence that Bulgaria and France are both in Cluster 4 highlights that they both have relatively low values for the Politics dimension and they both improved the Citizenship dimension over the considered years.

As future developments of this work, we expect, as the data will be available, to add to the analysis the Health dimension. This would be of particular interest especially during the last years of COVID-19 pandemic. Moreover, if as we expect, there will be changes in the migration policies of many of the countries considered, and, consequently, there will be changes over time in the trajectories of the considered indicators. Moreover, it will be of particular interest to estimate the probabilities to move trough the clusters along the time, through the application of Latent Markov models.

Chapter 4

Immigrants integration and detention in Europe

Abstract

Migrant integration and immigrants' behaviour have been central issues in the political debate for decades. In particular, the perceived link between crime and immigration is one of the hottest topics, which also boasts a wide literature. In this paper, we question whether there exists an association between countries' level of integration of immigrants and the proportion of immigrants in prison. To test the existence of such an association, we cluster 34 European countries for the year 2019, modelling the dimensions of the Migrant Integration Policy Index (MIPEX). Leveraging finite mixtures of multivariate Gaussian, we identify three groups of countries with a similar level of integration. Then, we estimate the relative proportion of immigrants held in prison among clusters, relying on UNODC and UNDESA data and exploiting Fisher's noncentral hypergeometric model. The analysis of the results shows that foreigners are 1.6 times more exposed to detention in that cluster which is less virtuous in terms of migrants' integration than in the others. Moreover, looking at the differences within clusters, we find that foreigners have a different propensity to be held in prison with respect to citizens. The proposed approach adds new valuable information to the MIPEX and provides a novel perspective to picture an important and highly debated phenomenon, such as immigrants in prison, through the lenses of migrants' integration.

4.1 Introduction

Since 2000, the number of international migrants has grown by more than 100 million, with Europe representing the destination area of the largest number of international migrants in the world (87 million in 2020; see UNDESA [2020b]). The intensification of immigrant flows has been a central issue in public debates over the last two decades, shaping people's attitudes towards immigration. The last wave of the Eurobarometer[1] highlights that immigration is among the top concerns of people in all Member States [Standard Eurobarometer, 2022]. Among the reasons for concern, the perceived link between crime and immigrants seems to be pervasive Mastrobuoni and Pinotti [2015], Mayda [2006], Card et al. [2012]. The political debate reflects people's opinions. Recent literature has stressed the role of immigration in the populist and demagogic debate that adopts the strategies of negative other-representation and criminalisation of immigrants: among others, see Greco and Polli [2019] and Combei and Giannetti [2020].

[1]The polling instrument used by EU institutions and agencies to monitor regularly the state of public opinion in Europe `https://europa.eu/eurobarometer/about/eurobarometer`

In particular, the propensity of immigrants to commit crimes inflames the political debate in Europe [Solivetti, 2018]. However, as highlighted in Drazanova et al. [2020], Europeans' attitudes towards immigration are quite heterogeneous among countries.

In this paper, we question whether different attitudes of countries' policies toward migrant integration correspond to different propensities to hold immigrants in prison.

We focus on crimes involving imprisonment as a penalty for data availability. The aim of this work is neither the quantification of the association between migrants' integration and immigrant propensity to commit crimes nor the assessment of a causal relation between these phenomena. Our objective is rather to describe a phenomenon through the lenses of another. We investigate the problem by focusing on European countries. Given the subjects' heterogeneity and the complexity of such a multidimensional phenomenon, we propose a clustering approach. Then, we compare the propensity of immigrants to land in prison among clusters of countries which have similar integration policies. We leverage data from multiple sources. To cluster European countries by their level of integration towards migrants, we rely on MIPEX data. As described in Solano and Huddleston [2020], Alaimo et al. [2021a], the MIPEX is a system of 167 policy indicators. It includes 52 countries and collects data from 2007 to 2019 in order to provide a view of integration policies across a broad range of differing environments. The values of the 167 basic indicators are chosen by experts from each country by means of a questionnaire, and through the arithmetic mean, they are first aggregated in 58 indicators, and then in 8 policy areas[2].The policy areas of integration covered are:

- Labour Market Mobility
- Family Reunion
- Education
- Political Participation
- Long-term Residence
- Access to Nationality
- Anti-discrimination
- Health

Each dimensional synthetic indicator is bounded between 0 and 100: the higher the value, the better the situation in that policy area. A description of MIPEX dimensions is given in Solano and Huddleston [2020] and briefly in Alaimo et al. [2021a]. To the aim of this work, we focus on the year 2019 and European countries. Concerning the migrant stock and the number of persons held in prison, we rely on the United Nations data from different agencies, namely

[2]For more information see: `http://mipex.eu/methodology`.

UNDESA (UN Department of Economic and Social Affairs) and the UNODC (UN Office on Drugs and Crime), respectively. We use 2020 data, which represents the closest year to the MIPEX's last wave.

This work is based on recent intuitions given in two papers presented at the 51^{st} Scientific Meeting of the Italian Statistical Society on June 2022 [Alaimo et al., 2022a, Ballerini and Liseo, 2022]. Indeed, to model the MIPEX dimensions, we rely on a model-based clustering approach, similarly to Alaimo et al. [2022a]; it offers the advantage of clearly stating the assumptions behind the clustering algorithm, and allows the analysis to benefit from the inferential framework of statistics to address some of the practical questions arising when performing clustering, i.e., determining the number of clusters, detecting and treating outliers, assessing uncertainty about which components each unit belongs to [Bouveyron et al., 2019a]. Then, we exploit Fisher's noncentral hypergeometric distribution (FNCH) to model the number of immigrants held in prison; it accounts for the possibility that different clusters have different propensities to hold foreigners in prison. The use of FNCH in social sciences has been proposed in Ballerini and Liseo [2022].

The article is organised as follows. Section 4.2 briefly reviews some of the existing literature on the association between immigrants and criminality and on migrant integration and describes data and their sources in detail (4.2.1). In Section 4.3, the methods used, namely model-based clustering via multivariate Gaussian mixtures (4.3.1) and Fisher's noncentral hypergeometric distribution (4.3.2), are briefly described. The analysis and the results are described in Section 4.4. Conclusions follow.

4.2 Migrant integration and crime

Countries' policies to regulate immigrants' integration affect their standard of living, their level of inclusion, and their ability to remain in the destination country Helbling et al. [2020], Solano and Huddleston [2020], Solano and De Coninck [2022]. The association between immigration and crimes is a long-studied issue among researchers. However, the results and their interpretations are mixed. In fact, social phenomena related to immigration are complex and change by the considered time and place. Thus, the association with crime is far from trivial, as the results could be influenced by many other unobserved factors. To conceptualise this association, as assessed in Solivetti [2018], some authors have adopted Merton's thesis within the "anomie" conceptual framework [Merton, 1938, and subsequent versions and revisions] that high social pressure to

succeed materially in the face of scarce legitimate opportunities leads to crime and other forms of deviance. Other authors have supported the so-called economic model of crime, which, following Becker's study Becker [1968], assumes that crime is a rational option whenever its benefit outweighs its cost. Crime costs and benefits, in turn, are influenced by economic conditions, which affect both legitimate opportunities and returns to crime. According to Bianchi et al. [2012], from a theoretical viewpoint, there are several reasons to expect a significant relationship between immigration and crime. This may happen because immigrants and natives face different legitimate earning opportunities, different probabilities of being convicted and different costs of conviction [Becker, 1968, Ehrlich, 1973]. For instance, LaLonde and Topel [1991] and Borjas [2000] document that immigrants in the United States experience worse labour market conditions, which would predict a higher crime propensity. Also, immigration may affect crime rates as a result of natives' response to the inflows of immigrants Borjas et al. [2010]. In Europe, Boateng et al. [2021] analyse aggregate-level data obtained from 21 European countries to assess the effects of immigrants on three different types of violent crimes. Their results indicated a null relationship between immigration and crime, suggesting that immigration is unrelated to all the three types of crimes assessed. Focusing on Italian provinces in a cross-sectional setting, Solivetti [2018] find that crime intensities are affected by time-invariant factors and marginally by immigration. On the contrary, the longitudinal analysis shows that variations in immigration had a positive impact on both the most serious and the most common offences, on property crimes as well as on crimes of violence. There is no evidence of indirect effects of immigration on crime or of a link with the native crime. Both Merton's anomie and the economic approach to crime place an association between immigration and utilitarian crimes, though an association also with non-utilitarian crimes mediated by frustration has been hypothesised regarding the anomie Blau and Blau [1982], Bjerregaard and Cochran [2008]. Hence, it is reasonable to investigate whether immigrants' propensity to criminality is different in countries with different levels of migrant integration. Throughout this paper, we focus on foreigners held in prison rather than "criminal foreigners"; we discuss the imperfect overlap of such populations in Section 4.2.1.

In this paper, we refer to the definition of migrant integration given in Solano and De Coninck [2022], which state that migrant integration refers to the process of settlement, interactions with the receiving society and social change due to immigration [Penninx, 2019, Garcés-Mascareñas and Penninx, 2016, Entzinger, 2000]. Integration policies relate to the conditions required to become and to remain part of a specific society and the entitlement rights as well as the support migrants receive [Garcés-Mascareñas and Penninx, 2016, Entzinger, 2000]. An overview of many

of the existing indexes and indicators on migration policies and their methods is given in Solano and Huddleston [2022].

4.2.1 Data sources

Migrant integration MIPEX data used are freely downloadable from the Migrant Integration Policy Index website[3]. We focus on 2019 data. For the sake of brevity, during the analysis and in all the Tables and Figures, we name the indicators using one-word labels, according to Alaimo et al. [2021a] notation.

Migrant stocks Data on international migrant stocks (at mid-year) by country of destination are made available at 5-years intervals by the United Nations Department of Economic and Social Affairs, Population Division (UNDESA) UNDESA [2020a]. The last available data are for the year 2020. We refer to these data, which are also those temporally closer to MIPEX 2019.

Immigrants in prison The United Nations Office on Drugs and Crime (UNODC) collects data on access and functioning of justice, including persons held in prison, on a yearly base. National data are usually submitted by Member States to UNODC through the United Nations Survey of Crime Trends and Operations of Criminal Justice Systems (UN-CTS). UNODC labels as “persons held in prison" all persons held in prisons, penal institutions or correctional institutions; non-criminal prisoners held for administrative purposes should be excluded. Data can be detailed by citizenship (citizens/foreigners). Although “immigrants in prison" may be a subpopulation of all “foreigners in prison" that may include travellers, we assume such overcoverage to be negligible. Notice that our target population is the one of foreigners in prison, which does not perfectly overlap with that of “criminal foreigners". Indeed, the latter also includes foreigners that commit crimes not leading to imprisonment, and because the unsentenced detainees might not be guilty. For the sake of coherence with migrant stocks, we refer to 2020 data. For a few countries[4], UNODC data for 2020 are not available. In such cases, we rely on data collected for the World Prison Brief (WPB) by the International Centre of Prison Studies in London WPB [2020], which are proven to be coherent with the UNODC data for those countries whose data are available on both sources; UNODC itself includes the WPB among its sources.

[3] https://www.mipex.eu/download-pdf.

[4] Cyprus, Germany, Ireland, Netherlands, United Kingdom.

4.3 Model setting

As assessed in the Introduction, our aim is to assess whether there exists a significant association between migrants' integration, measured using the MIPEX, and the number of immigrants held in prison. To this aim, we first cluster countries according to MIPEX dimensions; then, we estimate the relative exposure to imprisonment relying on Fisher's noncentral hypergeometric (FNCH) model, which arises naturally in this context. The next subsections briefly describe the methods.

4.3.1 Clustering via finite mixture of multivariate normal

Let $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a matrix of N multivariate observations independent and identically distributed (i.i.d.), each of dimension J, so that $\mathbf{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{iJ}\}$, with $i = 1, \ldots, N$. A finite mixture model represents the probability distribution or density function of one multivariate observation, $\mathbf{x}_i$, as a finite mixture or weighted average of G probability density functions, called *mixture components* [Bouveyron et al., 2019a]:

$$f(\mathbf{x}_i) = \sum_{g=1}^{G} \tau_g f_g(\mathbf{x}_i \mid \theta_g) \tag{4.1}$$

Where τ_g is the probability that an observation was generated by the g-th component, under the constraints that $\tau_g \geq 0$ for $g = 1, \ldots, G$, and $\sum_{g=1}^{G} \tau_g = 1$, while $f_g(\cdot \mid \theta_g)$ is the density of the g-th component given the values of its parameters θ_g.

We consider the case in which each component arises from a multivariate normal distribution $\mathbf{x}_i \mid \theta_g \sim f_g(\mathbf{x}_i \mid \theta_g) = MVN(\mu_g, \mathbf{\Sigma}_g)$, and has the form:

$$f_g(\mathbf{x}_i \mid \mu_g, \mathbf{\Sigma}_g) = \frac{1}{\mid 2\pi\mathbf{\Sigma}_g \mid^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mu_g)'\mathbf{\Sigma}_g^{-1}(\mathbf{x}_i - \mu_g)\right\} \tag{4.2}$$

where μ_g is the mean vector and $\mathbf{\Sigma}_i$ the covariance matrix. Model parameters are estimated by using the iterative Expectation-Maximization (EM) algorithm [Dempster et al., 1977].

Clustering via finite mixtures of multivariate normal [McLachlan and Basford, 1988, Fraley and Raftery, 2002, Melnykov et al., 2010, Fraley et al., 2007] allows to classify each observation x_{ij}, with $j = 1, \ldots, J$, into one of the G groups by computing the posterior probabilities. In the multivariate setting, the covariances' volume, shape, and orientation can be constrained to be equal or variable across groups. Thus, a parsimonious version of the model is considered, with

14 possible models with different geometric characteristics [Scrucca et al., 2016]. The optimal model and the optimal number of clusters are chosen according to the Bayesian information criterion (BIC).

4.3.2 Fisher's noncentral hypergeometric distribution

Let $\mathbf{Y} = (Y_1, \ldots, Y_N)$ be a N-dimensional vector of independent Binomial random variables, each of parameters M_i, π_i, $i = 1, \ldots, N$. Conditional on the observed sum of its elements, $\mathbf{Y}$ follows a Fisher's noncentral hypergeometric distribution (FNCH):

$$\mathbf{Y} \mid \sum_{i=1}^{N} Y_i = n \sim \text{FNCH}(\mathbf{M}, n, \mathbf{w}) \tag{4.3}$$

where $\mathbf{M} = (M_1, \ldots, M_N)$, and $\mathbf{w}$ is the $(N-1)$-dimensional vector of the odds ratios, the i^{th} element being

$$w_i = \frac{\pi_i/(1-\pi_i)}{\pi_{i^*}/(1-\pi_{i^*})} \quad \forall i \neq i^*, \tag{4.4}$$

with i^* indicating the reference category such that $w_{i^*} = 1$. FNCH has been underused in the statistical literature mainly because of the computational burden given by its probability mass function:

$$P\left(\mathbf{Y} = \mathbf{y} \mid \sum_{i=1}^{N} Y_i = n\right) = \frac{\prod_{i=1}^{N} \binom{M_i}{y_i} {w_i}^{y_i}}{\sum_{\mathbf{z} \in \mathcal{Z}} \prod_{i=1}^{N} \binom{M_i}{z_i} {w_i}^{z_i}} \tag{4.5}$$

where $\mathcal{Z} = \left\{ \mathbf{y} \in {\mathbb{N}_0}^N : \left[\sum_{i=1}^{N} y_i = n \right] \cap \left[0 \leq y_i \leq M_i \right], \forall i \right\}$.

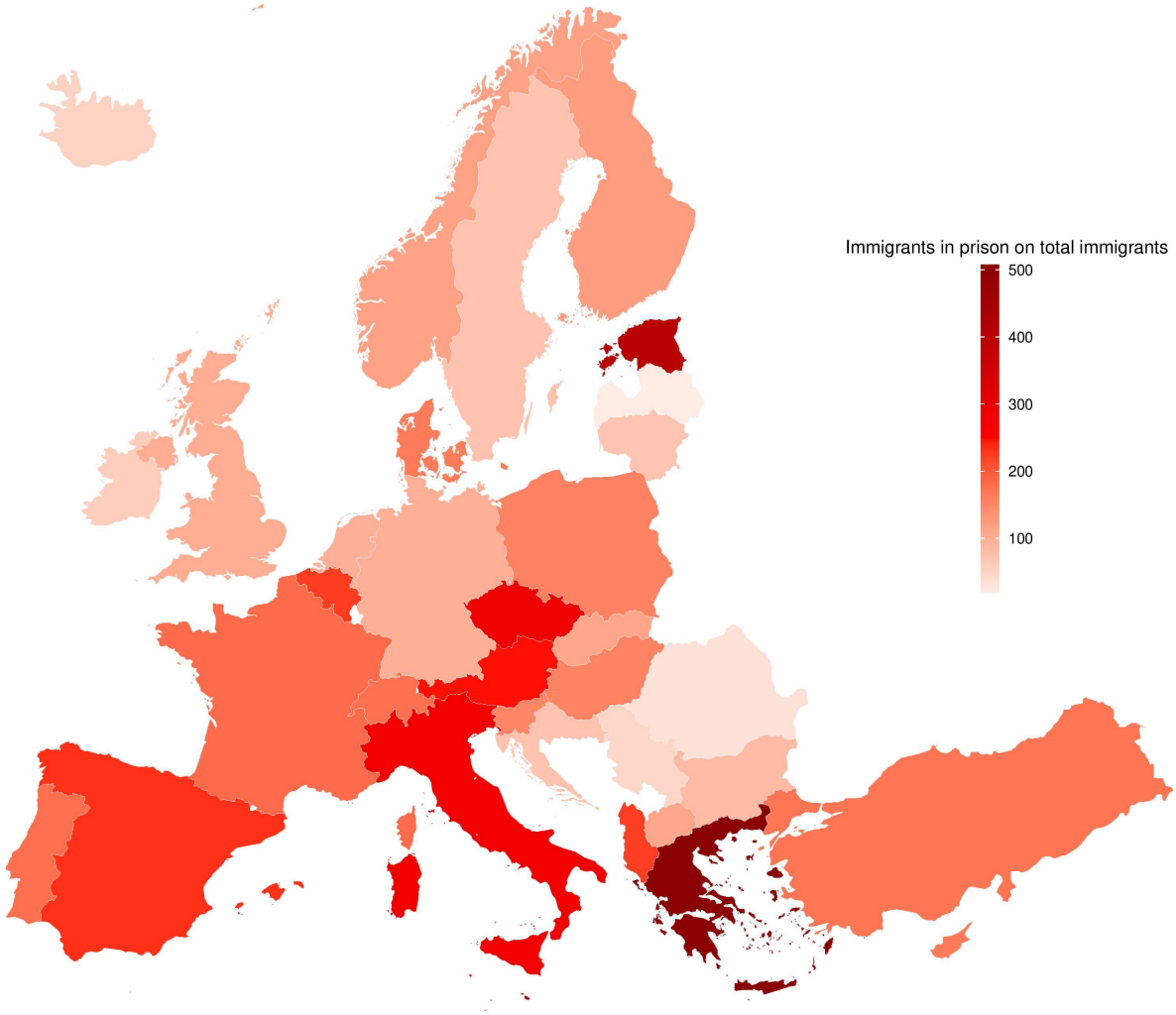
The sum at the denominator makes unfeasible the derivation of an MLE for w_i in closed form; numerical approximation methods are provided by [Fog, 2008]. In a Bayesian perspective, Ballerini and Liseo [2022] provides Markov Chain Monte Carlo (MCMC) methods to derive the posterior probability in the univariate case, also dealing with both weight w and size $\mathbf{M}$ parameters unknown, in the presence of informative prior information.

4.4 Analysis and results

As a first step of the analysis, we visualise the distribution of foreigners held in prison in Europe. Figure 4.1 shows the number of foreign persons held in prison over the total number of immigrants in European countries in 2020. Northern and Eastern Europe have the

lowest propensity to be held in prison for immigrants (less than 1 foreigner in prison over 1000 immigrants). On the other hand, we observe the highest rates in the Mediterranean countries, which usually are the first recipients of the immigration waves. It is possible to spot some high rates in Belgium, Austria, Czechia, and Estonia.

Figure 4.1. Foreigners held in prison per hundred thousand immigrants, by country.

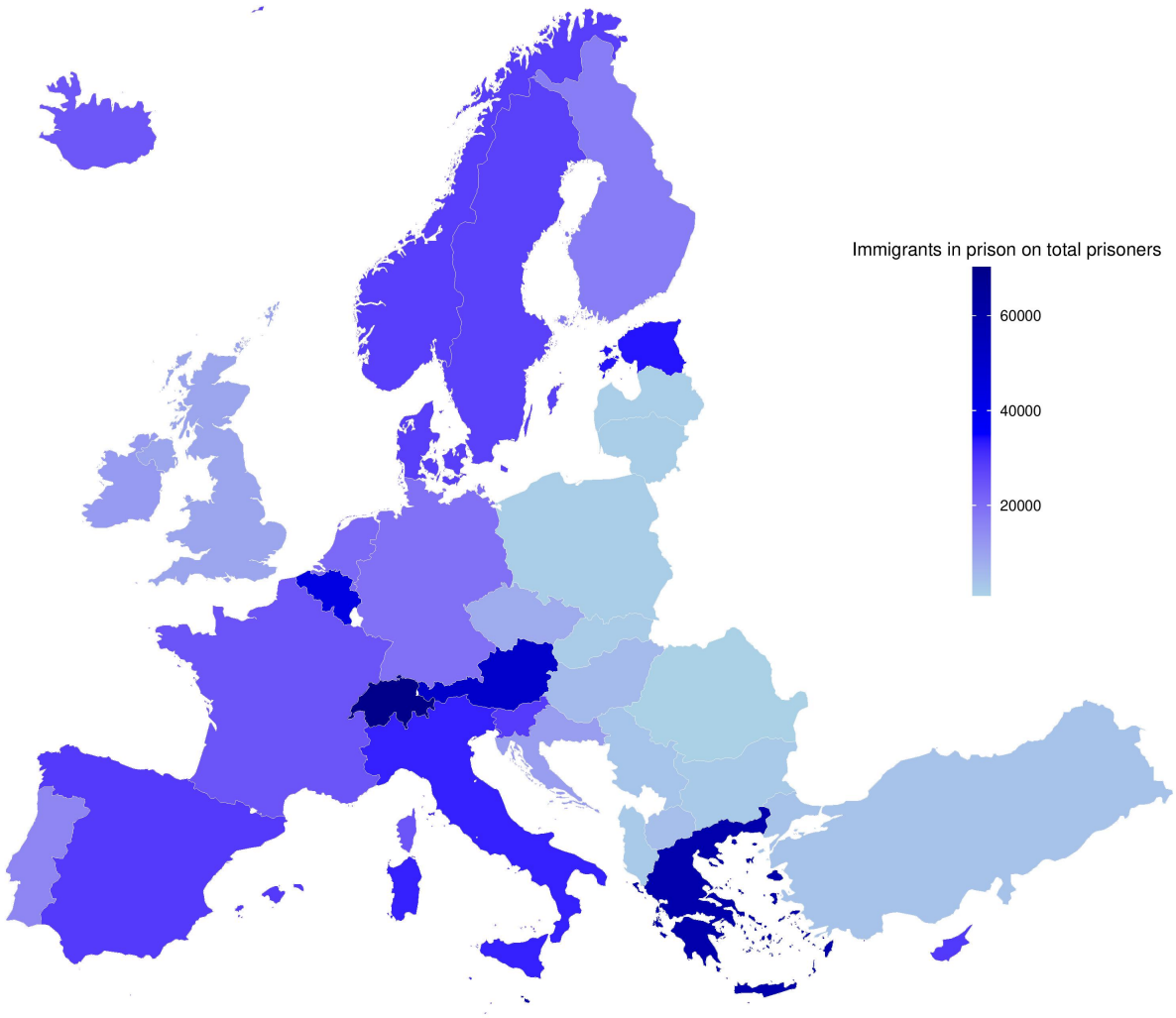


Sources: UNDESA, UNODC, WPB.

Figure 4.2 shows the number of foreigners held in prison over the total of prisoners. Eastern Europe records the lowest incidence of foreigners held in prisons; Albania, Bulgaria, Latvia, Lithuania, North Macedonia, Poland, Romania, Serbia, Slovakia and Turkey are all under 5%.

The exceptions are represented by Estonia (33.8%) and Cyprus (29.1%). On the other hand, Western countries show higher incidence, with peaks in Switzerland (70.2%), Greece (59.8%), and Austria (50.1%).

Figure 4.2. Foreigners held in prison per hundred thousand prisoners, by country.



Sources: UNDESA, UNODC, WPB.

4.4.1 Integration clusters

To cluster MIPEX dimensions data, we rely on a *model based clustering* approach based on parameterised finite Gaussian mixture models for its ability to approximate the density function

of any unknown distribution Titterington et al. [1985], Li and Barron [1999], computed via the package `Mclust` Scrucca et al. [2016] of the `R` statistical software[5]. It should be noted that in a longitudinal setting, MIPEX dimensions have already been clustered in Alaimo et al. [2021a], while in a cross-sectional setting, an attempt to cluster the MIPEX data for European countries has already been made in Hooghe and Reeskens [2009], even if the implemented model is not specified. We compare 14 models with different geometric characteristics of the covariances[6]; each model is applied for a different number of components, $1 \leq G \leq 9$. By means of the BIC, the selected model parametrisation is EVI (diagonal, equal volume, varying shape) with 3 components.

The cluster means are shown in Table 4.1.

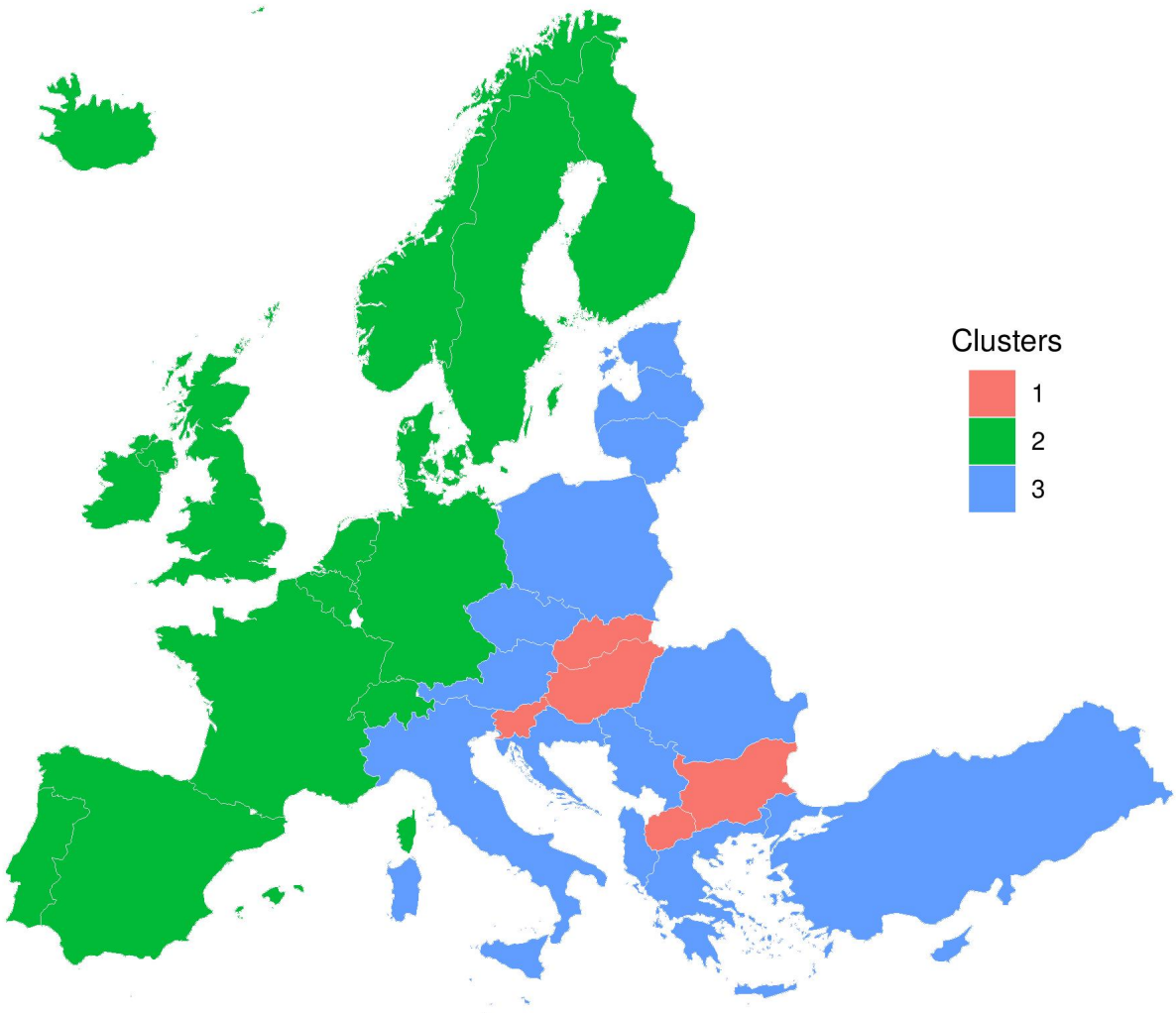
Table 4.1. Cluster parametrisation and means

Parametrisation - EVI	Cluster 1	Cluster 2	Cluster 3
Labour	3.40	4.11	3.82
Family	4.03	3.88	3.96
Education	1.86	4.00	3.70
Politics	0.62	4.14	2.57
Residence	4.27	4.15	3.96
Citizenship	3.05	4.01	3.51
Antidiscrimination	4.53	4.31	4.15
Health	3.56	4.24	3.69

[5]The data have been transformed to the log scale, to deal with variables defined in the real domain.

[6]As specified in Scrucca et al. [2016], in the multivariate setting, the volume, shape, and orientation of the covariances can be constrained to be equal (E) or variable (V) across groups. Thus, 14 possible models with different geometric characteristics can be specified.

Figure 4.3. Countries by cluster membership.



- **Cluster 1**: Slovakia, Slovenia, Hungary, North Macedonia, and Bulgaria.
  - It is characterized by very low scores in the integration-policy areas of education and political participation; moreover, quite low values characterize the citizenship and labour strands. However, Cluster 1 performs better than the others in the family reunions, long-term residence, and anti-discrimination areas. It must be noticed that the security of permanent residence may be a fundamental step on the path to full citizenship and better integration outcomes, according to Solano and Huddleston

[2020]. Hence, in the case of a significant association between integration and crime, the ease of long-term residency would likely represent an important determinant.

- **Cluster 2**: Portugal, Spain, France, Belgium, Netherlands, Germany, Switzerland, Ireland, United Kingdom, Iceland, Norway, Sweden, Finland, Denmark.
  - Overall, it is the most virtuous cluster toward migrant integration. It presents higher means than the other 2 clusters in the areas of labour market mobility, education, political participation, access to nationality and health, and a lower mean in the family reunion strand, in which only Portugal and Sweden present particularly high values. Countries that belong to this cluster are those that evaluate immigrants' impact on society most positively [Drazanova et al., 2020].
- **Cluster 3**: Estonia, Lithuania, Latvia, Poland, Czechia, Austria, Serbia, Romania, Italy, Malta, Croatia, Albania, Greece, Cyprus, and Turkey.
  - Countries in this cluster are characterized by generally quite low values in all the dimensions. The only particularly high values are those of Austria and Italy in the health area and of Serbia and Romania in the antidiscrimination area.

These results are perfectly in line with the results in Drazanova et al. [2020] concerning the attitude towards immigration in the different European countries. Indeed, countries in Cluster 2 also have the most positive attitude towards immigration in terms of the degree of belief that immigration has an overall positive impact on society. On the contrary, countries in Clusters 1 and 3 have a mostly negative perception.

4.4.2 Propensity to commit crimes

Let $\mathbf{M} = (M_1, \dots, M_N)$ be the vector of the number of individuals residing in countries $i = 1, \dots, N$. We denote with M_i^{c} and M_i^{f}, $i = 1, \dots, N$, the number of citizens and foreigners, respectively, summing to M_i. For each i, let Y_i^h be the number of persons held in prison; among them, Y_i^{c} will be citizens, and Y_i^{f} foreigners. We assume

$$Y_i^h \overset{ind}{\sim} \mathrm{Binom}(M_i^h, \pi_i^h) \quad h = \mathrm{c}, \mathrm{f}\,, \tag{4.6}$$

where π_i^h is the probability to be held in prison in country i for each status h. If we broadly assume that the population of those who commit crimes that involves imprisonment as a penalty

coincides with the population of individuals who are in prison, then we can see π_i^{c} and π_i^{f} as the propensity of citizens and foreigner people to commit such major crimes.

Once countries are clustered, we would be able to estimate the clusters' propensity to commit crimes as $\hat{\pi}_g^h = y_g^h / M_g^h$, $g = 1, \ldots, G$ being the cluster label and

$$y_g^h = \sum_{i \in g} y_i^h \quad M_g^h = \sum_{i \in g} M_i^h. \tag{4.7}$$

However, we should be interested in the relative propensity - in the form of odds ratios, for two reasons.

First, for the sake of interpretability; second, the odds ratio would not be affected by the presence of undercoverage in the population stocks if we assume the undercoverage rates to be constant over the clusters. For instance, consider the possibility that M_g^{f} is undercovered, i.e., that cluster g misses some units when counting the number of immigrants; above all, it may be due to illegal immigration. In such a case, we would observe a portion k_g^{f}, $k_g^{\mathrm{f}} \in (0, 1]$ of M_g^{f}. If we assume $k_i^{\mathrm{f}} = k^f$, the odds ratios remain constant whatever the level of k^f; that does not hold for the absolute propensity. Such an assumption is plausible; Mastrobuoni and Pinotti [2015] reports that in many developed countries the incidence of irregular immigrants is approximately 25% of the total immigrants (see also González-Enríquez [2009]).

Fisher's noncentral hypergeometric distribution arises naturally in this context. Indeed, the distribution of the number of people of type h held in prison in the different clusters, conditional on the observed sum n^h, will be

$$Y_1^h, \ldots, Y_G^h \mid \sum_{g=1}^{G} Y_g^h = n^h \sim \mathrm{FNCH}(\{M_1^h, \ldots, M_G^h\}, n^h, \{w_1^h, \ldots, w_G^h\}) \,. \tag{4.8}$$

Between clusters

Cluster analysis identified three clusters; we aggregate migrant stocks and count data of foreigners held in prison accordingly, as in (4.7). We provide estimates for the odds ratios w_g^{f}'s in (4.8) via numerical approximation, using the `R` package `BiasedUrn` [Fog, 2015]. Tables 4.2-4.4 show the estimates and their 95% confidence intervals, for $g^* = 1$, $g^* = 2$, $g^* = 3$, respectively.

Table 4.2. Point estimates and confidence intervals of the odds ratios comparing foreigners' propensity to commit crimes among clusters. Reference cluster $g^* = 1$.

	$\hat{w}_g^{\mathrm{f}}$	95% CI
Cluster 1	1	
Cluster 2	1.05	[0.998;1.096]
Cluster 3	1.67	[1.621;1.780]

Table 4.3. Point estimates and confidence intervals of the odds ratios comparing foreigners' propensity to commit crimes among clusters. Reference cluster $g^* = 2$.

	$\hat{w}_g^{\mathrm{f}}$	95% CI
Cluster 1	0.96	[0.913;1.002]
Cluster 2	1	
Cluster 3	1.62	[1.605;1.644]

Table 4.4. Point estimates and confidence intervals of the odds ratios comparing foreigners' propensity to commit crimes among clusters. Reference cluster $g^* = 3$.

	$\hat{w}_g^{\mathrm{f}}$	95% CI
Cluster 1	0.59	[0.562;0.617]
Cluster 2	0.62	[0.608;0.623]
Cluster 3	1	

While the odds of clusters 1 and 2 cannot be said to be different between them, they are significantly different from cluster 3. Indeed, the less virtuous cluster in terms of migrants integration is also the cluster with the highest proportion of immigrants held in prison. Recalling Figure 4.1, such a higher propensity of Cluster 3 might be led by the values of Italy, Austria, Czechia, Estonia, Albania and Greece, which showed the highest inflated incidence of foreigners over the total immigrant population. We do not aim to claim the existence of a causal relationship between the two phenomena. To investigate further, we look at the different propensities to detain foreigners and citizens within each cluster.

Within clusters

To test whether the immigrant propensity to be held in prison differs from that of citizens within each cluster, we still consider an FNCH model. In such a univariate case, for each cluster,

we consider an urn composed of two categories: foreigners and citizens. We assume the number of foreigners held in prison in a cluster, conditionally on the total number of people held in prison in that cluster, is FNCH distributed:

$$Y_g^{\mathrm{f}} \mid Y_g^{\mathrm{f}} + Y_g^{\mathrm{c}} = y_g \sim \mathrm{FNCH}(M_g^{\mathrm{f}}, M_g^{\mathrm{c}}, y_g, \phi_g), \quad \phi_g = \frac{p_g^f/(1-p_g^f)}{p_g^c/(1-p_g^c)} \tag{4.9}$$

Results are shown in Table 4.5. In all clusters, immigrants and citizens are differently exposed to being imprisoned.

In clusters 2 and 3, which we called the “most" and “less virtuous" clusters, respectively, foreigners are more inclined, or exposed, to be held in prison than citizens. Indeed, a person can be held in prison either after the delivery of the sentence or still unsentenced; immigrants generally suffer more from custodial pretrial measures than the so-called domestic detainees (e.g., for the Italian case, see Gonnella [2015]).

On the contrary, in Cluster 1, immigrants are slightly less exposed than citizens. A reason behind the latter result could lie in the fact that Cluster 1 groups transit countries for immigration. Such countries also have strong rejection policies, because of which they have been even sanctioned by the European Union[7]

Table 4.5. Point estimates and confidence intervals of the odds ratios that compare the propensity to commit crimes of foreigners and citizens within clusters (reference category: citizens).

	$g = 1$	$g = 2$	$g = 3$
$\hat{\phi}_g$	0.910	1.413	1.21
95% CI	[0.868;0.954]	[1.402;1.425]	[1.201;1.224]

4.5 Conclusions

Immigrant criminality and immigrants’ integration are long-studied issues. However, the association between the latter and immigrants’ propensity to commit crimes is poorly explored. In particular, to our knowledge, no quantitative analyses have been conducted on the relation between the two phenomena.

In this paper, we make a first attempt to test the existence of a link between immigrants’ integration and criminality in European countries. We leverage model-based clustering to group European countries according to the evaluation of their integration policies, modelling the eight

[7] E.g., see `https://ec.europa.eu/commission/presscorner/detail/en/ip_21_5801`

dimensions of MIPEX via finite mixtures of Gaussian densities. Then we compare the different exposures to criminality among and within clusters relying on Fisher's noncentral hypergeometric distribution, used to model clusters' counts of immigrants held in prison. We find that the cluster that shows the lowest mean values of the integration dimensions, i.e., Cluster 3, is also the cluster where the exposure of immigrants to imprisonment is higher. Moreover, in Clusters 3 and 2, the propensity to be held in prison among foreigners is also higher relatively to that of citizens. On the contrary, for Cluster 1, we observe that foreigners are slightly less exposed to detention than citizens. Such results might be either due to an actual lower propensity to commit crimes among the immigrants or led by unobserved mediators.

Indeed, a limitation of our work is that we do not consider possible covariates of interest in estimating the propensity to be held in prison. However, explaining the links through which integration policies might impact the immigrants' propensity to commit crimes goes beyond the scope of our work. Further research also relying on experts' sociological knowledge is needed. Another limitation of this work consists of neglecting the time dimension. Despite the proposed cross-sectional approach being helpful in providing a picture of the two phenomena together, it would be interesting to study the association between integration and crime from a panel perspective. Although the scarce availability of data makes the analysis not straightforward, for future development, we propose to explore the changes in the phenomena over time in a general model that accounts for possible covariates.

Chapter 5

Partial membership models for soft clustering of multivariate count data

Abstract

The standard mixture modelling framework has been widely used to study heterogeneous populations, by modelling them as being composed of a finite number of homogeneous subpopulations. However, the standard mixture model assumes that each data point belongs to one and only one mixture component, or cluster, but when data points have fractional membership in multiple clusters this assumption is unrealistic. It is in fact conceptually very different to represent an observation as partly belonging to multiple groups instead of belonging to one group with uncertainty. For this purpose, various soft clustering approaches, or individual-level mixture models, have been developed. In this context, Heller et al. [2008] formulated the Bayesian partial membership model (BPM) as an alternative structure for individual-level mixtures, which also captures partial membership in the form of attribute-specific mixtures, but does not assume a factorization over attributes. Our work proposes using the BPM for soft clustering of count data and compares the results with those achieved with the mixed membership model. Learning and inference are carried out using Markov chain Monte Carlo methods. The methods are demonstrated on simulated and real data, and it is applied on Capital Bike share data of Washington DC from 15 of June to 15 of July 2022, and on Serie A football players data, of the 2022/2023 football season.

5.1 Introduction

Model-based clustering has been widely used among researchers to study heterogeneous populations, by modelling them as being composed of a finite number of homogeneous subpopulations [Fraley and Raftery, 2002, Peel and MacLahlan, 2000]. Within this framework the observations in a dataset are modelled as they are drawn from one of several probability distributions. A clustering solution is sought whereby observations are partitioned into distinct groups, so that observations which have non-negligible posterior probability of belonging to more than one component are seen as having uncertain group membership, and are perhaps indicative of a poorly fit model. However, the standard mixture model assumes that each data point belongs to one and only one mixture component, or cluster, but when data points have fractional membership in multiple clusters this assumption is unrealistic; the idea of Mixed and Partial membership models accommodate partial membership. Following Heller et al. [2008]

example, let's consider an individual with a mixed ethnic background, say, partly Asian and partly European. It seems sensible to represent that individual as partly belonging to two different classes or sets. Being certain that a person is partly Asian and partly European, is very different than being uncertain about a person's ethnic background. The original idea for a mixed membership type of modeling goes back to at least the 1970s when the Grade of Membership (GoM) model was developed by mathematician Max Woodbury to allow for "fuzzy" classifications in medical diagnosis problems [Woodbury et al., 1978]. It was not until the early 2000s, with the widespread use of Bayesian methods and a better explanation of the duality between the discrete and continuous nature of latent structure in the GoM model, that a new Bayesian approach to the GoM model had been developed. Independently, within a short time of each other, three mixed membership models were developed to solve problems in three very different areas:

- Blei et al. [2003] – Latent Dirichlet Allocation (LDA)
- Erosheva [2003] – Grade of Membership model (GoM)
- Pritchard et al. [2000] – Admixture model

Mixed membership models unifies the LDA, GoM, and admixture models in a common framework and provides ways to construct other individual-level mixture models by varying assumptions on the population, sampling unit and latent variable levels, and the sampling scheme. In [Heller et al., 2008], Partial membership models are defined, which, albeit being part of the same framework, they overcome some of the drawbacks of mixed membership models. In the present paper we specify Partial membership models for count data, and so when component distributions are Poisson. We apply the method to Serie A football players data, of the 2022/2023 football season and to Capital Bike share data of Washington DC from 15 of June to 15 of July 2022. On the first application, the mixed membership model for count data, outlined in White and Murphy [2016], is also applied and the results achieved with the two models are compared. The comparison suggests that in the considered case, the partial membership model gives more realistic and interpretable results.

The article is organised as follows. Section 5.2 outlines the general partial membership model specification for Poisson components distribution, and compares the data generative process to those in mixed membership and mixture models. It also gives an overview on technical aspects like label switching and model selection, justifying the choice of the information criterion through literature and a simulation. In Section 5.3, both partial and mixed membership models are

applied to Serie A football players data and the results are compared[1]. Section 5.4 describes the application to Washington DC bike sharing data. Conclusions and future developments follow in 5.5.

5.2 Partial membership model

Consider a data set $\boldsymbol{X} = \{\boldsymbol{x}_{ij} : i = 1, 2, \ldots, N, j = 1, 2, \ldots, J\}$. In a finite mixture model, the density of a data point $\boldsymbol{x_i}$ given $\boldsymbol{\Theta}$, which contains the parameters for each of the K mixture components is

$$P(\boldsymbol{x_i}|\boldsymbol{\Theta}) = \sum_{\tau_i} P(\boldsymbol{\tau}_i) \prod_{k=1}^{K} P_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k)^{\tau_{ik}}. \tag{5.1}$$

Where $\boldsymbol{\tau}_i$ are the weights (or partial membership) which represents how much each data point belongs to each component, so $\tau_{ik} \in \{0, 1\}$ and $\sum_k \tau_{ik} = 1$. We relax the constrain $\tau_{ik} \in \{0, 1\}$ to take any continuous value in the range $[0, 1]$. So we change τ_{ik} from being binary to being in the simplex.

The complete data likelihood become.

$$P(\boldsymbol{x}_i|\boldsymbol{\Theta}) = \frac{1}{c} \int_{\boldsymbol{\tau}_i} P(\boldsymbol{\tau}_i) \prod_{k=1}^{K} P_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k)^{\tau_{ik}} d\boldsymbol{\tau}_i. \tag{5.2}$$

We integrate over all values of $\boldsymbol{\tau}_i$ instead of summing, and since the product over clusters K (in Equation 5.1) no longer normalizes, we put in a normalization constant c, which is a function of $\boldsymbol{\tau}_i$ and $\boldsymbol{\Theta}$. In this work we specify the case when the form of the distribution for each cluster $P_k(\boldsymbol{x}_i|\boldsymbol{\theta}_i)$ are Poisson

$$P_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k) = \prod_{j=1}^{J} P_k(\boldsymbol{x}_{ij}|\lambda_{kj}) = \prod_{j} \frac{\lambda_{kj}^{\boldsymbol{x}_{ij}} e^{-\lambda_{kj}}}{\boldsymbol{x}_{ij}!}. \tag{5.3}$$

Consider a model with K clusters and let $\boldsymbol{\delta}$ be a K-dimensional vector of positive hyperparameters ($\boldsymbol{\delta} \sim unif(a, b)$). We start by drawing mixture weights from a Dirichlet distribution:

$$\boldsymbol{\tau}_i \sim Dir(\boldsymbol{\delta}).$$

[1]For sake of completeness, a Poisson mixture model is also applied to this study. However, the results are only briefly explored, because a crisp clustering approach does not align with the specific objectives and purposes of the proposed application

that is the shorthand for

$$P(\boldsymbol{\tau}|\boldsymbol{\delta}) = c \prod_{k=1}^{K} \tau_k^{\delta_k - 1}. \tag{5.4}$$

Where $c = \frac{\Gamma(\sum_k \delta_k)}{\prod_k \Gamma(\delta_k)}$ is a normalization constant which can be expressed in terms of a Gamma function[2]. For each data point i, we draw a partial membership vector $\boldsymbol{\tau}_i$. We assumed that each cluster k is characterized by a Poisson distribution with natural parameters λ_{kj} and that

$$\lambda_{kj} \sim conj(\alpha, \beta).$$

A prior and likelihood are said to be *conjugate* when the resulting posterior distribution is the same type of distribution as the prior. Gamma distribution is a conjugate prior for the Poisson because they share the same functional form

$$P(\lambda_{kj}) \propto \lambda_{kj}^{\alpha-1} e^{-\beta\lambda_{kj}}. \tag{5.5}$$

Where α and β are hyperparameters of the prior[3]. Given all these latent variables[4], each data point is drawn from

$$x_{ij} \sim Pois(\exp(\sum_{k=1}^{K} \tau_{ik} \log \lambda_{kj})). \tag{5.6}$$

Which is the shorthand for

$$P(\boldsymbol{x}_{ij}|\boldsymbol{\tau}_i, \lambda_{kj}) \sim Pois(\prod_{k=1}^{K} \lambda_{kj}^{\tau_{ik}}) = Pois(\exp(\sum_{k=1}^{K} \tau_{ik} \log \lambda_{kj})). \tag{5.7}$$

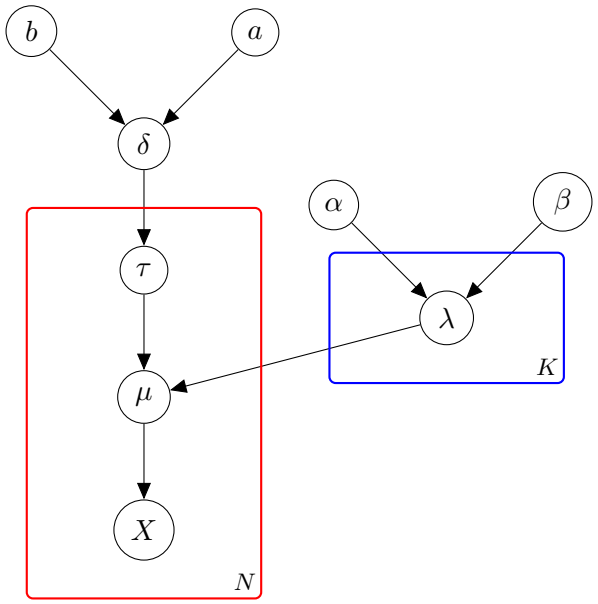
The generative process for $\boldsymbol{X}$ in the partial membership model, compared to those in mixed membership and mixture models is:

[2]The Gamma function generalizes the factorial to positive reals: $\Gamma(x) = (x-1)\Gamma(x-1)$, $\Gamma(n) = (n-1)!$ for integer n.

[3]The use of priors with different (sensible) choices of hyperparameters were found to have little effect on the clustering obtained by the application in Sect 5.3 and 5.4

[4]In the Bayesian framework the term latent variables could be used instead of parameters, to state that the model uses random variables that remain unobserved during inference.

Figure 5.1. Graphical model of the partial membership formulation



Mixture	**Mixed membership**	**Partial membership**
for(i in $1:N$)	for(i in $1:N$)	for(i in $1:N$)
$\boldsymbol{\tau}_i = \boldsymbol{\delta}$	$\boldsymbol{\tau}_i \sim Dirichlet(\boldsymbol{\delta})$	$\boldsymbol{\tau}_i \sim Dirichlet(\boldsymbol{\delta})$
$\boldsymbol{Z}_i \sim Multinomial(\boldsymbol{\tau}_i)$	for(j in $1:J$)	for(j in $1:J$)
for(j in $1:J$)	$\boldsymbol{Z}_{ij} \sim Multinomial(\boldsymbol{\tau}_i)$	$\mu_{ij} = \exp(\sum_{k=1}^{K} \tau_{ik} \log \lambda_{kj})$
$\boldsymbol{X}_{ij} \sim Poisson(\lambda_{Z_i,j})$	$\boldsymbol{X}_{ij} \sim Poisson(\lambda_{Z_{ij},j})$	$\boldsymbol{X}_{ij} \sim Poisson(\mu_{ij})$

A method for fitting mixed membership models to count data is outlined in White and Murphy [2016]. Partial membership model does not assume a factorization over attributes. More generally, mixed membership (MM) models, assume that each data attribute (for instance in a text analysis example the data attributes could be the words) of the data point (e.g. document) is drawn independently from a mixture distribution given the membership vector for the data point, $x_{nj} \sim \sum_k \tau_{nk} P(\boldsymbol{x}|\lambda_{kj})$. MM models only makes sense when the objects (e.g. documents) being modelled constitute bags of exchangeable sub-objects (e.g. words). Partial membership models make no such assumption.

The complete-data posterior takes the form

$$P(\boldsymbol{\tau}, \boldsymbol{\lambda}|\boldsymbol{x}, \alpha, \beta, a, b) \propto P(\boldsymbol{\lambda}|\alpha, \beta) P(\boldsymbol{\tau}|\boldsymbol{\delta}) \prod_{k=1}^{K} \prod_{j=1}^{J} P_k(\boldsymbol{x}_j|\lambda_{kj})^{\tau_k}. \tag{5.8}$$

Learning in the BPM consists of inferring all unknown variables given $\boldsymbol{X}$, for which we employ Monte Carlo Markov chain (MCMC). Another advantage of BPM over MM models, is that in the latter there is a discrete latent variable for every sub-object, corresponding to which mixture component that sub-object was drawn from. This large number of discrete latent variables makes MCMC sampling in MM potentially much more expensive than in BPM models.

5.2.1 Model selection

According to Watanabe and Opper [2010], a statistical model is said to be *regular* if the map taking parameters to probability distributions is one-to-one and if its Fisher information matrix is positive definite. If a model is not regular, then it is said to be *singular*. If a statistical model contains a hierarchical structure, hidden variables, or a grammatical rule, then the model is generally singular. In singular statistical models, the maximum likelihood estimator does not satisfy asymptotic normality. Consequently, AIC is not equal to the average generalization error [Hagiwara, 2002], and the Bayes information criterion (BIC) is not equal to the Bayes marginal likelihood [Watanabe, 2001], even asymptotically. In singular models, the maximum likelihood estimator often diverges, or even if it does not diverge, makes the generalization error very large. Therefore, the maximum likelihood method is not appropriate for singular models. On the other hand, Bayes estimation was proven to make the generalization error smaller if the statistical model contains singularities. WAIC [Watanabe and Opper, 2010], (Widely Applicable Information Criterion) could be used for estimating the predictive loss of Bayesian models, using a sample from the full-data posterior, and it is applicable to non-regular models, including non-identifiable models and non-realizable models.

In the present paper, WAIC is calculated from Equations 5, 12, and 13 in Gelman et al. [2014], and it is the log pointwise predictive density minus a correction for effective number of parameters to adjust for overfitting. According to Millar [2018], the **marginalized WAIC** might be more accurate for choosing the right model. We run the model in a simulated data scenario to verify the number of times the right model is chase by the WAIC conditional to all the parameters (WAICc) and marginalized for $\boldsymbol{\tau}$ (WAICm). We generated 100 random membership vectors from a Dirichlet($\boldsymbol{\delta}$) distribution with shape parameter $\boldsymbol{\delta} = (0.5, 0.5, 0.5, 0.5)$. Using these membership scores, we simulated 100 partial membership models with $N = 200$, $J = 25$ and $K = 4$ to match the football players application scenario. We ran the Gibbs sampling algorithm with a MCMC step for 20000 iterations, keeping every 50th draw. We discarded

the first 5000 of the retained draws as burn-in. We therefore run the model over a range of values of $K = 1, ..., 8$. To asses convergence, we examined trace plots. In this paper the label switching has been assessed permuting after the run of the model the labels to each MCMC draw, using the probabilistic relabelling algorithm of Sperrin et al. [2010], provided by the R package label.switching. The simulation results are illustrated in Table 5.1.

Table 5.1. Number of times each K correspond to the minimum WAIC conditional to all the parameters (WAICc) and marginalized for $\boldsymbol{\tau}$ (WAICm). N=200, J=25, Number of runs=100, true K=4

	WAICc	WAICm
K=1	0	-
K=2	0	0
K=3	0	0
K=4	79	99
K=5	16	1
K=6	3	0
K=7	2	0
K=8	0	0

The simulation confirm that the marginalized WAIC, is more accurate for choosing the right model, with 99% of success.

5.3 Serie A football players

We selected the stats of the 192 Serie A football players who played more than 1350 minutes during the 2022/2023 football season[5]. The analysis consider a set of 22 count variables recorded during the games, selected to encompass the essential skills associated with each player's role on the field. This application enables us to verify the reliability of the model's results by comparing them with each player's actual playing position. The partiality of the membership also allows us to estimate the positions on the field where the players tend to play, in addition to their primary position, based on their playing style. Both MM and Partial Membership (PM) models are applied to the dataset, with the WAIC suggesting that 4 and 5 profile fits are optimal, as illustrated in Table 5.2.

[5]The data are freely available at `https://fbref.com`.

Table 5.2. Profiles WAIC

WAIC	K=2	K=3	K=4	K=5	K=6	K=7	K=8
PM	52237.45	44274.20	37692.89	47919.76	42843.47	43603.61	41295.05
MM	35330.92	29833.80	28309.35	27898.17	28096.84	28480.93	29099.70

It follows a brief description of the variables analysed:

- **Gls** – Number of goals.
- **Ast** – Number of assists.
- **PrgC** – Progressive carries: carries that move the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes, or any carry into the penalty area. Excludes carries which end in the defending 50% of the pitch.
- **PrgP** – Progressive Passes: progressive Passes completed passes that move the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area. Excludes passes from the defending 40% of the pitch.
- **Sh** – Shots Total: does not include penalty kicks.
- **SoT** – Shots on Target: does not include penalty kicks.
- **KP** – Key Passes: Passes that directly lead to a shot (assisted shots).
- **PiFT** – Passes into Final Third: completed passes that enter the 1/3 of the pitch closest to the goal, not including set pieces.
- **PPA** – Passes into Penalty Area: completed passes into the 18-yard box, not including set pieces.
- **CrsPA** – Crosses into Penalty Area: completed crosses into the 18-yard box, not including set pieces.
- **SCA** – Shot-Creating Actions: the two offensive actions directly leading to a shot, such as passes, take-ons and drawing fouls.
- **PassLive** – SCA (PassLive): completed live-ball passes that lead to a shot attempt.
- **PassDead** – SCA (PassDead): completed dead-ball passes that lead to a shot attempt. Includes free kicks, corner kicks, kick offs, throw-ins and goal kicks.
- **TO** – SCA (TO): successful take-ons that lead to a shot attempt.
- **ShToSh** – SCA (Sh): shots that lead to another shot attempt.

- **Def** – SCA (Def): defensive actions that lead to a shot attempt.
- **GCA** – Goal-Creating Actions: the two offensive actions directly leading to a goal, such as passes, take-ons and drawing fouls. Note: a single player can receive credit for multiple actions and the shot-taker can also receive credit.
- **Tkl** – Tackles: Number of players tackled
- **Blocks** – Number of times blocking the ball by standing in its path
- **Int** – Interceptions
- **Clr** – Clearances
- **Err** – Errors: Mistakes leading to an opponent's shot

All the analysis has been carried out using NIMBLE, a system for programming statistical algorithms for general model structures within R [de Valpine et al., 2017].

5.3.1 Partial membership model application

In Table 5.3 and in Figure 5.2 are presented the profiles means.

Figure 5.2. Expected profiles means, conditional on profile membership, with 4 profiles.

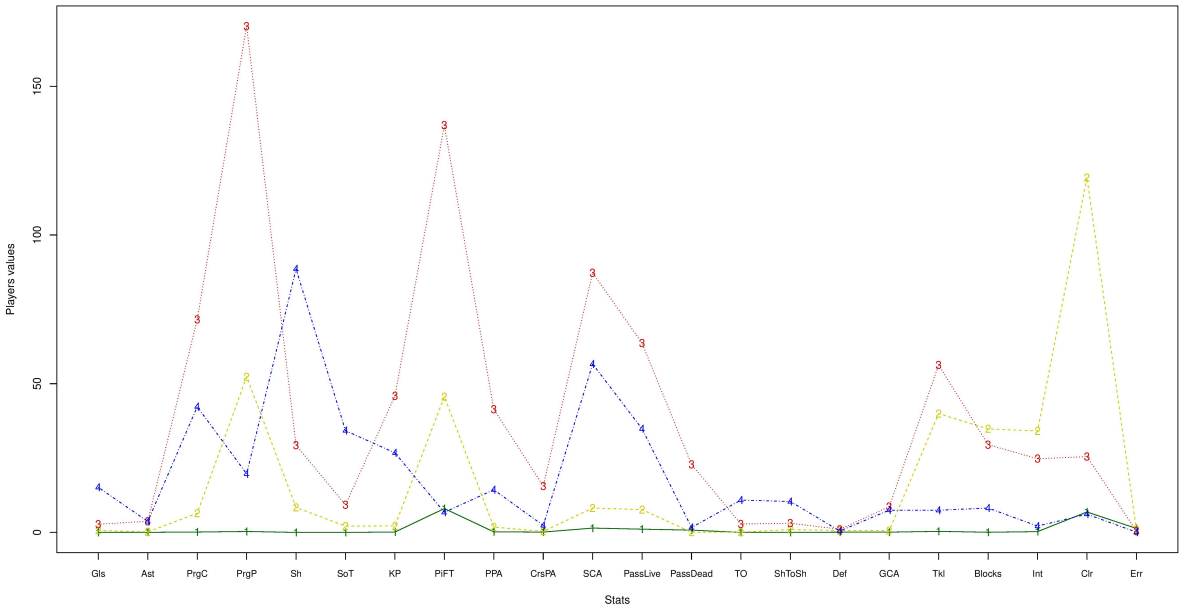


Table 5.3. PM model expected profiles means

	Gls	Ast	PrgC	PrgP	Sh	SoT	KP	PiFT	PPA	CrsPA	SCA
1	0.00	0.04	0.13	0.32	0.00	0.00	0.13	8.02	0.19	0.11	1.45
2	0.65	0.19	6.42	52.34	8.39	2.11	2.21	45.67	1.79	0.36	8.18
3	2.74	3.75	71.57	170.20	29.29	9.25	45.91	136.88	41.41	15.58	87.28
4	15.14	3.70	42.25	19.69	88.57	34.20	26.75	6.75	14.34	2.34	56.61
	PassLive	PassDead	TO	ShToSh	Def	GCA	Tkl	Blocks	Int	Clr	Err
1	1.09	0.77	0.01	0.00	0.07	0.09	0.36	0.06	0.26	6.87	1.34
2	7.68	0.04	0.11	0.95	0.63	0.59	39.94	34.80	34.11	119.34	1.01
3	63.60	22.84	2.82	3.11	1.05	8.64	56.30	29.47	24.78	25.48	0.43
4	34.81	1.66	10.87	10.39	0.48	7.48	7.45	8.20	2.13	6.13	0.10

In interpreting the profile means, we observe that:

- **Profile 1** consistently exhibits low values compared to the other profiles across almost every variable, except for *Err – Errors*. This profile can be interpreted as grouping the **goalkeepers**, as they typically have significantly lower average values in the considered variables. These variables are more relevant to players in more active roles involving ball possession.

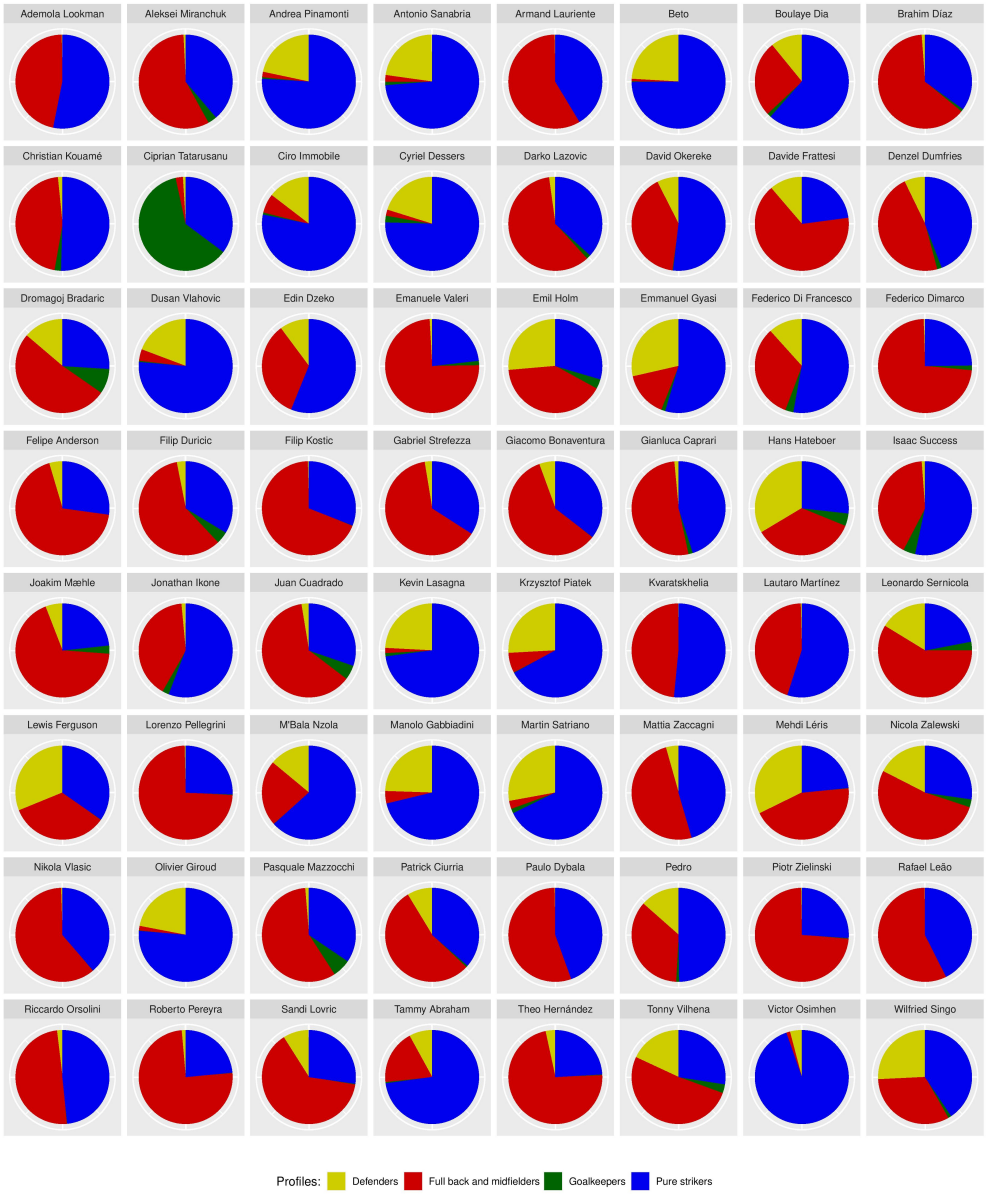
- **Profile 2** is characterized by notably high values in *Clr – Clearances*, *Int – Interceptions*, *Blocks – Number of times blocking the ball by standing in its path*, and *Tkl – Tackles*. Additionally, it exhibits relatively high values in *Err – Errors* and *Def – Defensive actions that lead to a shot attempt* compared to the other profiles. These characteristics are commonly associated with **defenders**.

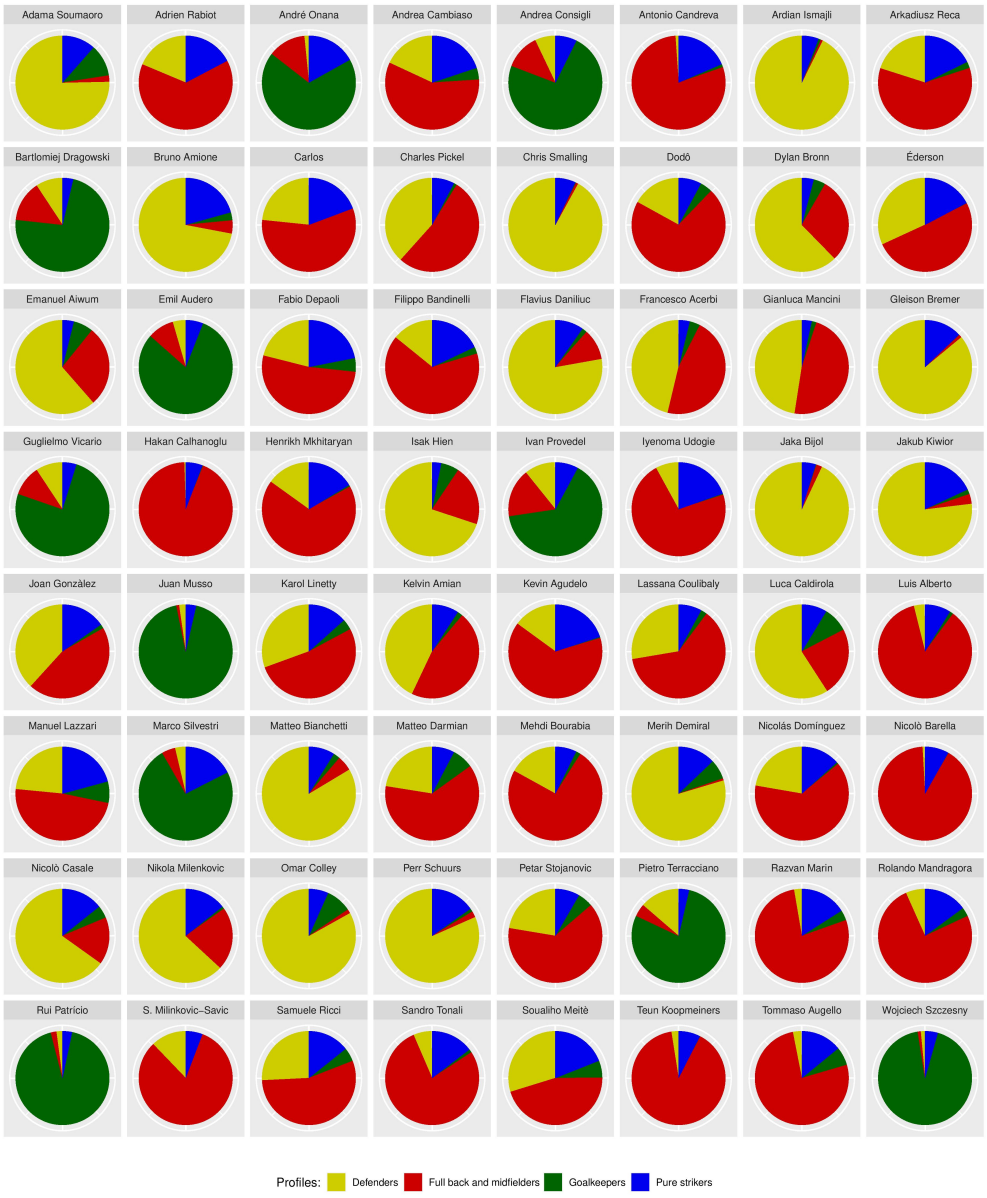
- **Profile 3** displays remarkably high values in *Tkl – Tackles*, *GCA – Goal-Creating Actions*, *Def – Defensive actions that lead to a shot attempt*, *PassDead – Completed dead-ball passes that lead to a shot attempt*, *PassLive – Completed live-ball passes that lead to a shot attempt*, *SCA – Shot-Creating Actions*, *CrsPA – Crosses into Penalty Area*, *PPA – Passes into Penalty Area*, *PiFT – Passes into Final Third*, *KP – Key Passes*, *PrgP – Progressive Passes*, *PrgC – Progressive carries*. Additionally, it exhibits high values in *Ast – Number of assists*. These characteristics, combined with the consistently high values across all these variables, suggest an association with **full-backs and midfielders**. These positions involve moving the ball around the field and are frequently positioned centrally during the game.

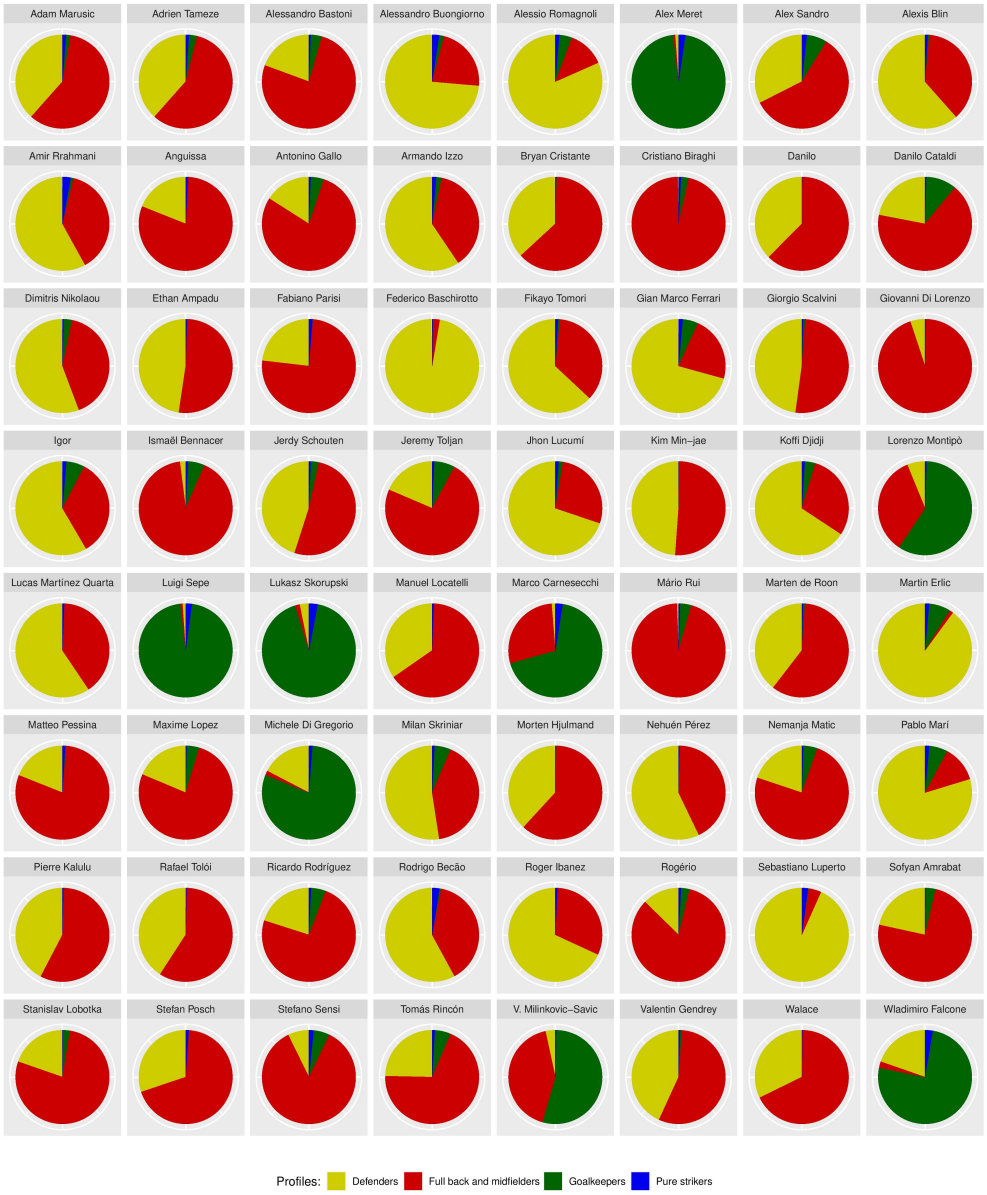
- **Profile 4** demonstrates remarkably high values in *ShToSh – Shots that lead to another shot attempt*, *TO – Successful take-ons that lead to a shot attempt*, *SoT – Shots on Target*, *Sh – Shots Total*, and *Gls – Number of goals*. These characteristics are typically associated with **pure strikers**.

Table .2 in Appendix 6 represent each football player profile membership and Figure 5.3 the corresponding pie chart.

Figure 5.3. Football players' pie charts of profiles membership.




Adama Soumaoro
Adrien Rabiot
André Onana
Andrea Cambiaso
Andrea Consigli
Antonio Candreva
Ardian Ismajli
Arkadiusz Reca
Bartlomiej Dragowski
Bruno Amione
Carlos
Charles Pickel
Chris Smalling
Dodô
Dylan Bronn
Éderson
Emanuel Aiwum
Emil Audero
Fabio Depaoli
Filippo Bandinelli
Flavius Daniliuc
Francesco Acerbi
Gianluca Mancini
Gleison Bremer
Guglielmo Vicario
Hakan Calhanoglu
Henrikh Mkhitaryan
Isak Hien
Ivan Provedel
Iyenoma Udogie
Jaka Bijol
Jakub Kiwior
Joan González
Juan Musso
Karol Linetty
Kelvin Amian
Kevin Agudelo
Lassana Coulibaly
Luca Caldirola
Luis Alberto
Manuel Lazzari
Marco Silvestri
Matteo Bianchetti
Matteo Darmian
Mehdi Bourabia
Merih Demiral
Nicolás Domínguez
Nicolò Barella
Nicolò Casale
Nikola Milenkovic
Omar Colley
Perr Schuurs
Petar Stojanovic
Pietro Terracciano
Razvan Marin
Rolando Mandragora
Rui Patrício
S. Milinkovic−Savic
Samuele Ricci
Sandro Tonali
Soualiho Meité
Teun Koopmeiners
Tommaso Augello
Wojciech Szczesny
Profiles:
Defenders
Full back and midfielders
Goalkeepers
Pure strikers



Based on the findings presented in Figure 5.3, it is evident that the model effectively captures the playing positions of the football players and successfully highlights the nuances of their playing styles. The results could have practical implications for coaches, talent scouts, team managers, and analysts. These stakeholders can utilize the findings to make informed decisions related to team strategy, talent acquisition, and statistical research, ultimately enhancing performance and understanding in the field of football.

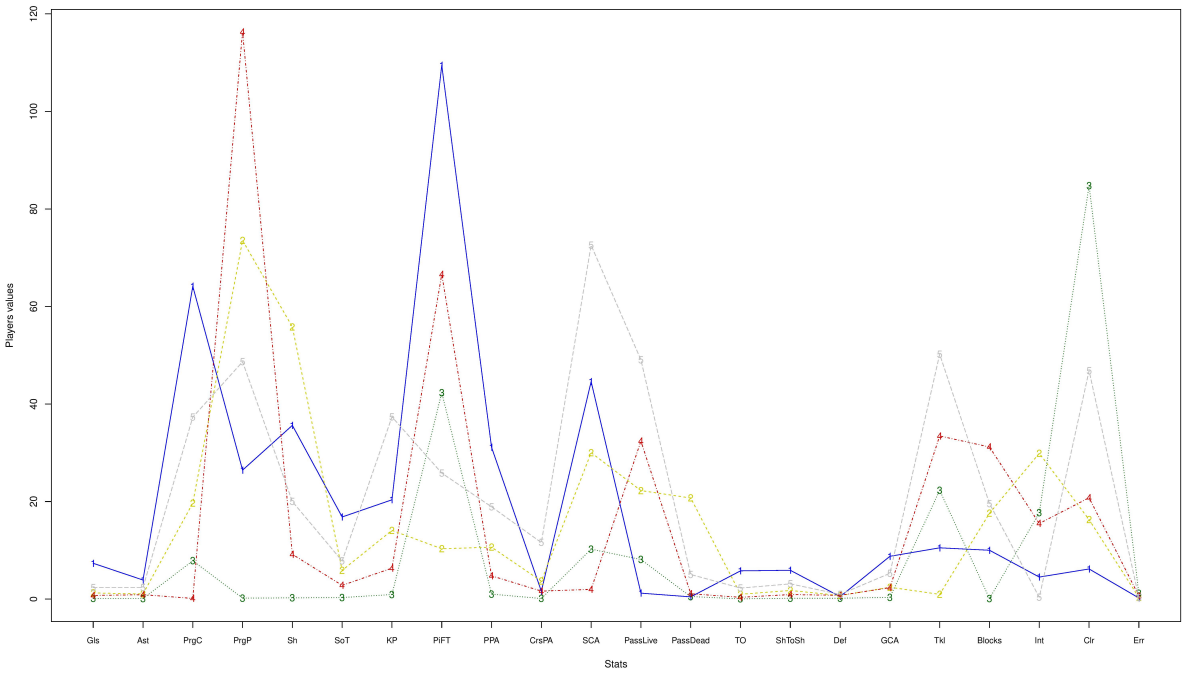
5.3.2 Mixed membership model application

Mixed membership models are applied to count data in White and Murphy [2016]. For a more in-depth explanation of the model than that given in this paper, please refer to that article. We performed the model and the profiles means are presented in Table 5.4 and Figure 5.4.

Table 5.4. MM model expected profiles means

	Gls	Ast	PrgC	PrgP	Sh	SoT	KP	PiFT	PPA	CrsPA	SCA
1	7.34	3.91	64.14	26.44	35.62	16.83	20.35	109.46	31.13	1.60	44.59
2	1.25	1.04	19.66	73.55	55.81	5.89	14.09	10.32	10.64	3.68	30.02
3	0.12	0.11	7.86	0.19	0.25	0.31	0.93	42.31	1.01	0.11	10.24
4	0.74	0.90	0.13	116.27	9.17	2.81	6.34	66.60	4.77	1.64	1.99
5	2.35	2.36	37.30	48.66	20.00	7.80	37.36	25.83	18.92	11.59	72.56
	PassLive	PassDead	TO	ShToSh	Def	GCA	Tkl	Blocks	Int	Clr	Err
1	1.21	0.45	5.79	5.90	0.50	8.77	10.51	10.01	4.51	6.17	0.22
2	22.25	20.73	0.98	1.76	0.72	2.42	0.96	17.60	29.90	16.28	0.33
3	8.11	0.52	0.04	0.17	0.17	0.35	22.30	0.10	17.71	84.78	1.17
4	32.39	1.09	0.36	0.93	0.76	2.35	33.45	31.19	15.49	20.77	0.64
5	49.05	5.04	2.25	3.12	0.96	5.28	50.20	19.57	0.36	46.82	0.27

Figure 5.4. MM model expected profiles means, conditional on profile membership, with 5 profiles.



The interpretation of the profile means in the MM model is less immediate compared to the PM model.

- **Profile 1** suggests a grouping of strikers, but it also exhibits high mean values in variables more typical of midfielders, such as *PrgC – Progressive carries.*
- **Profile 2** shows high mean values in *Int – Interceptions*, *Blocks – Number of times blocking the ball by standing in its path*, *PassDead – Completed dead-ball passes that lead to a shot attempt*, and *Sh – Shots Total.* These characteristics could be associated with defensive midfielders, although Sh - Shots Total is not typical for this role.
- **Profile 3** generally has low values in most variables, indicating a potential grouping of goalkeepers. However, it also presents high values in *Int — Interceptions* and *Tkl – Tackles*, which are more commonly associated with defenders.
- **Profile 4** exhibits very high values in *PrgP – Progressive Passes*, *Blocks – Number of times blocking the ball by standing in its path*, and *Tkl – Tackles.* This profile could be interpreted as grouping offensive midfielders, even though Blocks are not typically associated with this role.
- **Profile 5** might group defenders and full-backs, but these two positions have distinct characteristics, making the interpretation somewhat challenging.

5.3.3 Models comparison

When comparing clustering models, one of the key parameters to consider is the interpretability of the results. This aspect is extensively discussed in Fraley and Raftery [1998], Forgy [1965]. The ability to interpret the clusters and derive meaningful insights from them is crucial in various domains, including sports analysis.

When comparing the two models, it becomes evident that the interpretation of the clusters is significantly easier and more accurate in the PM model compared to the MM model. This disparity can be attributed to the inherent differences in the underlying assumptions of the two models. The PM model, unlike the MM model, does not assume factorization over attributes. This means that each data attribute of a given data point is not assumed to be drawn independently from a mixture distribution based on the membership vector. In contrast, MM models are designed to handle situations where the objects being modeled consist of exchangeable sub-objects. The lack of such assumptions in the PM model enhances its interpretability in our application. It

allows for a more straightforward and intuitive understanding of the clusters, as the model does not impose strict dependencies between the attributes. As a result, the PM model excels in capturing the nuances of different playing positions without forcing the interpretation to conform to specific attribute relationships. In summary, the ease and quality of cluster interpretation are superior in the PM model compared to the MM model. The flexibility of the PM model, driven by its independence assumptions between attributes, enables a more accurate representation of the diverse playing positions in football without imposing restrictive assumptions on the data.

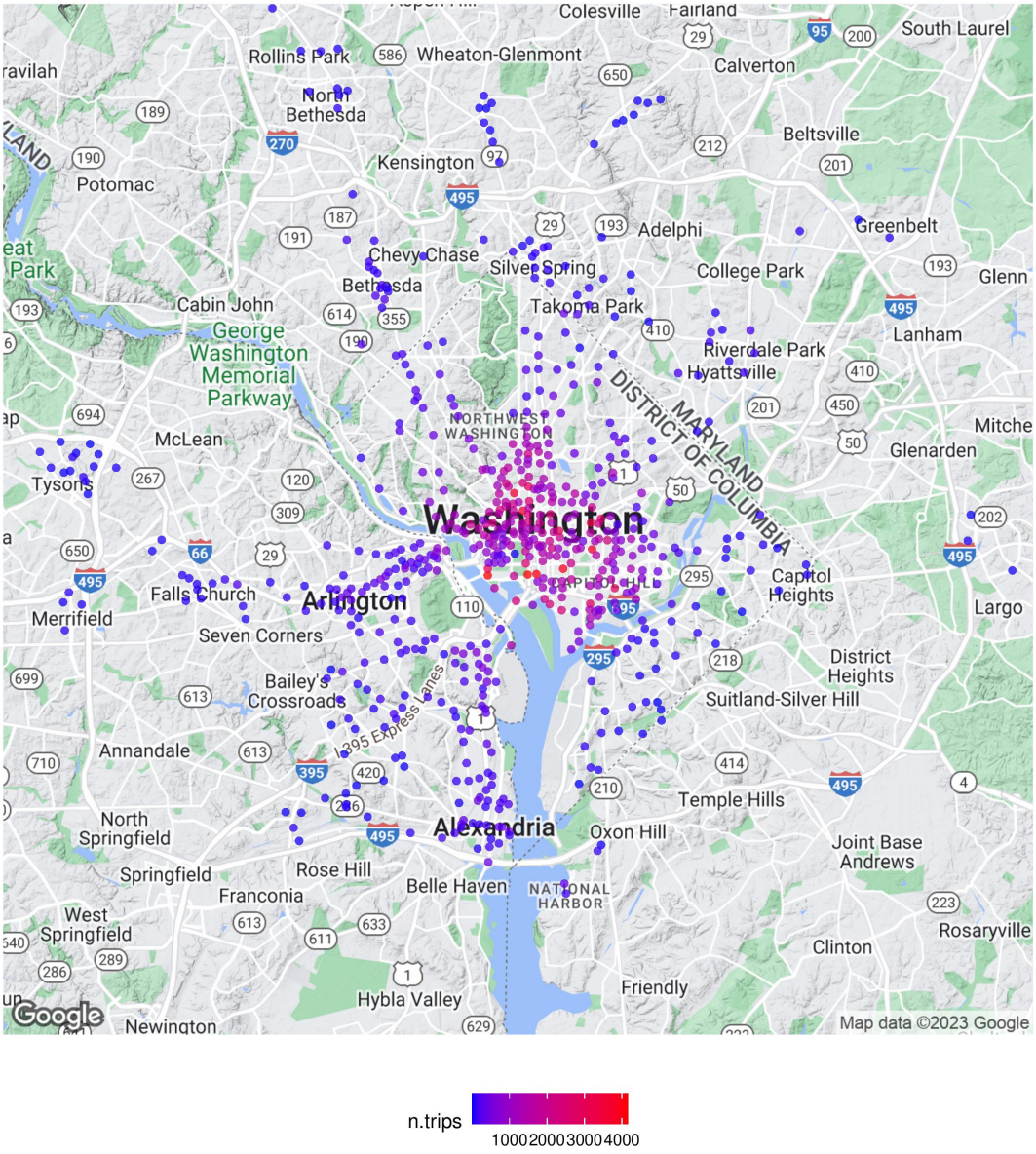
For sake of completeness, a Poisson mixture model was also applied in this study. The model was run for a range of K values from 1 to 30, and the best model selected using the BIC, was the model with 23 components. However, the results are not reported in this article, because a crisp clustering approach does not align with the specific objectives and purposes of the proposed applications, as explained in the Introduction 5.1.

5.4 Washington DC bikes data

We apply partial membership model on the data of the bike sharing company of Washington DC. The data are collected daily, from 15 of June to 15 of July 2022, and record each single ride: date and time of start of trip, date and time of end of trip, name, ID, longitude and latitude of starting station, name, ID, longitude and latitude of ending station[6]. Figure 5.5 shows the number of trips per station in the considered time period.

[6]The data are freely available at `https://capitalbikeshare.com/system-data`

Figure 5.5. Number of bicycle trips per bike sharing station from 15 of June to 15 of July 2022



We calculated the number of times bikes are collected from each of the 660 stations and we modelled these counts using a partial membership model, with the intent to explore the interactions between the bikes stations usage, to improve the allocation of the bikes. Partial membership model suits this type of application because the bikes move between the stations along the day, so the stations usage could vary and their membership could be partial. It should be noted that we do not addressed the temporal dependency as the temporal nature of the data would require. Nevertheless, the approaches appear to identify interesting behaviour in the data, and serve to illustrate the usefulness of the method. We run the model over a range of

$K = 1, \ldots, 6$. The model with the lowest WAIC is the one with 5 profiles (or components). For a better visualization, in Figure 5.6 are represented the natural log of the profiles means, while Figure 5.7 shows the marginal simplices representing stations' profile membership.

Figure 5.6. Log of the expected number of rides per day from 15 of June to 15 of July 2022, conditional on profile membership, with 5 profiles.

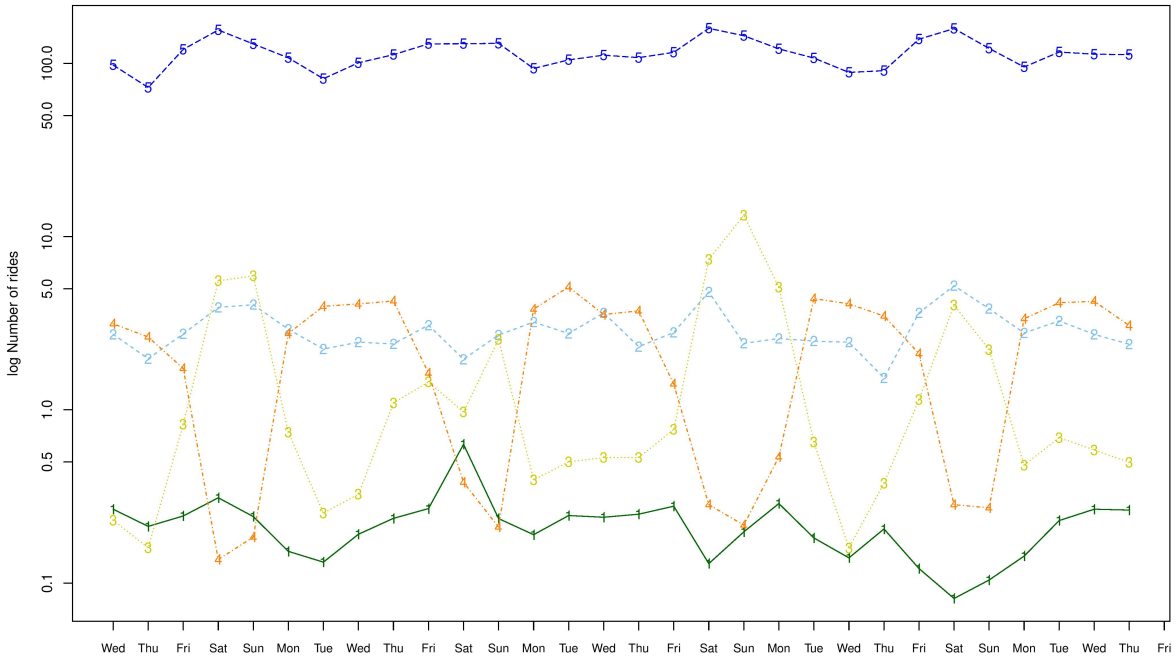


Figure 5.7. Marginal simplices representing bikes stations' profile membership

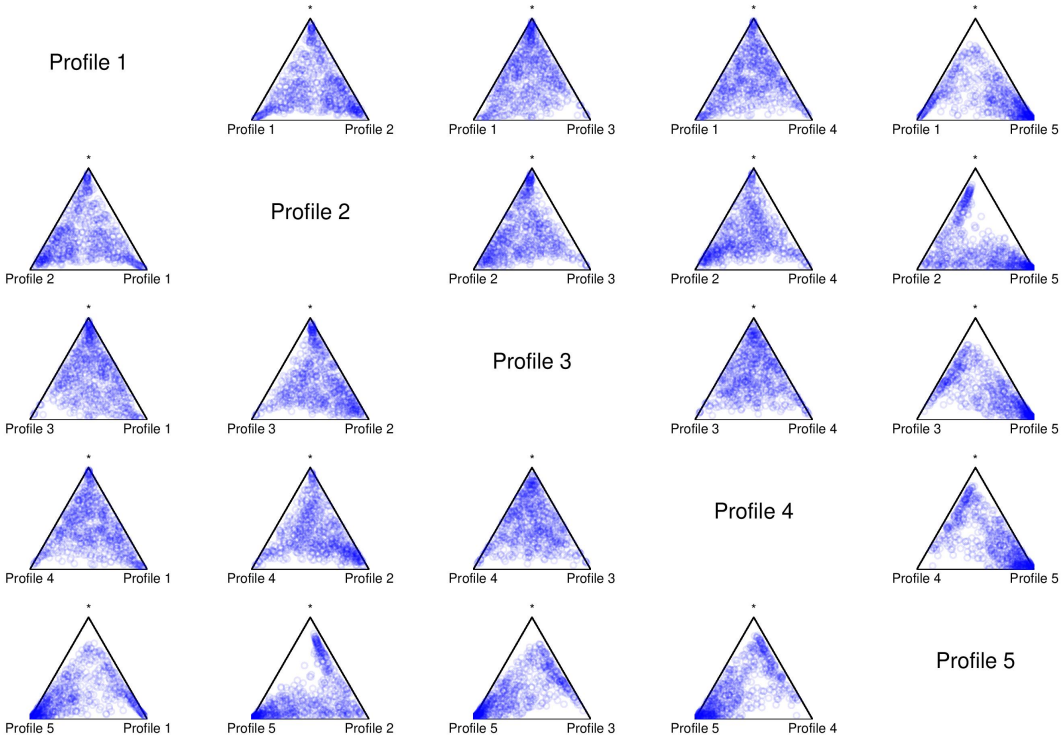
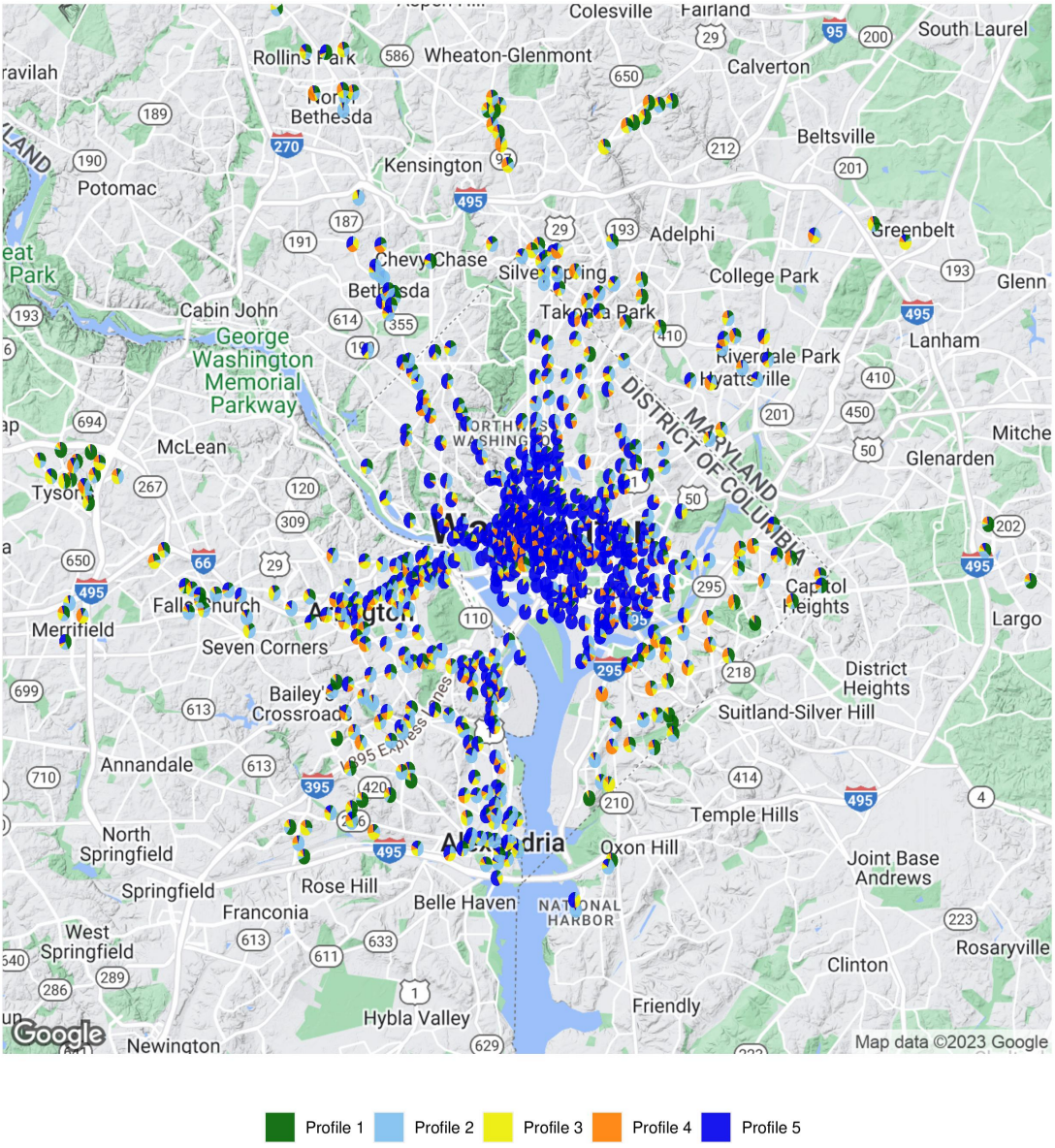


Figure 5.8 represent a pie chart on each bike stations, of the profiles memberships of each one.

Figure 5.8. Bike stations' pie charts of profiles membership.



It could be seen that profile 5 groups the busiest stations, which are mainly located in the center of the city. Profile 1 the less used ones, which are mainly in the outlying areas, profile 2 is an average usage stations cluster and looking at the map, it seems to connect the centre to the peripheral areas, profile 3 groups the stations mostly used during the weekends, with an high peak of usage during the holiday of Monday 4 of July, which is bank holiday in the States. The stations with an high membership to this profile, are often located near the river or green ares, or also in the outlying areas. Profile 4 is the group of the stations mostly used on working days.

5.5 Conclusions and future developments

Partial membership models provide the analyst with tools of greater flexibility than current model based clustering or standard distance-based clustering methods. We specified the model for count data and applied to Serie A football players data and bike sharing data of Washington DC. In the football players application, we also compared the results to those obtained with a mixed membership models for count data. We observed better and more interpretable results with the partial membership model. We think this model, while suffering from very high computational times can be of great use in many applications such as social sciences, genetics, natural sciences and textual analysis, and can overcome some of the limitations of mixed membership models. As a future development it would be interesting to explore solutions to assess the over-dispersion issue.

Chapter 6

Conclusions

This dissertation explored model based clustering and some of its possible uses in the social sciences. The presented works and their applications are examples of useful models in this framework, and although the models are not all completely original, they are nevertheless innovative.

Chapter 1 described some of the main issues when analysing social science data, and introduced the usefulness of clustering models in this area of topics. Chapter 2 focused on the models formulation and estimation. Firstly, finite mixture models were defined and how they are used for clustering. Then, a brief introduction to maximum likelihood estimation methods, the EM algorithm and the Bayesian analysis of mixtures and its common issues. The aim was to introduce the basic ingredients for the understanding of each methodological choice that has been undertaken in the following applications. Chapter 3 included the application of a Mixture of Matrix-Normals classification model for longitudinal data, to 7 MIPEX dimensions from 2014 to 2019. This work added new reading perspectives on the MIPEX data, by grouping countries which behave similarly across and within time, in order to facilitate the comparison and interpretation of the phenomenon. The work is currently published in Alaimo et al. [2021a]. While finite Mixture of Matrix-Normals models are not new, there exist only very few applications in literature, and they are all very recent. Chapter 4 presented a first attempt to test the existence of a link between immigrants' integration and criminality in European countries. European countries have been clustered according to the evaluation of their integration policies, modelling the eight dimensions of MIPEX for the year 2019, via finite mixtures of Gaussian densities. Then, the different exposures to criminality among and within clusters have been compared relying on Fisher's noncentral hypergeometric distribution, used to model clusters' counts of

immigrants held in prison. The novelty of the paper, currently under review, stands on linking model based clustering with the Fisher's noncentral hypergeometric distribution. This approach could be helpful to explore if there is a correspondence between the attitude of the units toward a phenomenon and the propensity to a certain behaviour for another, which could be supposed to be related to the first one. Moreover, to our knowledge, no quantitative analyses have been conducted on the relation between immigrant criminality and immigrants' integration. This work is currently under referral process. Chapter 5 introduced the specification of partial membership models for count data. In contrast to the other treated models, this methodology allows for a partial classification of units instead of a crisp one. It is therefore suitable for cases where units have fractional membership in multiple clusters. The model is applied to Serie A football players data, with the aim to estimate the playing positions of the football players and highlights the nuances of their playing styles. On this application it is also compared with the results achieved with mixed membership model. Also, the model is applied to the bike sharing data of Washington DC, with the aim to improve the allocations of the bikes, and so to improve the urban mobility.

Summing up, the methods presented in this thesis, demonstrated the potential of model-based clustering for addressing complex problems in the social sciences. However, further research is needed to fully explore the capabilities and limitations of these methods. Overall, the use of clustering models has the potential to provide valuable insights and information that can help bridge the gap between research and the wider public. By identifying patterns and groups within complex datasets, these methods can offer a more nuanced and comprehensive understanding of social phenomena. This, in turn, can inform policy decisions and facilitate communication between researchers, policymakers, and the general public. Moving forward, it will be important to continue developing and refining these methods to maximize their impact on the social sciences and beyond.

Bibliography

A. Agresti. *Categorical data analysis*, volume 792. John Wiley & Sons, 2002. doi: 10.1002/0471249688.

L. S. Alaimo. Complexity of social phenomena: Measurements, analysis, representations and synthesis. *Unpublished doctoral dissertation, University of Rome "La Sapienza", Rome, Italy*, 2020.

L. S. Alaimo. Complexity and knowledge. In F. Maggino, editor, *Encyclopedia of Quality of Life and Well-being Research*, pages 1–2. Cham: Springer, 2021a. doi: 10.1007/978-3-319-69909-7_104658-1.

L. S. Alaimo. Complex systems and complex adaptive systems. In F. Maggino, editor, *Encyclopedia of Quality of Life and Well-being Research*, pages 1–3. Cham: Springer, 2021b. doi: 10.1007/978-3-319-69909-7_104659-1.

L. S. Alaimo and F. Maggino. Sustainable Development Goals Indicators at Territorial Level: Conceptual and Methodological Issues — The Italian Perspective. *Social Indicators Research*, 147(2):383–419, 2020. doi: 10.1007/s11205-019-02162-4.

L. S. Alaimo and E. Seri. Monitoring the main aspects of social and economic life using composite indicators: A literature review. *Working papers Research group Economics, Policy Analysis, and Language; Ulster University*, W.P. 21-7:1–58, 2021.

L. S. Alaimo and E. Seri. Measuring human development by means of composite indicators: open issues and new methodological tools. *Quality & Quantity*, pages 1–33, 2023. doi: 10.1007/s11135-022-01597-1.

L. S. Alaimo, F. Amato, F. Maggino, A. Piscitelli, and E. Seri. A Comparison of Migrant

Integration Policies via Mixture of Matrix-Normals. *Social Indicators Research*, 12(3):327–337, 2021a. doi: 10.1007/s11205-022-03024-2.

L. S. Alaimo, A. Arcagni, M. Fattore, and F. Maggino. Synthesis of multi-indicator system over time: A poset-based approach. *Social Indicators Research*, 157(1):77–99, 2021b.

L. S. Alaimo, F. Amato, and E. Seri. A longitudinal cross country comparison of migrant integration policies via Mixture of Matrix-Normals. In C. C. A. Balzanella, M. Bini and R. Verde, editors, *Book of the Short Papers of the 51st Scientific Meeting of the Italian Statistical Society*, pages 1136–1141. Springer. ISBN-9788891932310, 2022a.

L. S. Alaimo, A. Arcagni, M. Fattore, F. Maggino, and V. Quondamstefano. Measuring Equitable and Sustainable Well-being in Italian Regions. The Non-aggregative Approach. *Social Indicators Research*, 161:711–733, 2022b. doi: 10.1007/s11205-020-02388-7.

A. Alesina and M. Tabellini. The political effects of immigration: Culture or economics? Technical report, National Bureau of Economic Research, 2022.

V. Ballerini and B. Liseo. Fisher's Noncentral Hypergeometric Distribution for Population Size Estimation. In C. C. A. Balzanella, M. Bini and R. Verde, editors, *Book of the Short Papers of the 51st Scientific Meeting of the Italian Statistical Society*, pages 1600–1605. Springer. ISBN-9788891932310, 2022.

J. D. Banfield and A. E. Raftery. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, 49(3):803–821, Sep 1993. doi: 10.2307/2532201.

K. G. Banting, W. Kymlicka, et al. *Multiculturalism and the welfare state: Recognition and redistribution in contemporary democracies*. Oxford University Press on Demand, 2006. doi: 10.1093/acprof:oso/9780199289172.001.0001.

F. Bartolucci, A. Farcomeni, and F. Pennoni. *Latent Markov models for longitudinal data*. New York: Chapman and Hall/CRC, 2012. doi: 10.1201/b13246.

K. E. Basford and G. J. McLachlan. The mixture method of clustering applied to three-way data. *Journal of Classification*, 2(1):109–125, 1985. doi: 10.1007/BF01908066.

G. S. Becker. Crime and punishment: An economic approach. In *The economic dimensions of crime*, pages 13–68. Springer, New York, 1968. doi: 10.1007/978-1-349-62853-7_2.

M. Beine, A. Boucher, B. Burgoon, M. Crock, J. Gest, M. Hiscox, P. McGovern, H. Rapoport, J. Schaper, and E. Thielemann. Comparing immigration policies: An overview from the impala database. *International Migration Review*, 50(4):827–863, 2016. doi: 10.1111/imre.12169.

M. Bianchi, P. Buonanno, and P. Pinotti. Do immigrants cause crime? *Journal of the European Economic Association*, 10(6):1318–1347, 2012. doi: 10.1111/j.1542-4774.2012.01085.x.

C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3):561–575, 2003. doi: 10.1016/S0167-9473(02)00163-9.

L. Bjerre, M. Helbling, F. Römer, and M. Zobel. Conceptualizing and measuring immigration policies: A comparative perspective. *International Migration Review*, 49(3):555–600, 2015. doi: 10.1111/imre.12100.

B. Bjerregaard and J. K. Cochran. Want amid plenty: Developing and testing a cross-national measure of anomie. *International Journal of Conflict and Violence (IJCV)*, 2(2):182–193, 2008. doi: 10.4119/ijcv-2764.

J. R. Blau and P. M. Blau. The cost of inequality: Metropolitan structure and violent crime. *American sociological review*, pages 114–129, 1982. doi: 10.2307/2095046.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

F. D. Boateng, D. K. Pryce, and J. L. Chenane. I may be an immigrant, but i am not a criminal: Examining the association between the presence of immigrants and crime rates in europe. *Journal of International Migration and Integration*, 22(3):1105–1124, 2021. doi: 10.1007/s12134-020-00790-1.

G. J. Borjas. The economic progress of immigrants. In *Issues in the Economics of Immigration*, pages 15–50. University of Chicago Press, Chicago, 2000.

G. J. Borjas, J. Grogger, and G. H. Hanson. Immigration and the economic status of African-American men. *Economica*, 77(306):255–282, 2010. doi: 10.1111/j.1468-0335.2009.00803.x.

G. Boushey and A. Luedtke. Immigrants across the us federal laboratory: Explaining state-level innovation in immigration policy. *State Politics & Policy Quarterly*, 11(4):390–414, 2011. doi: 10.2307/41575833.

C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery. *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, England, 2019a. doi: 10.1017/9781108644181.

C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery. *Model-based Clustering: Basic Ideas*, chapter 2, pages 15–78. Cambridge University Press, 2019b. doi: 10.1017/9781108644181.

C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery. *Model-based Clustering: Basic Ideas*, chapter 6, pages 163–197. Cambridge University Press, 2019c. doi: 10.1017/9781108644181.

D. Card, C. Dustmann, and I. Preston. Immigration, wages, and compositional amenities. *Journal of the European Economic Association*, 10(1):78–119, 2012. doi: 10.1111/j.1542-4774.2011.01051.x.

S. Castles and A. Davidson. *Citizenship and migration: Globalization and the politics of belonging*. New York: Routledge, 2000. doi: 10.4324/9781003061595.

G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28 (5):781–793, 1995. ISSN 0031-3203. doi: 10.1016/0031-3203(94)00125-6.

C. R. Combei and D. Giannetti. The immigration issue on twitter political communication. Italy 2018-2019. *Comunicazione politica*, 21(2):231–263, 2020. doi: 10.3270/97905.

M. Czaika and H. De Haas. The effectiveness of immigration policies. *Population and Development Review*, 39(3):487–508, 2013. doi: 10.1111/j.1728-4457.2013.00613.x.

P. de Valpine, D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik. Programming with models: writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics*, 26(2):403–413, 2017. doi: 10.1080/10618600.2016.1172487.

P. Dellaportas, D. Karlis, and E. Xekalaki. Bayesian analysis of finite poisson mixtures. *British Journal of Science*, 1(1):96–110, 2011.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

J. Diebolt and C. P. Robert. Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(2):363–375, 1994. doi: 10.1111/j.2517-6161.1994.tb01985.x.

L. Drazanova, T. Liebig, S. Migali, M. Scipioni, and G. Spielvogel. What are Europeans' views on migrant integration?: An in-depth analysis of 2017 Special Eurobarometer "Integration of immigrants in the European Union". *OECD Social, Employment and Migration Working Papers*, 2020. doi: 10.1787/f74bf2f5-en.

P. D'Urso. Dissimilarity measures for time trajectories. *Stat. Methods Appl.*, 9(1-3):53–83, 2000. doi: 10.1007/BF03178958.

C. Dustmann and I. P. Preston. Racial and economic factors in attitudes to immigration. *The BE Journal of Economic Analysis & Policy*, 7(1), 2007.

I. Ehrlich. Participation in illegitimate activities: A theoretical and empirical investigation. *Journal of political Economy*, 81(3):521–565, 1973. doi: 10.1086/260058.

H. Entzinger. The dynamics of integration policies: A multidimensional model. *Challenging immigration and ethnic relations politics: Comparative European perspectives*, pages 97–118, 2000.

E. A. Erosheva. Bayesian estimation of the grade of membership model. *Bayesian statistics*, 7: 501–510, 2003.

B. Everitt and D. Hand. *Finite Mixture Distribution*. Chapman and Hall, London, 1981. doi: 10.1007/978-94-009-5897-5.

S. B. Everitt, S. Landau, M. Leese, and D. Stahl. *Finite Mixture Densities as Models for Cluster Analysis*, chapter 6, pages 143–186. John Wiley & Sons, Ltd, 2011. doi: 10.1002/9780470977811.ch6.

M. Fattore. Synthesis of Indicators: The Non-aggregative Approach. In F. Maggino, editor, *Complexity in Society: From Indicators Construction to their Synthesis*, pages 193–212. Cham: Springer, 2017.

A. Fog. Sampling methods for Wallenius' and Fisher's noncentral hypergeometric distributions. *Communications in Statistics—Simulation and Computation*, 37(2):241–257, 2008. doi: 10.1080/03610910701790236.

A. Fog. *BiasedUrn: Biased Urn Model Distributions*, 2015. URL `https://CRAN.R-project.org/package=BiasedUrn`. R package version 1.07.

E. W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965.

C. Fraley and A. E. Raftery. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, 41(8):578–588, 1998. doi: 10.1093/comjnl/41.8.578.

C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631, 2002. doi: 10.1198/016214502760047131.

C. Fraley, A. E. Raftery, et al. Model-based methods of classification: using the mclust software in chemometrics. *Journal of Statistical Software*, 18(6):1–13, 2007. doi: 10.18637/jss.v018.i06.

M. Freudenber. *Composite Indicators of Country Performance*. Paris: OECD Publishing, 2003. doi: 10.1787/405566708255.

S. Frühwirth-Schnatter. *Finite mixture and Markov switching models*, volume 425. Springer New York, NY, 2006. doi: 10.1007/978-0-387-35768-3.

B. Garcés-Mascareñas and R. Penninx. *Integration processes and policies in Europe: Contexts, levels and actors*. Springer Nature, Cham, 2016. doi: 10.1007/978-3-319-21674-4.

A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016, 2014. doi: 10.1007/s11222-013-9416-2.

J. Gest, A. Boucher, S. Challen, B. Burgoon, E. Thielemann, M. Beine, P. McGovern, M. Crock, H. Rapoport, and M. Hiscox. Measuring and comparing immigration, asylum and naturalization policies across countries: challenges and solutions. *Global Policy*, 5(3):261–274, 2014. doi: 10.1111/1758-5899.12132.

P. Gonnella. Le identità e il carcere: donne, stranieri, minorenni. *Costituzionalismo.it*, 2, 2015.

C. González-Enríquez. Undocumented migration. *Counting the uncountable. Data and trends across Europe, Country Report Spain, Research DG European Commission*, 2009.

S. W. Goodman. Integration requirements for integration's sake? identifying, categorising and comparing civic integration policies. *Journal of Ethnic and Migration Studies*, 36(5):753–772, 2010. doi: 10.1080/13691831003764300.

S. W. Goodman. Conceptualizing and measuring citizenship and integration policy: Past lessons and new approaches. *Comparative Political Studies*, 48(14):1905–1941, 2015. doi: 10.1177/0010414015592648.

S. W. Goodman. Indexing immigration and integration policy: Lessons from europe. *Policy Studies Journal*, 47(3):572–604, 2019. doi: 10.1111/psj.12283.

F. Greco and A. Polli. The political debate on immigration in the election campaigns in Europe. In A. Przegalinska, F. Grippa, and P. A. Gloor, editors, *Collaborative innovation networks conference of Digital Transformation of Collaboration*, pages 111–123. Springer International Publishing, 2019. doi: 10.1007/978-3-030-48993-9_9.

P. J. Green. Introduction to finite mixtures. In *Handbook of Mixture Analysis*, pages 3–20. Chapman and Hall/CRC, 2019. doi: 10.1201/9780429055911.

C. A. Groenendijk, E. Guild, H. Dogan, et al. *Security of residence of long-term migrants: A comparative study of law and practice in European countries*. Strasbourg: Council of Europe, 1998.

A. Gupta and D. Nagar. *Matrix Variate Distributions*. New York: Chapman and Hall/CRC, 1 st edition, 1999. doi: 10.1201/9780203749289.

A. Hadjar and S. Backes. Migration background and subjective well-being a multilevel analysis based on the european social survey. *Comparative Sociology*, 12(5):645–676, 2013.

K. Hagiwara. On the problem in model selection of neural network regression in overrealizable scenario. *Neural Computation*, 14(8):1979–2002, 2002. doi: 10.1162/089976602760128090.

T. Hammar. *Democracy and the nation state : aliens, denizens and citizens in a world of international migration*. Aldershot, UK: Gower Publishing Company, 1990.

N. Harder, L. Figueroa, R. M. Gillum, D. Hangartner, D. D. Laitin, and J. Hainmueller. Multidimensional measure of immigrant integration. *Proceedings of the National Academy of Sciences*, 115(45):11483–11488, 2018. doi: 10.1073/pnas.1808793115.

M. Helbling. Validating integration and citizenship policy indices. *Comparative European Politics*, 11(5):555–576, 2013. doi: 10.1057/cep.2013.11.

M. Helbling and D. Leblang. Controlling immigration? how regulations affect migration flows. *European Journal of Political Research*, 58(1):248–269, 2019. doi: 10.1111/1475-6765.12279.

M. Helbling, L. Bjerre, F. Römer, and M. Zobel. Measuring immigration policies: The impic database. *European Political Science*, 16:79–98, 2017. doi: 10.1057/eps.2016.4.

M. Helbling, S. Simon, and S. D. Schmid. Restricting immigration to foster migrant integration? A comparative study across 22 European countries. *Journal of ethnic and migration studies*, 46(13):2603–2624, 2020. doi: 10.1080/1369183X.2020.1727316.

K. A. Heller, S. Williamson, and Z. Ghahramani. Statistical models for partial membership. In *Proceedings of the 25th International Conference on Machine learning*, pages 392–399, 2008. doi: 10.1145/1390156.1390206.

C. Hennig, M. Meila, F. Murtagh, and R. Rocci. *Handbook of cluster analysis*. New York: Chapman and Hall/CRC, 1 st edition, 2015. doi: 10.1201/b19706.

M. Hooghe and T. Reeskens. Exploring Regimes of Immigrant Integration: Clustering Countries on the Basis of the MIPEX Data. In *Legal Frameworks for the Integration of Third-Country Nationals*, pages 95–112. Brill Nijhoff, Leida, Netherlands, 2009. doi: 10.1163/ej.9789004170698.i-246.23.

D. Ingleby, R. Petrova-Benedict, T. Huddleston, and E. Sanchez. The mipex health strand: a longitudinal, mixed-methods survey of policies on migrant health in 38 countries. *European journal of public health*, 29(3):458–462, 2019.

D. Karlis. An em algorithm for multivariate poisson distribution and related models. *Journal of Applied Statistics*, 30(1):63–77, 2003. doi: 10.1080/0266476022000018510.

D. Karlis and E. Xekalaki. Mixed poisson distributions. *International Statistical Review/Revue Internationale de Statistique*, 73(1):35–58, 2005. doi: 10.2307/25472639.

A. Kerber. Evaluation, Considered as Problem Orientable Mathematics over Lattices. In M. Fattore and R. Brüggemann, editors, *Partial Order Concepts in Applied Sciences*, pages 87–103. Dordrecht: Springer, 2017.

A. Kerber and R. Brüggemann. Problem Driven Evaluation of Chemical Compounds and Its Exploration. *MATCH Commun Math Comput Chem*, 73:577–618, 2015.

R. Koopmans, I. Michalowski, and S. Waibel. Citizenship rights for immigrants: National political processes and cross-national convergence in western europe, 1980–2008. *American journal of sociology*, 117(4):1202–1245, 2012. doi: 10.1086/662707.

R. J. LaLonde and R. H. Topel. Immigrants in the american labor market: Quality, assimilation, and distributional effects. *The American economic review*, 81(2):297–302, 1991. URL `http://www.jstor.org/stable/2006873`.

J. Q. Li and A. R. Barron. Mixture Density Estimation. In *NIPS*, volume 12, pages 279–285, 1999.

F. Maggino. Developing Indicators and Managing the Complexity. In F. Maggino, editor, *Complexity in Society: From Indicators Construction to their Synthesis*, pages 87–114. Cham: Springer, 2017.

F. Maggino and L. S. Alaimo. Complexity and wellbeing: measurement and analysis. In L. Bruni, A. Smerilli, and D. D. Rosa, editors, *A Modern Guide to the Economics of Happiness*, pages 113–128. Cheltenham, UK: Edward Elgar Publishing, 2021.

F. Maggino and L. S. Alaimo. Measuring complex socio-economic phenomena. conceptual and methodological issues. In S. Valaguzza and M. A. Hughes, editors, *Interdisciplinary Approaches to Climate Change for Sustainable Growth*, pages 43–59. Cham: Springer, 2022.

F. Maggino, R. Bruggemann, and L. S. Alaimo. Indicators in the framework of partial order. In R. Bruggemann, L. Carlsen, T. Beycan, C. Suter, and F. Maggino, editors, *Measuring and Understanding Complex Phenomena*, pages 17–29. Cham: Springer, 2021.

R. Maitra. Initializing partition-optimization algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(1):144–157, 2009. doi: 10.1109/TCBB.2007.70244.

J.-M. Marin, K. Mengersen, and C. P. Robert. *Bayesian Modelling and Inference on Mixtures of Distributions*, volume 25 of *Handbook of Statistics*, pages 459–507. Elsevier, 2005. doi: 10.1016/S0169-7161(05)25016-2.

D. S. Massey, J. Arango, G. Hugo, A. Kouaouci, and A. Pellegrino. *Worlds in motion: understanding international migration at the end of the millennium.* Oxford, Clarendon Press, 1998.

G. Mastrobuoni and P. Pinotti. Legal status and the criminal activity of immigrants. *American Economic Journal: Applied Economics*, 7(2):175–206, 2015. doi: 10.1257/app.20140039.

A. M. Mayda. Who is against immigration? A cross-country investigation of individual attitudes toward immigrants. *The review of Economics and Statistics*, 88(3):510–530, 2006. doi: 10.1162/rest.88.3.510.

G. J. McLachlan and K. E. Basford. *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker, New York, 1988.

P. D. Mcnicholas and T. B. Murphy. Model-based clustering of longitudinal data. *Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 38(1):153–168, 2010. doi: 10.2307/27805221.

V. Melnykov and I. Melnykov. Initializing the em algorithm in gaussian mixture models with an unknown number of components. *Computational Statistics & Data Analysis*, 56(6):1381–1395, 2012. ISSN 0167-9473. doi: 10.1016/j.csda.2011.11.002.

V. Melnykov, R. Maitra, et al. Finite mixture models and model-based clustering. *Statistics Surveys*, 4:80–116, 2010. doi: DOI:10.1214/09-SS053.

R. K. Merton. Social structure and anomie. *American Sociological Review*, 3(5):672–682, 1938.

R. B. Millar. Conditional vs marginal estimation of the predictive loss of hierarchical models using waic and cross-validation. *Statistics and Computing*, 28(2):375–385, 2018. doi: 10.1007/s11222-017-9736-8.

M. Nardo, M. Saisana, A. Saltelli, and S. Tarantola. Tools for composite indicators building. *European Commission, Ispra*, 15(1):19–20, 2005.

J. Niessen and T. Huddleston. *Legal frameworks for the integration of third-country nationals.* Leiden, The Netherlands: Brill| Nijhoff, 2009. doi: 10.1163/ej.9789004170698.i-246.

J. Niessen, T. Huddleston, L. Citron, A. Geddes, and D. Jacobs. Migrant integration policy index. Technical report, Brussels: British Council and Migration Policy Group, 2007.

OECD. Handbook on Constructing Composite Indicators. Methodology and User Guide, 2008.

P. Papastamoulis and G. Iliopoulos. An artificial allocations based solution to the label switching problem in bayesian analysis of mixtures of distributions. *Journal of Computational and Graphical Statistics*, 19(2):313–331, 2010. doi: 10.2307/25703571.

E. Pearson. Some aspects of the geometry of statistics: the use of visual presentation in understanding the theory and application of mathematical statistics. *Journal of the Royal Statistical Society. Series A (General)*, 119(2):125–146, 1956. doi: 10.2307/2342880.

K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

P. J. Pedersen, M. Pytlikova, and N. Smith. Selection and network effects—migration flows into oecd countries 1990–2000. *European Economic Review*, 52(7):1160–1186, 2008.

D. Peel and G. MacLahlan. *Finite mixture models*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, 2000. doi: 10.1002/0471721182.

R. Penninx. Problems of and solutions for the study of immigrant integration. *Comparative Migration Studies*, 7(1):1–11, 2019. doi: 0.1186/s40878-019-0122-x.

R. Penninx and M. Martiniello. *Integration processes and policies: State of the art and lessons*, pages 139–164. Aldershot, UK: Ashgate, 1 st edition, 2004.

J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly. Association mapping in structured populations. *The American Journal of Human Genetics*, 67(1):170–181, 2000. doi: 10.1086/302959.

B. Rainer and H. Marc. Which Indicators are Most Useful for Comparing Citizenship Policies? EUI-RSCAS Working Papers 54, European University Institute (EUI), Robert Schuman Centre of Advanced Studies (RSCAS), 2011.

K. S. Ramakrishnan. Incorporation versus assimilation. *Outsiders No More?: Models of Immigrant Political Incorporation*, page 27, 2013. doi: 10.1093/acprof:oso/9780199311316.003.0002.

A. Rau, C. Maugis-Rabusseau, M.-L. Martin-Magniette, and G. Celeux. Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics*, 31(9):1420–1427, 01 2015. doi: 10.1093/bioinformatics/btu845.

G. Rayp, I. Ruyssen, and S. Standaert. Measuring and explaining cross-country immigration policies. *World Development*, 95:141–163, 2017. doi: 10.1016/j.worlddev.2017.02.026.

S. Richardson and P. J. Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792, 1997. doi: 10.1111/1467-9868.00095.

J. Rousseau, C. Grazian, and J. E. Lee. *Bayesian mixture models: Theory and methods*, chapter 4, pages 53–72. Chapman and Hall/CRC, 2019. doi: 10.1201/9780429055911.

D. Ruedin. Increasing validity by recombining existing indices: Mipex as a measure of citizenship models. *Social Science Quarterly*, 96(2):629–638, 2015.

S. Sarkar, X. Zhu, V. Melnykov, and S. Ingrassia. On parsimonious models for modeling matrix data. *Computational Statistics & Data Analysis*, 142:106822, 2020. doi: 10.1016/j.csda.2019.106822.

L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R journal*, 8(1):289, 2016. doi: 10.32614/RJ-2016-021.

N. B. Simpson. Demographic and economic determinants of migration. *IZA World of Labor*, 2017.

S.-E. Skaaning. Measuring the rule of law. *Political Research Quarterly*, 63(2):449–460, 2010. doi: 10.1177/1065912909346745.

G. Solano and D. De Coninck. Explaining migrant integration policies: A comparative study across 56 countries. *Migration Studies*, 2022. doi: 10.1093/migration/mnac036.

G. Solano and T. Huddleston. Migrant Integration Policy Index 2020. *Barcelona Center for International Affairs (CIDOB)*, 2020.

G. Solano and T. Huddleston. Beyond immigration: Moving from western to global indexes of migration policy. *Global Policy*, 12(3):327–337, 2021. doi: 10.1111/1758-5899.12930.

G. Solano and T. Huddleston. Migration policy indicators. In *Introduction to Migration Studies*, pages 389–407. Springer, Cham, 2022. doi: 10.1007/978-3-030-92377-8_24.

L. M. Solivetti. Immigration, socio-economic conditions and crime: A cross-sectional versus cross-sectional time-series perspective. *Quality & Quantity*, 52(4):1779–1805, 2018. doi: 10.1007/s11135-017-0566-8.

M. Sperrin, T. Jaki, and E. Wit. Probabilistic relabelling strategies for the label switching problem in bayesian mixture models. *Statistics and Computing*, 20(3):357–366, 2010. doi: 10.1007/s11222-009-9129-8.

Standard Eurobarometer. EUROBAROMETER 97 - Summer 2022, 2022.

M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000. doi: 10.1111/1467-9868.00265.

D. M. Titterington, S. Afm, A. F. Smith, U. Makov, et al. *Statistical analysis of finite mixture distributions*, volume 198. John Wiley & Sons Incorporated, New York, 1985.

S. D. Tomarchio, S. Ingrassia, and V. Melnykov. Modelling students' career indicators via mixtures of parsimonious matrix-normal distributions. *Australian & New Zealand Journal of Statistics*, 2022. doi: 10.1111/anzs.12351.

UNDESA. International Migrant Stock 2020, 2020a. United Nations Department of Economic and Social Affairs - Population Division `https://www.un.org/development/desa/pd/content/international-migrant-stock`.

UNDESA. International migration 2020 highlights, 2020b.

Undp. *Human development report 1997*. Oxford University, 1997.

V. Viallefont, S. Richardson, and P. J. Green. Bayesian analysis of poisson mixtures. *Journal of Nonparametric Statistics*, 14(1-2):181–202, 2002. doi: 10.1080/10485250211383.

C. Viroli. Model based clustering for three-way data structures. *International Society for Bayesian Analysis*, 6(4):573–602, 2011a. doi: 10.1214/11-BA622.

C. Viroli. Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, 21(4):511–522, 2011b. doi: 10.1007/s11222-010-9188-x.

H. Waldrauch and C. Hofinger. An index to measure the legal obstacles to the integration of migrants. *Journal of Ethnic and Migration Studies*, 23(2):271–285, 1997. doi: 10.1080/1369183X.1997.9976590.

Y. Wang and V. Melnykov. On variable selection in matrix mixture modelling. *Stat*, 9(1):e278, 2020. doi: 10.1002/sta4.278.

S. Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13 (4):899–933, 2001. doi: 10.1162/089976601300014402.

S. Watanabe and M. Opper. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12), 2010.

A. White and T. B. Murphy. Exponential family mixed membership models for soft clustering of multivariate data. *Advances in Data Analysis and Classification*, 10(4):521–540, 2016. doi: 10.1007/s11634-016-0267-5.

M. A. Woodbury, J. Clive, and A. Garson Jr. Mathematical typology: a grade of membership technique for obtaining disease definition. *Computers and biomedical research*, 11(3):277–298, 1978. doi: 10.1016/0010-4809(78)90012-5.

WPB. Persons held in prison, 2020. World Prison Brief `https://www.prisonstudies.org/world-prison-brief-data`.

X. Zhu, S. Sarkar, and V. Melnykov. MatTransMix: an R Package for Matrix Model-Based Clustering and Parsimonious Mixture Modeling. *Journal of Classification*, 39:147—170, 2022. doi: 10.1007/s00357-021-09401-9.

G. Zincone, R. Pennix, and M. Borkert. Migration policymaking in europe: The dynamics of actors and contexts in past and present. *Migration Policymaking in Europe*, 2011. doi: 10.2307/j.ctt46n178.

Appendix A

Table .1. MMN clusters' means over time.

Cluster 1	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
Labour	42.34	42.38	43.62	45.40	45.91	47.02
Family	65.96	66.00	67.24	69.02	69.53	70.64
Education	46.68	46.72	47.96	49.74	50.25	51.36
Politics	19.14	19.18	20.42	22.21	22.72	23.83
Residence	71.85	71.89	73.13	74.92	75.43	76.54
Citizenship	16.64	16.68	17.92	19.71	20.22	21.33
Anti-discrimination	66.12	66.16	67.40	69.19	69.70	70.81
Cluster 2	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
Labour	48.34	48.34	48.50	48.50	48.56	48.56
Family	64.09	64.09	64.25	64.25	64.31	64.31
Education	39.51	39.51	39.67	39.67	39.73	39.73
Politics	32.52	32.52	32.68	32.68	32.74	32.74
Residence	66.79	66.79	66.95	66.95	67.00	67.00
Citizenship	49.45	49.45	49.60	49.60	49.66	49.66
Anti-discrimination	71.20	71.20	71.36	71.36	71.42	71.42
Cluster 3	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
Labour	51.65	52.37	52.37	52.95	53.41	53.53
Family	49.99	50.71	50.71	51.29	51.75	51.87
Education	44.42	45.14	45.14	45.71	46.18	46.30
Politics	39.46	40.18	40.18	40.75	41.22	41.34
Residence	58.94	59.65	59.65	60.23	60.70	60.81
Citizenship	49.02	49.74	49.74	50.31	50.78	50.90
Anti-discrimination	65.39	66.11	66.11	66.68	67.15	67.27
Cluster 4	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**

Labour	38.85	39.28	41.76	43.25	43.63	44.56
Family	46.45	46.87	49.35	50.84	51.22	52.16
Education	28.89	29.32	31.80	33.29	33.67	34.60
Politics	11.13	11.56	14.03	15.53	15.91	16.84
Residence	50.61	51.04	53.51	55.01	55.39	56.32
Citizenship	37.59	38.02	40.49	41.99	42.36	43.30
Anti-discrimination	67.18	67.61	70.09	71.58	71.96	72.89
Cluster 5	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
Labour	52.29	51.65	53.22	54.82	54.03	53.85
Family	62.32	61.69	63.26	64.85	64.06	63.89
Education	34.76	34.13	35.69	37.29	36.50	36.32
Politics	40.98	40.34	41.91	43.51	42.72	42.54
Residence	61.75	61.12	62.69	64.28	63.50	63.32
Citizenship	65.94	65.30	66.87	68.46	67.68	67.50
Anti-discrimination	74.32	73.68	75.25	76.85	76.06	75.88

Appendix B

Table .2. PM model. Players' profiles membership

	Goalkeepers	Defenders	Full back and midfielders	Pure strikers
Tammy Abraham	0.005	0.079	0.187	0.729
Francesco Acerbi	0.037	0.461	0.463	0.038
Kevin Agudelo	0.004	0.149	0.647	0.200
Emanuel Aiwum	0.070	0.614	0.278	0.039
Luis Alberto	0.013	0.037	0.861	0.089
Kelvin Amian	0.018	0.430	0.460	0.092
Bruno Amione	0.028	0.720	0.045	0.207
Ethan Ampadu	0.002	0.476	0.517	0.004
Sofyan Amrabat	0.035	0.215	0.749	0.002
Felipe Anderson	0.001	0.045	0.682	0.272
Emil Audero	0.806	0.044	0.090	0.060
Tommaso Augello	0.064	0.029	0.765	0.141
Filippo Bandinelli	0.024	0.141	0.655	0.180
Nicolò Barella	0.002	0.007	0.908	0.083
Federico Baschirotto	0.002	0.973	0.022	0.002
Alessandro Bastoni	0.039	0.193	0.762	0.006
Rodrigo Becão	0.006	0.578	0.390	0.025
Ismaël Bennacer	0.058	0.019	0.915	0.008
Beto	0.003	0.240	0.008	0.750
Matteo Bianchetti	0.024	0.836	0.051	0.089
Jaka Bijol	0.003	0.931	0.019	0.048
Cristiano Biraghi	0.027	0.003	0.963	0.008
Alexis Blin	0.006	0.615	0.370	0.009
Giacomo Bonaventura	0.002	0.055	0.588	0.355

Mehdi Bourabia	0.017	0.168	0.741	0.073
Domagoj Bradarić	0.087	0.137	0.515	0.260
Gleison Bremer	0.002	0.861	0.009	0.129
Dylan Bronn	0.042	0.623	0.294	0.041
Alessandro Buongiorno	0.017	0.736	0.223	0.024
Luca Caldirola	0.086	0.592	0.235	0.087
Hakan Çalhanoğlu	0.003	0.004	0.935	0.058
Andrea Cambiaso	0.041	0.181	0.580	0.199
Antonio Candreva	0.011	0.010	0.793	0.186
Gianluca Caprari	0.015	0.014	0.519	0.451
Carlos	0.001	0.232	0.573	0.193
Marco Carnesecchi	0.682	0.012	0.281	0.025
Nicolò Casale	0.044	0.652	0.162	0.143
Danilo Cataldi	0.106	0.219	0.671	0.004
Patrick Ciurria	0.007	0.086	0.543	0.364
Omar Colley	0.089	0.831	0.013	0.067
Andrea Consigli	0.735	0.070	0.121	0.074
Lassana Coulibaly	0.022	0.276	0.622	0.080
Bryan Cristante	0.003	0.367	0.627	0.002
Juan Cuadrado	0.051	0.025	0.621	0.303
Flavius Daniliuc	0.018	0.778	0.106	0.098
Danilo	0.001	0.375	0.623	0.001
Matteo Darmian	0.074	0.225	0.625	0.077
Merih Demiral	0.069	0.796	0.006	0.129
Fabio Depaoli	0.048	0.210	0.523	0.219
Cyriel Dessers	0.024	0.201	0.020	0.755
Federico Di Francesco	0.029	0.117	0.326	0.529
Michele Di Gregorio	0.801	0.173	0.014	0.012
Giovanni Di Lorenzo	0.002	0.050	0.946	0.002
Boulaye Dia	0.015	0.108	0.261	0.616
Brahim Díaz	0.009	0.012	0.629	0.350
Federico Dimarco	0.017	0.003	0.731	0.248
Koffi Djidji	0.037	0.658	0.296	0.010
Dodô	0.043	0.170	0.709	0.079
Nicolás Domínguez	0.007	0.222	0.637	0.134
Bartlomiej Dragowski	0.731	0.092	0.140	0.037

Denzel Dumfries	0.017	0.071	0.470	0.442
Paulo Dybala	0.001	0.002	0.552	0.446
Edin Džeko	0.001	0.101	0.336	0.562
Éderson	0.002	0.318	0.507	0.173
Martin Erlic	0.075	0.899	0.012	0.014
Wladimiro Falcone	0.757	0.196	0.021	0.026
Lewis Ferguson	0.003	0.311	0.340	0.346
Davide Frattesi	0.001	0.112	0.658	0.229
Manolo Gabbiadini	0.001	0.245	0.042	0.713
Antonino Gallo	0.046	0.159	0.789	0.006
Valentin Gendrey	0.007	0.431	0.557	0.005
Olivier Giroud	0.002	0.220	0.015	0.764
Joan Gonzàlez	0.015	0.382	0.450	0.153
Emmanuel Gyasi	0.015	0.285	0.154	0.546
Hans Hateboer	0.044	0.335	0.353	0.268
Theo Hernández	0.003	0.034	0.723	0.241
Isak Hien	0.063	0.698	0.207	0.031
Morten Hjulmand	0.003	0.380	0.615	0.001
Emil Holm	0.034	0.263	0.407	0.296
Roger Ibanez	0.001	0.681	0.311	0.007
Igor	0.065	0.583	0.339	0.013
Jonathan Ikone	0.024	0.013	0.403	0.559
Ciro Immobile	0.008	0.145	0.065	0.782
Ardian Ismajli	0.009	0.927	0.006	0.058
Armando Izzo	0.019	0.594	0.373	0.014
Pierre Kalulu	0.003	0.423	0.571	0.003
Jakub Kiwior	0.017	0.769	0.033	0.181
Teun Koopmeiners	0.002	0.024	0.899	0.075
Filip Kostić	0.002	0.003	0.685	0.310
Christian Kouamé	0.023	0.015	0.457	0.505
Khvicha Kvaratskhelia	0.000	0.001	0.483	0.516
Kevin Lasagna	0.013	0.241	0.017	0.730
Armand Lauriente	0.000	0.002	0.584	0.414
Darko Lazović	0.016	0.021	0.597	0.367
Manuel Lazzari	0.075	0.235	0.483	0.208
Rafael Leão	0.000	0.001	0.573	0.426

Mehdi Léris	0.002	0.322	0.442	0.234
Karol Linetty	0.037	0.306	0.523	0.134
Stanislav Lobotka	0.024	0.197	0.778	0.002
Manuel Locatelli	0.002	0.346	0.646	0.005
Ademola Lookman	0.001	0.002	0.465	0.532
Maxime Lopez	0.043	0.187	0.767	0.003
Sandi Lovrić	0.003	0.088	0.635	0.273
Jhon Lucumí	0.015	0.698	0.276	0.011
Sebastiano Luperto	0.003	0.933	0.043	0.021
Gianluca Mancini	0.017	0.476	0.474	0.034
Rolando Mandragora	0.032	0.067	0.753	0.148
Gian Marco Ferrari	0.054	0.706	0.225	0.014
Pablo Marí	0.065	0.798	0.124	0.014
Răzvan Marin	0.036	0.026	0.778	0.160
Lautaro Martínez	0.000	0.004	0.447	0.550
Lucas Martínez Quarta	0.001	0.595	0.399	0.005
Adam Marušić	0.016	0.384	0.588	0.012
Nemanja Matić	0.049	0.198	0.748	0.005
Pasquale Mazzocchi	0.064	0.012	0.579	0.345
Soualiho Meïté	0.059	0.297	0.454	0.190
Alex Meret	0.960	0.011	0.006	0.023
Nikola Milenković	0.009	0.631	0.217	0.143
Sergej Milinković-Savić	0.001	0.121	0.820	0.058
Vanja Milinković-Savić	0.543	0.033	0.421	0.003
Kim Min-jae	0.001	0.488	0.510	0.001
Aleksei Miranchuk	0.032	0.007	0.573	0.388
Henrikh Mkhitaryan	0.006	0.149	0.680	0.164
Lorenzo Montipò	0.590	0.061	0.344	0.005
Juan Musso	0.933	0.022	0.009	0.036
Joakim Mæhle	0.029	0.058	0.680	0.233
Dimitris Nikolaou	0.029	0.556	0.410	0.004
M'Bala Nzola	0.002	0.140	0.225	0.634
David Okereke	0.003	0.075	0.404	0.518
André Onana	0.687	0.015	0.129	0.169
Riccardo Orsolini	0.002	0.018	0.496	0.484
Victor Osimhen	0.001	0.041	0.012	0.947

Fabiano Parisi	0.002	0.232	0.755	0.012
Rui Patrício	0.928	0.021	0.018	0.033
Pedro	0.011	0.134	0.357	0.498
Lorenzo Pellegrini	0.003	0.004	0.738	0.256
Roberto Pereyra	0.002	0.013	0.750	0.235
Nehuén Pérez	0.002	0.571	0.424	0.003
Matteo Pessina	0.004	0.189	0.798	0.009
Krzysztof Piatek	0.003	0.259	0.068	0.671
Charles Pickel	0.011	0.382	0.529	0.077
Andrea Pinamonti	0.006	0.216	0.018	0.760
Stefan Posch	0.002	0.301	0.687	0.010
Ivan Provedel	0.648	0.106	0.166	0.079
Adrien Rabiot	0.001	0.186	0.640	0.173
Arkadiusz Reca	0.023	0.200	0.600	0.177
Samuele Ricci	0.048	0.257	0.552	0.143
Tomás Rincón	0.056	0.246	0.689	0.009
Ricardo Rodríguez	0.052	0.201	0.741	0.007
Rogério	0.030	0.126	0.836	0.008
Alessio Romagnoli	0.045	0.816	0.126	0.013
Marten de Roon	0.007	0.395	0.594	0.004
Amir Rrahmani	0.010	0.579	0.383	0.028
Mário Rui	0.039	0.005	0.951	0.005
Antonio Sanabria	0.012	0.226	0.023	0.738
Alex Sandro	0.071	0.324	0.588	0.017
Martin Satriano	0.017	0.278	0.027	0.678
Giorgio Scalvini	0.008	0.479	0.509	0.004
Jerdy Schouten	0.028	0.450	0.516	0.006
Perr Schuurs	0.012	0.817	0.019	0.152
Stefano Sensi	0.056	0.072	0.856	0.016
Luigi Sepe	0.962	0.010	0.007	0.021
Leonardo Sernicola	0.030	0.163	0.588	0.220
Marco Silvestri	0.743	0.036	0.046	0.175
Wilfried Singo	0.012	0.257	0.325	0.406
Łukasz Skorupski	0.924	0.032	0.015	0.030
Milan Škriniar	0.057	0.524	0.411	0.008
Chris Smalling	0.003	0.921	0.007	0.070

Adama Soumaoro	0.111	0.753	0.021	0.115
Petar Stojanović	0.053	0.223	0.640	0.084
Gabriel Strefezza	0.001	0.025	0.633	0.341
Isaac Success	0.043	0.011	0.412	0.534
Wojciech Szczesny	0.932	0.014	0.010	0.044
Adrien Tameze	0.028	0.384	0.577	0.011
Ciprian Tatarusanu	0.615	0.009	0.024	0.352
Pietro Terracciano	0.785	0.135	0.043	0.036
Jeremy Toljan	0.072	0.185	0.736	0.007
Rafael Tolói	0.001	0.407	0.588	0.003
Fikayo Tomori	0.007	0.631	0.354	0.009
Sandro Tonali	0.012	0.063	0.778	0.146
Iyenoma Udogie	0.004	0.079	0.721	0.196
Emanuele Valeri	0.016	0.009	0.744	0.232
Guglielmo Vicario	0.756	0.092	0.104	0.048
Tonny Vilhena	0.031	0.180	0.514	0.275
Dušan Vlahović	0.005	0.192	0.039	0.764
Nikola Vlašić	0.002	0.004	0.607	0.387
Walace	0.001	0.322	0.675	0.002
Mattia Zaccagni	0.001	0.043	0.501	0.455
Nicola Zalewski	0.028	0.174	0.525	0.273
Anguissa	0.001	0.189	0.802	0.008
Piotr Zieliński	0.003	0.003	0.734	0.261
Filip Duricic	0.043	0.030	0.590	0.337

Appendix C

Measuring Human Development by means of composite indicators: open issues and new methodological tools

Abstract

Over the years, the Human Development Index has become a reference measure of quality of life and well-being. Its growing importance has been accompanied by a lively debate in the literature concerning the pros and cons of this index. Many works have attempted to provide solutions to Human Development Index related problems. In this paper, we will focus on some of these problems, which are typical not only for the measurement of human development, but for the construction of composite indicators. We will try to provide an answer by proposing two new methodological tools, the $Min - BoD$ interval of synthesis and the mid aggregation point, which present interesting potentialities to be used in empirical analyses and for policy evaluations, not only in the human development measurement. The proposed tool have been applied to the Human Development Index data collected for 189 countries in 2019.

1. Introduction

The Human Development Index - HDI has become over the years one of the main indicators for measuring wellbeing, linked to the so-called *Beyond GDP* debate (see, for instance: Bleys, 2012; Boarini et al., 2006; D'Urso et al., 2022), focusing on developing indicators that are more inclusive of environmental and social aspects than Gross Domestic Product (GDP) is. The reasons for the success of HDI are manifold. Certainly, the simplicity with which it is calculated together with the transparency of data published by international organizations have made

it easily understandable to all, academics and policy makers alike. It is, indeed, a composite indicator that measures the average achievement in a country in three basic dimensions of human development: health, education, and standard of living (Harttgen & Klasen, 2012). However, there has also been much criticism of this measure, both conceptual and methodological (Anand & Sen, 1997; Sagar & Najam, 1998; Alkire, 2002).

The measurement process in social sciences is associated with the construction of systems of indicators. Indicators within a system are not simple collections of measures; they are interconnected. A system of indicators allows the measurement of a complex concept that would not otherwise be measurable by taking into account the indicators individually. They play a key role in describing and understanding complex socio-economic phenomena. The complex nature of systems of indicators requires approaches allowing more concise views in order to analyse and understand them. The guiding concept is *synthesis*. Synthesising data responds to a range of cognitive and practical needs and it is related to the need for "reduction" in knowledge of complex phenomena.

The synthesis of indicators' systems has become a main issue in the literature. A variety of statistical methods useful for this purpose have been defined and used. From a technical perspective, these methods can be classified into two different approaches: the aggregative-compensative (OECD, 2008) and the non-aggregative (Brüggemann & Patil, 2011; Fattore, 2017; Alaimo et al., 2021; Maggino et al., 2021). In this paper, we focus on the previous one. Methods belonging to the first approach, known as composite indicators (CIs), constitute the dominant framework in the literature and they have been widely used. Within this approach, a variety of methods and applications have been proposed over the years (see Section 3 for details). Despite its success, the aggregative-compensative

approach has been criticised and a series of conceptual and methodological issues have been posed. These questions are still open and inflame the debate in the literature on this topic. Obviously, HDI is constructed using such an approach and, consequently, its problems also affected this measure. Starting from the presentation and analysis of some of these open questions, the aim of this paper is twofold. On the one hand, we present and discuss two new methodological tools, by explaining in detail the different steps to apply them. Specifically, an interval of synthesis, in which for each unit we identify an upper bound and a lower bound representing the best and the worst performance that each unit could obtain by aggregating a system of indicators, and starting from that interval, we propose the mid aggregation point (*map*) as a new synthetic measure. On the other hand, the application of these tools to the HDI is interesting in itself, providing a measure and an analysis of human development in countries.

The paper is structured as follows. In Section 2, we report a brief literature review about the main criticisms on HDI. Section 3 presents some issues typical of CIs and the research questions addressed in this work. In Section 4.1 we present the main aspect of interval of synthesis. Section 4.2 reports the main characteristics of the mid aggregation point. In Section 5 we present an application to Human Development Index (HDI) data and Section 6 concludes.

2. Human Development Index: a brief analysis of critical literature

The concept of human development has become over the years a reference in the so-called *Beyond GDP* debate. The United Nations Programme for Development (UNDP) played a central role in this. With its Human Development Reports (HDRs), it has put the concept of human development at the center of intellectual, academic and, albeit to a lesser extent, policy makers' debate.

Moreover, UNDP gave a preliminary framework for defining and measuring this concept by means of the HDI. Since the publication of the first HDR (UNDP, 1990), there has been a heated debate about this index and how to calculate it. Criticisms generally regard two aspects: the definition of human development, its components and determinants; the ways in which the different indicators could be aggregated to obtain HDI. Over the years, UNDP has attempted to partially address some of these issues, by substituting, for instance, those variables that approach the achievements in education and material wellbeing (UNDP, 2010) or with the use of the geometric mean instead of the arithmetic one (Herrero et al., 2012). However, the changes made over the years have mostly been minor adjustments to the composite's calculation rather than real conceptual and methodological advances.

Human development is a complex phenomenon that requires an approach capable of grasping their complex and multidimensional nature (Alaimo, 2021b,a). It is made up from a network of elements, which interact both with one another and with the environment. Evaluating the achievement in terms of human development requires, consequently, the construction of multidimensional indicators. This poses the question of identifying the most relevant dimensions which ensure a solid epistemological and empirical basis for the concept of human development (Alkire, 2002). HDI is based on Sen's capabilities-functionings approach (Sen, 1999) and considers three dimensions: a long and healthy life, knowledge and a decent standard of living. Some scholars criticised this framework, considering it inadequate and too simplistic (Gasper, 2002) or incomplete (Anand & Sen, 1997, 2000a; Nussbaum, 2000; Ranis et al., 2006; Chhibber & Laajaj, 2007; Biggeri & Mauro, 2018). However, on this point it should be pointed out that the measurement of a complex phenomenon needs to consider different but not necessarily

all dimensions. HDI is based on a very robust conceptual definition (Sen, 1999) which gives meaning and relevance to the identified dimensions and according to which they are sufficient for the definition of human development. "The concept of human development is broader than any measure of human development. Thus although the HDI is a constantly evolving measure, it will never perfectly capture human development in its full sense" (UNDP, 1993, 104). The objective of HDI is not to provide a comprehensive measure of human development and wellbeing, but an alternative to purely economic ones (Kovacevic, 2010). The capabilities approach is a partial theory of well-being that does not claim to provide a complete description of all the components of a good life (Klugman, 2009). It is not the purpose of this paper to critically analyse the HDI framework. Consequently, we consider it valid and adequate.

Assuming that human development can be measured by means of the three dimensions considered above, another controversial point is related to the selection of basic indicators (Anand & Sen, 2000b; Hicks, 1997; Foster et al., 2005; Herrero et al., 2012)[1]. After the changes introduced in 2010 (UNDP, 2010), the health dimension is assessed by the life expectancy at birth (LEB); the education dimension is measured by means of the mean years of schooling for adults aged 25 years and more (MYS) and the expected years of schooling for children of school entering age (EYS) and the standard of living dimension is measured by the gross national income per capita (GNI)[2]. The choice of elementary indicators is one

[1]In particular, some authors focused on the lack of concern for distributive issues (Anand & Sen, 2000b), emphasising the need to identify methods capable to incorporate distributional inequalities of income, education, and longevity (Hicks, 1997; Foster et al., 2005). In order to try to overcome this limitation, in 2010 UNDP introduced the Inequality-adjusted Human Development Index - IHDI, a measure that accounts for inequality in the society, following the preliminary analyses made by Alkire & Foster (2010), and that is obtained by combining the estimate of the basic HDI to the Atkinson measure of inequality (Atkinson, 1970).

[2]Three of the four indicators were revised: GDP per capita was replaced by GNI per capita

of the crucial points in the construction of a synthetic measure. A clarification is needed. In the case of human development, we are dealing with a concept that must be measured through a formative measurement model. The debate on measurement models refers to the relationship between constructs and indicators. Two different conceptual approaches can be identified: *reflective* and *formative* (Blalock, 1964; Diamantopoulos & Winklhofer, 2001; Diamantopoulos & Siguaw, 2006; Diamantopoulos et al., 2008; Alaimo, 2022)[3]. In a formative model, indicators are causes of the construct rather than its effects (like in the reflective one) and they determine the latent variable giving it its meaning (Blalock, 1964, 1968). Accordingly, indicators are not interchangeable: omitting an indicator is omitting part of the construct (Bollen, 1984). Thus, the choice of indicators determines what we want to measure. This clarification allows us to make it clear that changing the elementary indicators selected to measure HDI would be tantamount to changing the concept of human development. For this reason, in this paper we will use the same indicators selected by UNDP.

The methodological choices for the construction of the HDI were another element of criticism. In particular, the methods of normalisation, weighting and aggregation have been and continue to be among the most controversial issues.

Since UNDP (1994)[4], the normalisation method used has been a Min-Max including the use of fixed goalposts (the procedure is illustrated in Section 5).

(both valued in PPP US), literacy and gross enrolments were replaced by mean years of schooling and expected years of schooling.

[3]We need to clarify that, as shown in different studies (see, for instance, Edwards & Bagozzi (2000); Bollen (2007); Bollen & Diamantopoulos (2017)), the choice of the measurement model only depends on the nature of the latent variable, the appropriateness to the phenomenon to be measured and the direction of relationships between constructs and measures. It is not a personal choice of the researcher.

[4]Until 1993, the normalisation method used was a Min-Max, with minimum and maximum values for all three components based on variable criteria, like the actual minimum and maximum in the current year, or an average threshold value, as with income (Stanton, 2007). This choice was strongly criticised in the literature (Kelley, 1991; McGillivray, 1991; McGillivray & White,

The choice of normalisation has significant implications for the index values and rankings, as shown in different papers (Klugman, 2009; Kovacevic, 2010). Another critique regards the choice of equal weights for all the components. At the basis of this choice, there is the consideration that the three dimensions are equally important in determining the level of human development (Sagar & Najam, 1998). Indeed, giving the same weight does not mean "not-weighting", but giving all indicators the same importance. Different weighting methods have been proposed in the literature, such as equal weighting, principal component analysis, experts' judgements (Slottje, 1991; Paul, 1996; Mazumdar, 2003). However, it was found that by using different weighting systems for the calculation of the HDI, the resulting rankings of the synthetic measures obtained were very similar (UNDP, 1993; Noorbakhsh, 1998; Biswas & Caliendo, 2002). Generally, all weighting schemes are questionable and controversial (Anand & Sen, 1997); the choice of equal weights were justified "on the simple premise that all these choices were very important and that there was no a-priori rationale for giving a higher weight to one choice than to another" (Ul Haq, 1995, 48). It should be pointed out that these are critical aspects not only of the HDI, but generally of composite indicators. There is no "perfect" method of normalisation, just as no agreed methodology exists to weight basic indicator. These choices have a large impact on the values of the composites, influencing their results and, consequently, the interpretation of the phenomenon. Thus, we feel able to affirm that the choices made by UNDP are, albeit questionable, shareable and acceptable and will be considered as such in this paper. Certainly, one of the most controversial issues is the choice of the aggregation method, i.e. the type of mathematical function to

1993; Doessel & Gounder, 1994; Paul, 1996), mainly because it did not allow any temporal comparability.

be used to aggregate the previously normalised indicators. Starting from UNDP (2010)[5], the dimensional indices were aggregated with a geometric mean, in order to overcome the limitations of the arithmetic mean. In this way, HDI attains a compromise by adopting a functional form (the geometric mean) that is between the extremes of perfect substitutability and perfect complementarity (Klugman, 2009). Compensability is a main issue in composite indicators construction. This is a much more conceptual than methodological question, which often has no clear or unambiguous answer. The choice of the geometric mean for HDI was considered an advancement, because it made it possible to overcome the problem of full compensability. But at the same time it poses other problems. For instance, while we know that the arithmetic mean is fully compensatory, we do not know how much the geometric mean is[6]. In general, the choice of geometric mean is also arbitrary, though more or less agreeable. Other authors (Diener & Suh, 1997) criticised the construction of a single composite index, pointing out that a more appropriate choice would be to use a dashboard. This is an open issue in the literature and we can find arguments in favour of the composite indices and against them[7]. In this paper, we focus on the latter two questions, the compensability and the possibility of approaching synthesis in a different way than by constructing a single composite indicator. These are open issues typical

[5]Until 2010, the HDI was calculated as an unweighted arithmetic mean of the normalised elementary indicators. This choice has been strongly criticised in the literature (Desai, 1991; Palazzi & Lauri, 1998; Sagar & Najam, 1998), because it implied perfect substitutability allowing that a deficit in one dimension could be compensated by a surplus in another.

[6]The more variable the distribution, the smaller the geometric mean with respect to the arithmetic mean. Consequently, if two units have the same arithmetic mean in the elementary indicators, the unit with higher variability will have a geometric mean smaller than that of the unit with lower variability.

[7]For instance, a dashboard allows to avoid an arbitrary choice of the functional form and the weighting scheme and to observe a phenomenon from multiple points of view, However, it does not allow a simple and direct understanding of the phenomenon under consideration (Saisana & Tarantola, 2002; OECD, 2008).

of the synthesis, in particular of the aggregative-compensation approach.

3. Composite indicators: some conceptual and methodological open issues

In its simplest form, a system of indicators is a bi-dimensional data matrix[8] typical of multivariate statistics, $\mathbf{X} \equiv \left\{ x_{ij} : i = 1...N; j = 1...J \right\}$, where the generic x_{ij} unit represents the determination of the j-th indicator in the i-th unit. Generally, given the data matrix $\mathbf{X} \equiv \{x_{ij}\}$, the objective is to synthesize it in a vector $\mathbf{v} \equiv \{v_i\}$, with N statistical units, in which the generic element v_i represents the synthetic value of the i-th unit with respect to all the indicators of the original matrix $\mathbf{X}$. In an aggregative-compensative approach, the synthesis of $\mathbf{X}$ is performed by means of a mathematical function that combines the previously standardised basic indicators. In other words, it consists in the mathematical combination (or aggregation) of the set of indicators, obtained by applying specific methodologies (Nardo et al., 2005). Over the years, these methodologies, known as CIs, have been widely used in literature for measuring and evaluating a great variety of socio-economic phenomena, such as human development (Despotis, 2005; Mariano et al., 2015; Rogge, 2018a,b), well-being and quality of life (Ülengin et al., 2001; Morais & Camanho, 2011; Ciommi et al., 2017; Cataldo et al., 2017; Dardha & Rogge, 2020), sustainable development (Krajnc & Glavič, 2005; Kondyli, 2010; Alaimo, 2018; Alaimo & Maggino, 2020; Cataldo et al., 2021), and so on. These are only a few examples of the enormous

[8] In most cases, the indicator systems are in the form of *three-way data time arrays*. These data structures are characterized by a greater complexity of information, consisting in the fact that multivariate data are observed at different *times* (D'Urso, 2000). The statistical tools presented in this paper are applied to multi-indicator systems in a specific year, i.e. a specific *slice* of a three-way time data array. Application to temporal data will be the subject of a future work.

multidisciplinary academic production on this topic. CIs are particularly useful in facilitating the reading of phenomena and the comparison and evaluation of performance of different statistical units (Archibugi & Coco, 2004; Filippetti & Peyrache, 2011; Sehnbruch et al., 2020; Masset & García-Hombrados, 2021). Consequently, they are a key tool in decision making and policy evaluation. Any policy intervention is a *choice*, which is expressed in terms of a precise allocation of resources, not only and not necessarily economic. Dealing with limited resources, an essential issue for policy makers, at different levels, is evaluating and monitoring the efficiency and effectiveness of their choices. These choices, in order to be as effective as possible, must be made in light of information about the phenomenon and the area covered by that specific policy. In this view, the CIs are a key factor. The main purpose of their success is informative. It is easier for the public to understand a synthetic indicator (one single measure) than many elementary indicators. Despite its success, the aggregative-compensative methods pose some conceptual and methodological questions that have been extensively analysed in literature (Fattore, 2017; Maggino, 2017; Alaimo & Maggino, 2020; Alaimo et al., 2022b,c,a). These are still open questions, on which the debate continues to be very intense. In this paper, we propose some new methodological tools that we think could be a possible solution to some of these issues.

The synthesis of a multi-indicator system has the objective of obtaining a synthetic measure for each unit. Switching from multi-dimensional to unidimensional implicitly involves a loss of information, justified by the need to give an easy-to-read information about a phenomenon. In many cases, this loss of information is excessive. Generally, synthesising a complex phenomenon into a single number is not straightforward and can lead to misleading results and conclusions, which increase if the indicator is poorly defined and constructed. The

consequence could be an over-simplistic interpretation of a phenomenon (OECD, 2008). As discussed in Section 2, these questions have also been raised with regard to HDI, leading some authors to support the adoption of a dashboard of indicators rather than a single measure. From those considerations, our first research question derives:

- *Should the synthesis of a multi-indicators system necessarily be a single number assigned to each statistical unit?*

We can anticipate that the answer to this question is negative. In this paper we propose an *interval of synthesis*, in which for each unit we identify an upper bound and a lower bound representing the best and the worst performance that each considered unit could obtain aggregating a system of indicators. The upper bound is calculated using the Benefit of the Doubt (BoD) approach and the lower bound is the minimum (min) between the basic indicators of the considered unit.

In all phases of the construction of a CI, *subjectivity* is involved. One of the main issues of aggregative methods is related to the way in which they are calculated, i.e. as a combination of basic indicators. Different methods of aggregation exist and can be used, leading to different results and interpretations. Of course, this choice is subjective, although it must be guided by knowledge of the phenomenon and based on clear assumptions. In particular, one of the main assumptions concerns the degree of compensation or substitutability allowed between basic indicators (Giarlotta, 2001; OECD, 2008; Munda, 2012; Mazziotta & Pareto, 2017). Generally, the basic indicators of a composite index are called substitutable if a deficit in one may be compensated by a surplus in another; on the contrary, the basic indicators are called non-substitutable. Consequently, aggregation methods can be *compensative* or *non-compensative*, depending on the

adoption or not of compensation. The compensability issue is not only methodological but also, and above all, conceptual. Looking at the indicators of HDI, if we admit, for instance, full compensability, we implicitly affirm that a surplus in education can compensate for a deficit in health. This is, at least, highly questionable. On the other hand, if we affirm the non-compensability of the basic indicators, we risk crushing the results of our synthesis. A possible solution identified in literature (Tarabusi & Guarini, 2013; Mazziotta & Pareto, 2017) is the adoption of a *partially compensative method*, i.e. allowing it "up to a certain point"; however, the question would arise as to what is the permissible and tolerable threshold of compensability. A very similar issue arises with the use of the geometric mean for the HDI. Choosing one approach over another influences the results. Moreover, this choice is often made arbitrarily by the researcher, without taking into account the conceptual assumptions that may justify compensative or non-compensative approach. But even where it is chosen with respect to assumptions, nobody ensures that this is the 'right' method. There are many methods for constructing syntheses using the composite indicators approach (Saisana & Tarantola, 2002), each of which has strengths and weaknesses. There is *no best method*. Different choices and different methods lead to different syntheses that often give a different interpretation of the phenomena studied. These considerations lead to the other research question:

- *Is it possible to identify a synthesis that is, in some way, representative of all the possible ones?*

Starting from the interval defined and using the properties of the interval data (Gioia & Lauro, 2005; Moore et al., 2009), we propose a new synthetic measure, the *mid aggregation point* (*map*). It can be considered as a "not bad solution"

among the infinite possible syntheses obtainable from a set of elementary indicators by adopting an aggregative-compensative approach. Obviously, this is not the best method in absolute terms. However, as highlighted in Section 4.2, *map* has characteristics that make it a particularly interesting choice.

4. Methodological tools

4.1. The Min-BoD interval of synthesis

Given the matrix $\mathbf{X} \equiv \{x_{ij}\}$, the first operational step in composites construction is the normalisation of data in order to obtain the matrix $\mathbf{R} \equiv \{r_{ij}\}$, in which the generic element r_{ij} is the normalised value of the generic x_{ij} of the matrix $\mathbf{X}$. The normalisation is necessary to allow the comparison between different indicators for measurement unit and variability. This step makes the indicators comparable and mathematically operational in aggregation (Talukder et al., 2017) and allows all basic indicators have positive *polarity*[9], i.e. an increase in the normalised indicators corresponds to an increase in the composite index. Normalisation is a very delicate step of composites construction, because it can change the distribution and the internal variability of indicators and it can obscure their original purpose. There are different methods, each of which presents advantages and drawbacks. In this paper, we use the re-scaling or Min-Max, one of the most common in the literature and the same adopted by UNDP for the HDI. It normalises indicators to be bounded in [0,1] by subtracting the minimum value and dividing by the range of the indicator values (OECD, 2008):

[9]The direction of the relation between the indicator and the phenomenon defines polarity; this, it depends on the type of composite. Some indicators can present *positive polarity* (i.e. they have the same direction of the phenomenon), others *negative polarity* (i.e. they are negatively related with the phenomenon).

$$r_{ij} = \frac{x_{ij} - \min\limits_{i}(x_{ij})}{\max\limits_{i}(x_{ij}) - \min\limits_{i}(x_{ij})} \tag{1}$$

where $\min\limits_{i}(x_{ij})$ and $\max\limits_{i}(x_{ij})$ are, respectively, a minimum and a maximum value (commonly the observed ones; but in the case of HDI they are the goalposts) that represent the possible range of the indicator j[10]. Obviously, this method presents pros and cons. For instance, it is particularly sensitive to the outliers. At the same time, reporting all indicators at the range [0,1], it facilitates the reading of the phenomenon. Starting from the matrix $\mathbf{R}$, we construct not a single vector $\mathbf{v}$, but an interval-valued variable (Moore et al., 2009), $\mathbf{I} \equiv \{I_I = [\underline{I_i}, \overline{I_i}]\}$, namely a variable assuming an interval of values on each statistical unit:

$$\mathbf{I} \equiv \left\{ I_i = [\underline{I_i}, \overline{I_i}] : i = 1...N \right\} \equiv \begin{pmatrix} I_1 = [\underline{I_1}, \overline{I_1}] \\ \vdots \\ I_i = [\underline{I_i}, \overline{I_i}] \\ \vdots \\ I_N = [\underline{I_N}, \overline{I_N}] \end{pmatrix} \tag{2}$$

in which $I_I = [\underline{I_i}, \overline{I_i}]$ is the interval of the generic unit i-th and $\underline{I_i}$ and $\overline{I_i}$ are, respectively, the left and right endpoints of I_i. Given a generic unit i-th, the two endpoints are constructed as follows. The left endpoint, $\underline{I_i}$, is equal to the minimum value (Min_i) between the normalised indicators of the considered unit:

$$Min_i = \min_{i}(r_{ij}) \tag{3}$$

The right endpoint is constructed by using the BoD approach, an aggrega-

[10] It applies if indicator has positive polarity, otherwise we compute the complement to respect to 1 to equaation 1.

tive method for composite indicators construction (Rogge et al., 2006; Cherchye et al., 2007) based on the Data Envelopment Analysis (DEA), a linear programming technique, useful to measure the relative efficiency of decision making units (DMU) on the basis of multiple inputs and outputs (Charnes et al., 1978; Emrouznejad & Yang, 2018). The efficiency of a set of indicators can be adapted to construct a synthetic indicator using an input-oriented DEA. Units want to maximise efficiency, defined as the ratio of outputs to inputs. When applied to the problem of calculating a composite measure, the outputs are the elementary indicators and the input for each facility is a "dummy variable" set equal to 1 (Shwartz et al., 2009). In BoD approach with DEA weighting scheme, the composite score is constrained to be ≤ 1, thus, the composite score represent the proportion of the maximum possible score that the unit has achieved. The synthetic measure is obtained as the weighted sum of the normalised indicators relatively to a benchmark; more precisely, it is defined as the performance of the single unit divided for the performance of the benchmark:

$$BoD_i = \frac{\sum_{j=1}^{J} r_{ij} w_{ij}}{r_{ij}^*} \tag{4}$$

where r_{ij} is the normalised value and w_{ij} is the corresponding weight, specific for each unit and each indicator. The benchmark r_{ij}^* is defined as follows:

$$r_{ij}^* = \max_{r_{i\in[1\dots N]}} \sum_{j=1}^{J} r_{ij} w_{ij} \tag{5}$$

The identification of the optimal set of weights guarantees that each unit is associated to the best possible position compared to all the others and it is obtained as follows:

$$BoD_i^* = \max_{w_{ij}} \frac{\sum_{j=1}^{J} r_{ij} w_{ij}}{\max\limits_{k \in [1...N]} \sum_{j=1}^{J} r_{kj} w_{kj}}, \forall i = 1...N \tag{6}$$

under the constraints that the weights are non-negative and the result is bounded [0,1]. "In the absence of an *a priori* weighting scheme, the method thus selects the weights which maximise the composite indicator for each country under investigation" (Cherchye et al., 2007, 120). The most favourable weights are always applied to all observations. The composite score depends exclusively on the frontier's distance and not on the relationship between basic indicators. Obviously, this method presents drawbacks, the main of which are related to the DEA solution[11]. However, it has been and continues to be widely used in different fields and many methodological innovations have been proposed (Rogge, 2018a,b; Verbunt & Rogge, 2018; Fusco et al., 2018; Färe et al., 2019).

Thus, we obtain the $Min - BoD$ interval of synthesis

$$\mathbf{I} \equiv \left\{ I_i = [\underline{Min_i}, \overline{BoD_i}] : i = 1...N \right\}.$$

For each unit, this interval is narrower or, in extreme cases, equal to that of the Min-Max, [0,1]. We chose BoD in order to have an upper bound higher or at least equal to the maximum between the basic indicators (Cherchye et al., 2007; De Witte & Rogge, 2009; Vidoli & Mazziotta, 2013), but hypothetically reachable if units allocate their resources optimally. As lower extreme, we used the minimum among the indicators for each unit because, in the time considered,

[11] For example, since the weights are specific for each unit, cross-unit comparisons are not possible and the values of the scoreboard depend on the benchmark performance. Moreover, another drawback is the multiplicity of equilibria. Hiding a problem of multiple equilibria makes the weights not uniquely determined (even if the composite indicator is unique). The optimisation process could lead to many 0-weights if no restrictions are imposed on the weights. For a detailed analysis, please see Vidoli & Mazziotta (2013).

it is not conceivable that a unit can fall or desire to fall below this minimum level. The proposed $Min - BoD$ interval of synthesis presents some interesting properties and advantages.

The interval defines a range of variation of the performances for each unit, obtained starting from the values of the basic indicators and in which each point is hypothetically reachable. Within that range, the values of all the mean-based aggregation methods are included. Consequently, this makes them comparable with respect to the minimum achieved and the maximum hypothetically reachable by the unit. Given a generic unit i-th, $\underline{Min}_i$ represents the worst value obtainable from a composite, given that specific set of basic indicators. $\overline{BoD}_i$ is a point on the frontier of the unit i-th, representing the maximum value it could aspire to achieve given the starting set of basic indicators. Obviously, each unit must aspire to improve its performance and, consequently, the $\overline{BoD}_i$ represents an actual benchmark to which the unit i must strive to arrive. Hence an interesting use of the $Min - BoD$ interval of synthesis. Given a synthetic vector $\mathbf{v}$, obtained by any aggregative-compensative method (using the Min-Max normalisation), for each generic unit i-th we can calculate the quantity:

$$ {}_a dist_i = \overline{BoD}_i - v_i \tag{7}$$

which expresses the distance of the synthetic measure (obtained by means of a specific method) from the optimal performance to which the unit can aspire (on the basis of the specific set of basic indicators). Obviously, the benchmarks are different for each unit. The resulting vector $\mathbf{{}_a dist}$ is constituted by the distances of each unit from a benchmark that we can consider "true", i.e. effectively achievable. The higher the value of ${}_a dist_i$, the greater the distance of the syn-

thetic value of the generic unit i-th from the "true" benchmark. Similarly, we can calculate the distance from the minimum:

$$_{b}dist_i = v_i - \underline{Min}_i \tag{8}$$

where $_{b}dist$ expresses how much the unit's synthetic value deviates from its worst performance. In this way, we can evaluate the performance of each unit (the value of the synthesis calculated by a specific aggregative method) not only relatively to the others, but also to itself, by comparing the value of a chosen aggregation method with the endpoints of the $Min - BoD$ interval of that unit.

4.2. The mid aggregation point - map

Starting from the $Min - BoD$ interval of synthesis, we can use the midpoint (Gioia & Lauro, 2005) between the two endpoints for each unit as a synthetic measure. Given the interval $\mathbf{I}$, we obtain the vector of the *mid aggregation points*, $\mathbf{map} \equiv \{map_i : i = 1 \dots N\}$, where the generic element map_i is given by:

$$map_i = \frac{\underline{Min}_i + \overline{BoD}_i}{2} \tag{9}$$

map_i presents two interesting properties, useful to propose a possible solution to two problems related to the aggregative-compensative approach. It is equally-spaced from the two endpoints of the interval $\mathbf{I_i}$; consequently, it is at the center of a symmetrical interval. For this reason, it gives an easy-to-read representation of the distance from the best and worst possible performance; indeed, the relative distance (i.e., the distance of map_i from the endpoints of $\mathbf{I_i}$) is the same for each unit. The values of different synthetic measures depend on each subjective decision made; different choices lead to different results. It would be helpful to

have a criterion for choosing one method over others. In most cases, this criterion does not exist and the choice is made arbitrarily by the researcher. *map* gives a possible criterion of choice. It can be chosen if one wants to represent a "not bad solution" among the infinite possible syntheses obtainable from a set of basic indicators by using an aggregative-compensative method. Another interesting property of *map* is that it is only partially affected by compensability. It is the midpoint between the extreme calculated with the BoD method, affected by high compensability (Vidoli & Mazziotta, 2013), and the Min that is totally non-compensative. Choosing it, then, allows the consideration of a partially non-compensatory method, a good compromise between a totally compensatory and a totally non-compensatory one.

5. An application to Human Development Index (HDI) data

Data used are freely downloadable from the UNDP website[12] and refer to indicators collected for 189 countries in 2019. To ensure that the measures obtained are comparable with the index produced by UNDP, we followed the same data processing procedures[13]. We use a Min-Max normalisation; the minimum and the maximum values (the goalposts in Table 1) are chosen according to the theoretical framework and act as the "natural zeros" and "aspirational targets," respectively, from which component indicators are normalised[14]. After the normalisation, for the education dimension, the arithmetic mean of the two normalised indices is

[12] http://hdr.undp.org/en/content/download-data.

[13] For detailed information, please see: `http://hdr.undp.org/sites/default/files/hdr2019_technical_notes.pdf`.

[14] When a variable exceeds the upper bound of its dimension, the value is truncated at the upper bound so that the range of normalised indicators is always between 0 and 1 and none of the dimensional sub-indices exceeded 1.

taken. For GNI, the natural logarithm is used. The HDI is the geometric mean of the three dimensional indices: $HDI = \sqrt[3]{I_{health} * I_{education} * I_{income}}$

Table 1: HDI basic indicators: goalposts.

Dimension	Indicator	Minimum	Maximum
Health	Life expectancy at birth	20	85
Education	Expected years of schooling at birth	0	18
	Mean years of schooling at birth	0	15
Standard of living	Gross national income per capita	100	75,000

The system of indicators used in this work includes the dimensional indices of HDI as constructed according to the UNDP procedure[15]. Table 2 reports the summary statistics (Table 5 in Appendix reports the summary statistics of the basic indicators used for constructing dimensional indicators). All the considered indicators present low coefficients of variation, negative skewed and platykurtic distributions. Figure 1 shows that all the indicators are highly correlated and all the Pearson's linear correlation coefficients are statistically significance with confidence level $\alpha = 0.001$.

Table 2: Dimensional indices of HDI: summary statistics.

	Min	Max	1^{st} Quartile	Median	3^{rd} Quartile	Mean	cv	Skewness β_1	Kurtosis β_2
I_{health}	0.512	0.998	0.733	0.832	0.891	0.812	0.140	-0.565	2.627
$I_{education}$	0.249	0.943	0.531	0.682	0.791	0.660	0.260	-0.362	2.269
I_{income}	0.305	1.000	0.589	0.732	0.859	0.715	0.242	-0.245	2.141

The 189 countries are very different from one another. In order to facilitate the reading of the results, we decided to classify them into homogeneous groups according to the basic indicators. We used a *model based clustering* approach based

[15]In this application, we consider only 3 dimensions and 4 indicators. It should be made clear that the methods presented are applicable for any indicator system regardless of the number of dimensions and indicators.

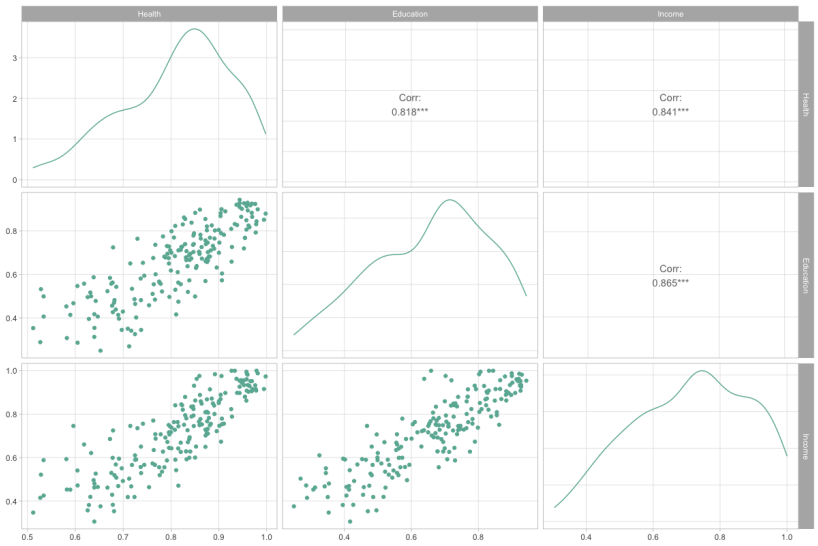
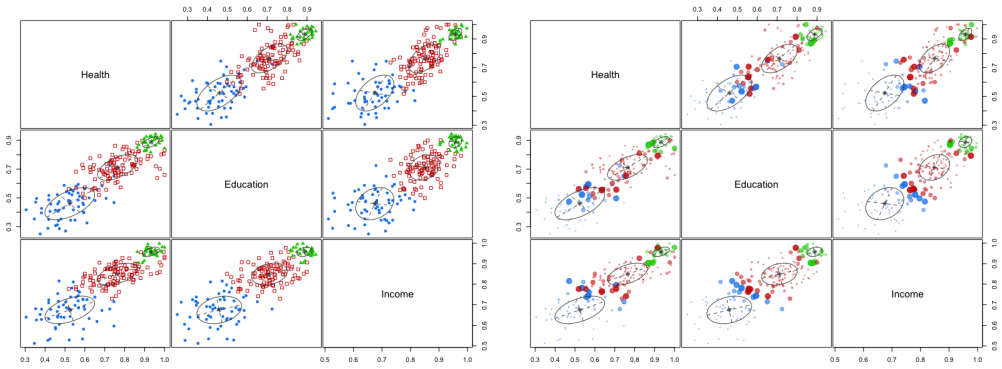


Figure 1: Scatterplots, density plots and Pearson's linear correlation coefficients of the three dimensional indices of HDI.

on parameterised finite Gaussian mixture models (McLachlan & Basford, 1988; Fraley & Raftery, 2002), computed via the package *Mclust* (Scrucca et al., 2016) of the **R** statistical software. The model-based clustering offers the advantage of clearly stating the assumptions behind the clustering algorithm. Moreover, it allows cluster analysis to benefit from the inferential framework of statistics to address some of the practical questions arising when performing classification: determining the number of clusters, detecting and treating outliers, assessing uncertainty in the clustering (Bouveyron et al., 2019). Finally, we focus on mixture of multivariate Gaussian densities for its ability to approximate the density function of any unknown distribution (Titterington et al., 1985). We compare 14 models with different geometric characteristics of the covariances[16]; each model was applied for different number of components, $2 \leq G \leq 8$. The number of

[16]As specified in Scrucca et al. (2016), in the multivariate setting, the volume, shape, and orientation of the covariances can be constrained to be equal (E) or variable (V) across groups. Thus, 14 possible models with different geometric characteristics can be specified.

mixing components (G) and the covariance parameterisation are selected using the Bayesian Information Criterion (BIC). We select the model VEE (for details, please see Scrucca et al. (2016)), i.e. variable - V volume, equal - E shape, equal - E orientation and ellipsoidal distribution, with 3 components (Table 6 in Appendix - B reports the values of the BIC for the different models estimated). Figure 2a show the clusters produced for each couple of indicators and Figure 2b show the classification uncertainty, which indicates that the most of observations are well classified. The ellipses superimposed on the classification and uncertainty plots correspond to the covariances of the components.



(a) Clustering classification plot (b) Clustering uncertainty plot

Figure 2: Classification of countries; dimensional indices of HDI; world countries. Year 2019.

In Table 3, we report the indicators' means of each component. Cluster 1 presents values quite low in the three dimensions of HDI. It includes 59 countries (for instance, Gambia, Haiti, India, Kenya), predominantly African and Asian. We indicate this cluster as *Cluster 1 - Low HDI*. Cluster 3 comprises 29 countries (for instance, Great Britain, Canada, Australia, Japan) and has high values in all the indicators considered. We label this cluster as *Cluster 3 - High HDI*. The 101 remaining countries (for instance, USA, Italy, Brasil, China) are classified in Cluster 2, characterised by intermediate values in the indicators between those

of Cluster 1 and Cluster 3. We indicate this cluster as *Cluster 2 - Medium HDI*. Figure 3 shows the subdivision of the world countries according to the clustering partition.

Table 3: GMM partition: means of each component; number of units.

	Health	Education	Income	Total units
Cluster 1 - Low HDI	0.676	0.460	0.524	58
Cluster 2 - Medium HDI	0.850	0.711	0.764	102
Cluster 3 - High HDI	0.957	0.887	0.933	29

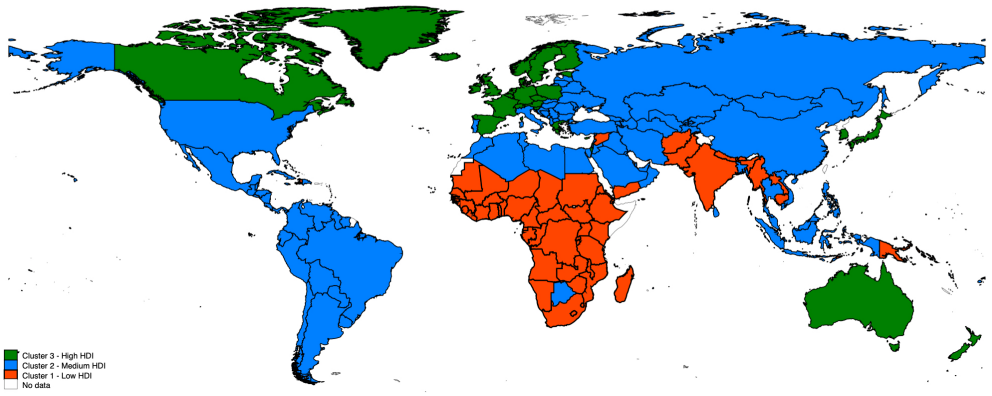


Figure 3: HDI dimensional indices: clusters' composition of world countries. Year 2019

At this point, we calculate the $Min - BoD$ interval of synthesis and the mid aggregation point *map*. At the same time, we compute other syntheses by using methods widely used in literature: minimum, maximum, arithmetic mean, geometric mean, quadratic mean, cubic mean (results in Figure 5 in Appendix).

Figures 5–10 in Appendix report all the synthetic measures constructed for Cluster 1, 2 and 3. The main characteristics discussed in the previous pages appear evident. The $Min - BoD$ interval is significantly narrower than the Min-Max and it includes all the synthetic measures obtained by means of different methods. BoD is higher or at least equal to the maximum, as clearly shown in

Figures (for instance, see Canada - CAN, Sweden - SWE, United States - USA, Italy - ITA) and it can be conceived as an actual benchmark to which each unit must strive to arrive. Within the range of variation of the performances defined by the $Min-BoD$ interval, the values of all the mean-based aggregation methods are included. Thus, we can compare them with respect to the minimum achieved and the maximum hypothetically reachable by each unit. This interval allows an assessment of countries' human development levels that takes into account their potential improvements according to their achievements in the basic indicators. Let us take an example. Suppose we want to assess the level of human development achieved by 4 countries, Italy (ITA), France (FRA), Spain (ESP) and Germany (DEU)[17], compared to the maximum level they can achieve based on their combinations in the basic indicators (i.e., with respect to the BoD). We consider 3 synthetic measures: HDI, the mid aggregation point (map) and the arithmetic mean of the normalised indicators (AM).

Comparing the units in Figure 4, we can observe that DEU always performs better than all the other countries, regardless of the synthesis considered; similarly, ITA has the worst values. In addition to evaluating each country with the others, we can also compare it with itself using the interval. DEU presents a very narrow $Min - BoD$ interval of synthesis ($[0.943, 1.00]$) and this indicates that it can assume very similar values regardless of the synthesis used; on the contrary, ITA presents a much wider range ($[0.793, 0.979]$). If we look at the upper bounds of the intervals (BoD), ITA has an optimal aspirational performance higher than the one of FRA and equal to that of ESP, despite the fact that its synthetic indices are lower than those of the other two nations. We can interpret

[17]We have chosen these four nations because of their similarity and because they are often used as comparisons with each other.

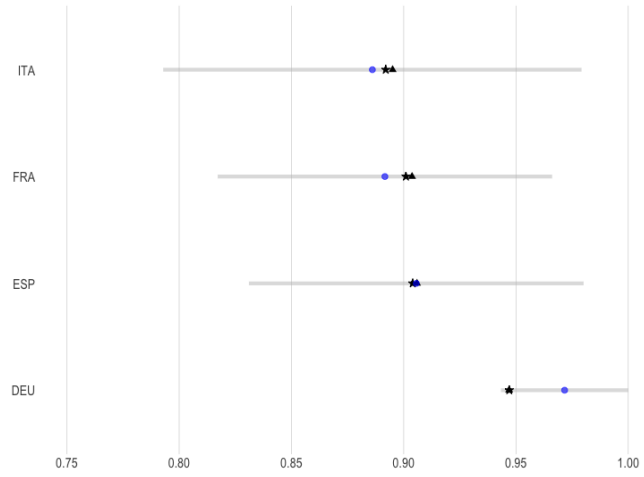


Figure 4: Example of comparison of synthetic measure. Countries: Italy (ITA), France (FRA), Spain (ESP) and Germany (DEU); measures: — $BoD - Min$ interval of synthesis; • mid aggregation point - *map*; ★ HDI of UNDP; ▲ arithmetic mean. Year 2019.

this situation in terms of Italy's greater potential compared to France, given the starting situation (the combination in the basic indicators). At this point, we can highlight the properties of the *map*. By definition, it is the midpoint of the interval and, therefore, equidistant from the two bounds. This is certainly an interesting property and may be a potential criterion for choosing the *map* over all other synthetic measures. Table 4 reports, for each considered country, the distances ($_i dist_a$ as defined in Section 4.1) of the different synthetic measures from the optimal performances and the corresponding percentages.

Table 4: Example of comparison of synthetic measure. Countries: Italy (ITA), France (FRA), Spain (ESP) and Germany (DEU); measures: index of UNDP (HDI), Arithmetic mean (AM), mid aggregation point (*map*), minimum (Min), benefit of the doubt (BoD); distance and percentages of distance from the AM ($_{AM}dist_a$ - $Pct_{dist_{AM}}$), from HDI ($_{HDI}dist_a$ - $Pct_{dist_{HDI}}$) and from *map* ($_{map}dist_a$ - $Pct_{dist_{map}}$). Year 2019.

	HDI	AM	*map*	Min	BoD	Range	$_{AM}dist_a$	$_{HDI}dist_a$	$_{map}dist_a$	$Pct_{dist_{AM}}$	$Pct_{dist_{HDI}}$	$Pct_{dist_{map}}$
ITA	0.892	0.895	0.886	0.793	0.979	0.186	0.084	0.087	0.093	45.16%	46.80%	50.00%
FRA	0.901	0.904	0.892	0.817	0.966	0.149	0.062	0.065	0.074	41.88%	43.70%	50.00%
DEU	0.947	0.947	0.972	0.943	1.000	0.057	0.053	0.053	0.028	93.57%	93.49%	50.00%
ESP	0.904	0.906	0.906	0.831	0.980	0.149	0.074	0.076	0.075	49.78%	51.07%	50.00%

The results show that there are profound differences between units. Looking at the AM, in percentage, FRA is 41.9% distant from its maximum, ITA 45.1%, ESP 49.8% and DEU, even, 93.5%. A similar situation applies to the HDI. On the contrary, *map* is, for all units, proportionally equidistant from the maximum, 50%. This does not mean that the other measures are wrong, but only that choosing the *map* we have a synthesis that for all the considered units is exactly between the worst and the best performance hypothetically achievable. This, undoubtedly simplify the interpretation of the results and can be a factor in choosing the *map*, especially if there is a lack of useful information about the phenomenon.

6. Concluding remarks

Over the years, the Human Development index has risen to prominence as one of the main indicators for measuring wellbeing and quality of life in countries. Because of its success, changes and improvements have been proposed. At the same time, conceptual and methodological problems and weaknesses were highlighted, which continue to be debated today. Many of these open issues are more generally related to the composite indicators. The latter are an indispensable tool for measuring and understanding complex socio-economic phenomena. They have become the focus of attention of researchers, as tools for reading reality as well as policy makers, for their ease of reading and usefulness for decision making and policy evaluation. The increasing use of composite indicators has been accompanied by the conceptual and methodological debate on the problems associated with their construction. In particular, the oversimplification and excessive loss of information and the subjectivity of choices that characterises the entire construction process are drawbacks that has been and continues to

be much debated (Freudenber, 2003; Nardo et al., 2005; Maggino, 2017; Fattore, 2017; Alaimo, 2020). These problems are also typical for HDI. Accordingly, in this paper we have tried to address two research questions, valid both for the HDI and for the general topic of composite indicators.

We started with the question whether the synthesis should necessarily be a number. As highlighted in a consistent literature, the synthesis can be an object, a map, an image (Tufte, 1983; Lima, 2011). It must be an informative patrimony capable of describing the observed reality (Alaimo et al., 2022c). In line with this, we have proposed the $Min - BoD$ interval of synthesis. This interval defines a range of variation of the performances for each unit, obtained starting from the values of the basic indicators and in which each point is hypothetically reachable. We choose as lower bound the minimum among the indicators for each unit because, in the time considered, it is not conceivable that a unit can fall or desire to fall below this minimum level. As upper bound, we use the BoD, in order to have an upper bound higher or at least equal to the maximum between the basic indicators (Cherchye et al., 2007; Vidoli & Mazziotta, 2013), but hypothetically reachable. BoD has often been criticized in literature because it is not a "truthful" synthesis, but an expression of an ideal performance to which a unit can aspire. Moreover, another criticism is related to the fact that the weights are different from one unit to another. By using BoD as the limit of the proposed interval of synthesis, however, these problems do not arise. It acts not as a synthesis, but as a "benchmark", the maximum level to which the unit can aim to achieve, given its starting situation. Moreover, the difference in weights between one unit and another in this perspective is an advantage: it is absolutely acceptable that the units have different possibilities depending on their observed situation. The interval of synthesis includes all the mean-based

aggregation methods. In this way, the latter can be compared by taking as a reference the endpoints of the interval, representing two benchmarks hypothetically achievable. The distance of a given CI from the benchmark (minimum or maximum) is an important element, which enriches the information provided by the synthesis, giving a more complete picture of the phenomenon and allowing a more precise evaluation of the statistical units. This can be seen from the application to HDI in Section 5.

Since different choices produce different syntheses and different results, the second research question was if it is possible to identify a synthesis that is "representative" of all possible ones. We proposed the *map* as a possible choice. It is at the center of a symmetrical interval and, thus, it gives an easy-to-read representation of the distance from the best and worst possible performance. This point can be considered as a "not bad solution" among the infinite possible syntheses obtainable from a set of basic indicators by using an aggregative-compensative method.

Of course, it should be remembered that the methods proposed in this paper belong to the aggregative-compensative approach, and, therefore, present some of its limitations and weaknesses, such as its in-applicability when non-cardinal indicators are present (Fattore, 2017; Alaimo et al., 2022c). In addition, the proposed methods are currently not applicable in the case of multi-indicators systems over time.

The application of the proposed methods on the Human Development Index, addresses and tries to give a solution to some of the most common criticisms in the literature. Obviously, such methods can also be used for the construction of other composite indicators.

As a future development of this work, we will propose the application of the

interval of synthesis to longitudinal data. This means to address the problem of optimising the weights in time, for the upper limit of the interval (BoD).

Declarations

- The authors have no relevant financial or non-financial interests to disclose.
- The authors have no competing interests to declare that are relevant to the content of this article.
- All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.
- The authors have no financial or proprietary interests in any material discussed in this article.

Data availability statement

The data used for the analysis conducted in this paper are available and freely accessible at the following link: `https://hdr.undp.org/en/data`.

References

Alaimo, L. S. (2018). Sustainable development and national differences: an european cross-national analysis of economic sustainability. *RIEDS-Rivista Italiana di Economia, Demografia e Statistica-Italian Review of Economics, Demography and Statistics*, *72*, 101–123.

Alaimo, L. S. (2020). Complexity of Social Phenomena: Measurements, Analysis, Representations and Synthesis. *Unpublished Doctoral Dissertation, University of Rome" La Sapienza", Rome, Italy*, .

Alaimo, L. S. (2021a). Complex systems and complex adaptive systems. In F. Maggino (Ed.), *Encyclopedia of Quality of Life and Well-being Research* (pp. 1–3). Cham: Springer. doi:10.1007/978-3-319-69909-7_104659-1.

Alaimo, L. S. (2021b). Complexity and knowledge. In F. Maggino (Ed.), *Encyclopedia of Quality of Life and Well-being Research* (pp. 1–2). Cham: Springer. doi:10.1007/978-3-319-69909-7_104658-1.

Alaimo, L. S. (2022). Open issues in composite indicators construction. In A. Balzanetta, M. Bini, C. Cavicchia, & R. Verde (Eds.), *Book of Short Papers SIS 2022* (pp. 176–185). Milano: Pearson.

Alaimo, L. S., Arcagni, A., Fattore, M., & Maggino, F. (2021). Synthesis of multi-indicator system over time: A poset-based approach. *Social Indicators Research*, *157*, 77–99. Https://doi.org/10.1007/s11205-020-02398-5.

Alaimo, L. S., Arcagni, A., Fattore, M., Maggino, F., & Quondamstefano, V. (2022a). Measuring Equitable and Sustainable Well-being in Italian Regions. The Non-aggregative Approach. *Social Indicators Research*, *161*, 711–733. Doi: https://doi.org/10.1007/s11205-020-02388-7.

Alaimo, L. S., Fiore, M., & Galati, A. (2022b). Measuring consumers' level of satisfaction for online food shopping during covid-19 in italy using posets. *Socio-Economic Planning Sciences*, *82*, 101064. DOI: 10.1016/j.seps.2021.101064.

Alaimo, L. S., Ivaldi, E., Landi, S., & Maggino, F. (2022c). Measuring and evaluating socio-economic inequality in small areas: An application to the urban units of the municipality of genoa. *Socio-Economic Planning Sciences*, *83*, 101170.

Alaimo, L. S., & Maggino, F. (2020). Sustainable development goals indicators at territorial level: Conceptual and methodological issues—The Italian perspective. *Social Indicators Research*, *147*, 383–419.

Alkire, S. (2002). Dimensions of human development. *World development*, *30*, 181–205.

Alkire, S., & Foster, J. E. (2010). Designing the inequality-adjusted human development index. *Oxford Poverty & Human Development Initiative (OPHI) Working Paper*, *10*.

Anand, S., & Sen, A. (1997). Concepts or human development and poverty. a multidimensional perspective. *United Nations Development Programme, Poverty and human development: Human development papers*, (pp. 1–20).

Anand, S., & Sen, A. (2000a). Human development and economic sustainability. *World Development*, *28*, 2029–2049.

Anand, S., & Sen, A. (2000b). The income component of the human development index. *Journal of Human Development*, *1*, 83–106.

Archibugi, D., & Coco, A. (2004). A new indicator of technological capabilities for developed and developing countries (arco). *World development*, *32*, 629–654.

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*, *2*,

244–263.

Biggeri, M., & Mauro, V. (2018). Towards a more 'sustainable'human development index: Integrating the environment and freedom. *Ecological indicators*, *91*, 220–231.

Biswas, B., & Caliendo, F. (2002). A multivariate analysis of the human development index. *Economics Research Institute Study Paper*, *11*, 1.

Blalock, H. M. (1964). *Causal Inferences in Nonexperimental Research*. N.C.: University of North Carolina Press.

Blalock, H. M. (1968). The Measurement Problem: A Gap between the Languages of Theory and Research. In F. Kerlinger (Ed.), *Methodology in Social Research* (pp. 5–27). New York: McGraw-Hill.

Bleys, B. (2012). Beyond GDP: Classifying Alternative Measures for Progress. *Social Indicators Research*, *109*, 355–376.

Boarini, R., Johansson, Å., & d'Ercole, M. M. (2006). *Alternative Measures of Well-being*. 33. OECD. URL: `https://www.oecd-ilibrary.org/content/paper/713222332167`. doi:`https://doi.org/10.1787/713222332167`.

Bollen, K. A. (1984). Multiple Indicators: Internal consistency or No Necessary relationship? *Quality and Quantity*, *18*, 377–385.

Bollen, K. A. (2007). Interpretational Confounding is due to Misspecification, not to Type of Indicator: Comment on Howell, Breivik, and Wilcox (2007). *Psychological Methods*, *12*, 219–228.

Bollen, K. A., & Diamantopoulos, A. (2017). In Defense of Causal-formative Indicators: A Minority Report. *Psychological Methods*, *22*, 581–596.

Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge University Press.

Brüggemann, R., & Patil, G. P. (2011). *Ranking and prioritization for multi-indicator systems: Introduction to partial order applications*. Dordrecht: Springer Science & Business Media.

Cataldo, R., Crocetta, C., Grassia, M. G., Lauro, N. C., Marino, M., & Voytsekhovska, V. (2021). Methodological pls-pm framework for sdgs system. *Social Indicators Research*, *156*, 701–723.

Cataldo, R., Grassia, M. G., Lauro, N. C., & Marino, M. (2017). Developments in higher-order pls-pm for the building of a system of composite indicators. *Quality & Quantity*, *51*, 657–674.

Charnes, A., Cooper, W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*, 429–444.

Cherchye, L., Moesen, W., Rogge, N., & Van Puyenbroeck, T. (2007). An Introduction to 'Benefit of the Doubt'Composite Indicators. *Social Indicators Research*, *82*, 111–145.

Chhibber, A., & Laajaj, R. (2007). A multi-dimensional development index: Extending the human development index with environmental sustainability and security. *UNDP[online] Available at: http://www. aae. wisc. edu/events/papers/DevEcon/2008/laajaj*, *10*.

Ciommi, M., Gigliarano, C., Emili, A., Taralli, S., & Chelli, F. M. (2017). A new class of composite indicators for measuring well-being at the local level: An application to the equitable and sustainable well-being (bes) of the italian provinces. *Ecological indicators*, *76*, 281–296.

Dardha, E., & Rogge, N. (2020). How's life in your region? measuring regional material living conditions, quality of life and subjective well-being in oecd countries using a robust, conditional benefit-of-the-doubt model. *Social Indicators Research*, *151*, 1015–1073.

De Witte, K., & Rogge, N. (2009). Accounting for exogenous influences in a benevolent performance evaluation of teachers. *Available at SSRN 1462690*, .

Desai, M. (1991). Human development: concepts and measurement. *European Economic Review*, *35*, 350–357.

Despotis, D. (2005). Measuring human development via data envelopment analysis: the case of asia and the pacific. *Omega*, *33*, 385–390.

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing Formative Measurement Models. *Journal of Business Research*, *61*, 1203–1218.

Diamantopoulos, A., & Siguaw, J. A. (2006). Formative versus Reflective Indicators in Organizational Measure Development: A Comparison and Empirical Illustration. *British Journal of Management*, *17*, 263–282.

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research*, *38*, 269–277.

Diener, E., & Suh, E. (1997). Measuring quality of life: Economic, social, and subjective indicators. *Social indicators research*, *40*, 189–216.

Doessel, D. P., & Gounder, R. (1994). Theory and measurement of living levels: Some empirical results for the human development index. *Journal of International Development*, *6*, 415–435.

D'Urso, P. (2000). Classificazione fuzzy per matrici a tre vie temporali. *Unpublished doctoral*

dissertation, University of Rome "La Sapienza", Rome, Italy, .

D'Urso, P., Alaimo, L. S., De Giovanni, L., & Massari, R. (2022). Well-Being in the Italian Regions Over Time. *Social Indicators Research*, *161*, 599–627.

Edwards, J. R., & Bagozzi, R. P. (2000). On the Nature and Direction of Relationships between Constructs and Measures. *Psychological Methods*, *5*, 155–174.

Emrouznejad, A., & Yang, G.-l. (2018). A survey and analysis of the first 40 years of scholarly literature in dea: 1978–2016. *Socio-economic planning sciences*, *61*, 4–8.

Färe, R., Karagiannis, G., Hasannasab, M., & Margaritis, D. (2019). A benefit-of-the-doubt model with reverse indicators. *European Journal of Operational Research*, *278*, 394–400.

Fattore, M. (2017). Synthesis of Indicators: The Non-aggregative Approach. In F. Maggino (Ed.), *Complexity in Society: From Indicators Construction to their Synthesis* (pp. 193–212). Cham: Springer.

Filippetti, A., & Peyrache, A. (2011). The patterns of technological capabilities of countries: a dual approach using composite indicators and data envelopment analysis. *World Development*, *39*, 1108–1121.

Foster, J. E., Lopez-Calva, L. F., & Szekely, M. (2005). Measuring the distribution of human development: methodology and an application to mexico. *Journal of Human Development*, *6*, 5–25.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, *97*, 611–631.

Freudenber, M. (2003). *Composite Indicators of Country Performance*. Paris: OECD Publishing. URL: `https://www.oecd-ilibrary.org/content/paper/405566708255`. doi:`https://doi.org/https://doi.org/10.1787/405566708255`.

Fusco, E., Vidoli, F., & Sahoo, B. K. (2018). Spatial heterogeneity in composite indicator: A methodological proposal. *Omega*, *77*, 1–14.

Gasper, D. (2002). Is sen's capability approach an adequate basis for considering human development? *Review of political economy*, *14*, 435–461.

Giarlotta, A. (2001). Multicriteria compensability analysis. *European Journal of Operational Research*, *133*, 190–209.

Gioia, F., & Lauro, C. N. (2005). Basic statistical methods for interval data. *Statistica applicata*, *17*, 75–104.

Harttgen, K., & Klasen, S. (2012). A household-based human development index. *World Development*, *40*, 878–899.

Herrero, C., Martínez, R., & Villar, A. (2012). A newer human development index. *Journal of Human Development and Capabilities*, *13*, 247–268.

Hicks, D. A. (1997). The inequality-adjusted human development index: a constructive proposal. *World Development*, *25*, 1283–1298.

Kelley, A. C. (1991). The human development index:" handle with care". *Population and Development Review*, (pp. 315–324).

Klugman, J. (2009). Human development report 2009. overcoming barriers: Human mobility and development. *Overcoming Barriers: Human Mobility and Development (October 5, 2009). UNDP-HDRO Human Development Reports*, .

Kondyli, J. (2010). Measurement and evaluation of sustainable development: A composite indicator for the islands of the north aegean region, greece. *Environmental Impact Assessment Review*, *30*, 347–356.

Kovacevic, M. (2010). Review of hdi critiques and potential improvements. *Human development research paper*, *33*, 1–44.

Krajnc, D., & Glavič, P. (2005). A model for integrated assessment of sustainable development. *Resources, Conservation and Recycling*, *43*, 189–208.

Lima, M. (2011). *Visual Complexity: Mapping Patterns of Complexity*. New York: Princeton Architectural Press.

Maggino, F. (2017). Dealing with Syntheses in a System of Indicators. In F. Maggino (Ed.), *Complexity in Society: From Indicators Construction to their Synthesis* (pp. 115–137). Cham: Springer.

Maggino, F., Bruggemann, R., & Alaimo, L. S. (2021). Indicators in the framework of partial order. In R. Bruggemann, L. Carlsen, T. Beycan, C. Suter, & F. Maggino (Eds.), *Measuring and Understanding Complex Phenomena: Indicators and their Analysis in Different Scientific Fields* (pp. 17–29). Cham: Springer International Publishing.

Mariano, E. B., Sobreiro, V. A., & do Nascimento Rebelatto, D. A. (2015). Human development and data envelopment analysis: A structured literature review. *Omega*, *54*, 33–49.

Masset, E., & García-Hombrados, J. (2021). Sensitivity matters. comparing the use of multiple indicators and of a multidimensional poverty index in the evaluation of a poverty eradication

program. *World Development*, *137*, 105162.

Mazumdar, K. (2003). A new approach to human development index. *Review of Social Economy*, *61*, 535–549.

Mazziotta, M., & Pareto, A. (2017). Synthesis of Indicators: The Composite Indicators Approach. In F. Maggino (Ed.), *Complexity in Society: From Indicators Construction to their Synthesis* (pp. 159–191). Cham: Springer.

McGillivray, M. (1991). The human development index: Yet another redundant composite development indicator? *World Development*, *19*, 1461–1468.

McGillivray, M., & White, H. (1993). Measuring development? the undp's human development index. *Journal of international development*, *5*, 183–192.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering* volume 38. M. Dekker New York.

Moore, R. E., Kearfott, R. B., & Cloud, M. J. (2009). *Introduction to interval analysis*. SIAM.

Morais, P., & Camanho, A. S. (2011). Evaluation of performance of european cities with the aim to promote quality of life improvements. *Omega*, *39*, 398–409.

Munda, G. (2012). Choosing aggregation rules for composite indicators. *Social Indicators Research*, *109*, 337–354.

Nardo, M., Saisana, M., Saltelli, A., & Tarantola, S. (2005). Tools for composite indicators building. *European Commission, Ispra*, *15*, 19–20.

Noorbakhsh, F. (1998). The human development index: some technical issues and alternative indices. *Journal of International Development: The Journal of the Development Studies Association*, *10*, 589–605.

Nussbaum, M. C. (2000). *Women and human development: The capabilities approach* volume 3. Cambridge University Press.

OECD (2008). Handbook on Constructing Composite Indicators. Methodology and User Guide.

Palazzi, P., & Lauri, A. (1998). The human development index: Suggested corrections. *PSL Quarterly Review*, *51*, 193–221.

Paul, S. (1996). A modified human development index and international comparison. *Applied Economics Letters*, *3*, 677–682.

Ranis, G., Stewart, F., & Samman, E. (2006). Human development: beyond the human development index. *Journal of Human Development*, *7*, 323–358.

Rogge, N. (2018a). Composite indicators as generalized benefit-of-the-doubt weighted averages. *European Journal of Operational Research*, *267*, 381–392.

Rogge, N. (2018b). On aggregating benefit of the doubt composite indicators. *European Journal of Operational Research*, *264*, 364–369.

Rogge, N., Cherchye, L., Moesen, W., & Van Puyenbroeck, T. (2006). 'Benefit of the Doubt' Composite Indicators. In *European Conference on Quality in Survey Statistics*.

Sagar, A. D., & Najam, A. (1998). The human development index: a critical review. *Ecological economics*, *25*, 249–264.

Saisana, M., & Tarantola, S. (2002). *State-of-the-art report on current methodologies and practices for composite indicator development*. EUR 20408 EN, European Commission-JRC: Italy.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, *8*, 289.

Sehnbruch, K., González, P., Apablaza, M., Méndez, R., & Arriagada, V. (2020). The quality of employment (qoe) in nine latin american countries: A multidimensional perspective. *World Development*, *127*, 104738.

Sen, A. (1999). *Commodities and Capabilities*. Oxford: Oxford University Press.

Shwartz, M., Burgess, J. F., & Berlowitz, D. (2009). Benefit-of-the-doubt approaches for calculating a composite measure of quality. *Health Services and Outcomes Research Methodology*, *9*, 234–251.

Slottje, D. J. (1991). Measuring the quality of life across countries. *The Review of Economics and Statistics*, (pp. 684–693).

Stanton, E. A. (2007). The human development index: A history. *PERI Working Papers*, (p. 85).

Talukder, B., W Hipel, K., W vanLoon, G. et al. (2017). Developing composite indicators for agricultural sustainability assessment: Effect of normalization and aggregation techniques. *Resources*, *6*, 66.

Tarabusi, E. C., & Guarini, G. (2013). An unbalance adjustment method for development indicators. *Social indicators research*, *112*, 19–45.

Titterington, D. M., Afm, S., Smith, A. F., Makov, U. et al. (1985). *Statistical analysis of finite mixture distributions* volume 198. John Wiley & Sons Incorporated.

Tufte, E. R. (1983). *The Visual Display of quantitative information*. Cheshire: Graphics Press.

Ul Haq, M. (1995). *Reflections on human development*. Oxford University Press.

Ülengin, B., Ülengin, F., & Güvenç, Ü. (2001). A multidimensional approach to urban quality of life: The case of istanbul. *European Journal of Operational Research*, *130*, 361–374.

UNDP (1990). *Human Development Report 1990*. New York: Oxford University Press.

UNDP (1993). *Human Development Report 1993*. New York: Oxford University Press.

UNDP (1994). *Human Development Report 1994*. New York: Oxford University Press.

UNDP (2010). *Human Development Report 2010*. New York: Oxford University Press.

Verbunt, P., & Rogge, N. (2018). Geometric composite indicators with compromise benefit-of-the-doubt weights. *European Journal of Operational Research*, *264*, 388–401.

Vidoli, F., & Mazziotta, C. (2013). Robust Weighted Composite Indicators by means of Frontier Methods with an Application to European Infrastructure Endowment. *Italian Journal of Applied Statistics*, *23*, 259–282.

7. Appendix

Table 5: Basic indicators of HDI: summary statistics.

	Min	Max	1^{st} Quartile	Median	3^{rd} Quartile	Mean	cv	Skewness β_1	Kurtosis β_2
Life expectancy	53.280	84.860	67.440	74.050	77.910	72.712	0.102	-0.557	2.614
Expected years of schooling	5.005	21.954	11.431	13.188	15.227	13.325	0.221	-0.116	3.101
Mean years of schooling	1.644	14.152	6.437	9.032	11.326	8.728	0.354	-0.313	2.033
GNI	753.909	131,031.590	4,910.208	12,707.366	29,497.232	20,219.726	1.050	1.773	6.971

Table 6: GMMs selection: number of components; models with different parameters; BIC values.

	EII	VII	EEI	VEI	EVI	VVI	EEE	VEE	EVE	VVE	EEV	VEV	EVV	VVV
2	840.946	835.776	858.003	853.111	853.205	848.101	1030.637	1047.837	1046.161	1045.900	1016.490	1033.568	1031.710	1030.930
3	974.904	997.624	982.684	1014.812	967.135	1007.616	1022.705	**1067.988**	1027.522	1062.609	1000.760	1034.480	995.545	1042.169
4	1005.753	1029.549	1003.283	1023.699	992.879	1004.715	1035.686	1054.641	1011.278	1029.950	996.775	1035.011	970.561	1007.836
5	1008.138	1032.999	1007.501	1031.028	990.206	1000.100	1017.996	1048.847	1001.968	1014.633	975.195	1007.256	945.789	988.947
6	997.347	1019.907	986.491	1022.129	958.719	1000.366	992.389	1034.255	978.217	NA	962.8670	975.257	911.320	968.946
7	991.899	1006.758	970.082	1013.469	942.256	989.419	956.184	1014.235	943.940	943.314	926.137	962.505	910.540	920.952
8	975.488	994.326	977.077	998.163	945.645	959.482	981.783	989.749	936.348	926.987	922.107	934.277	866.786	889.554
9	954.911	988.877	968.653	987.393	918.274	927.783	976.577	971.767	908.653	927.983	894.549	907.483	820.821	840.943

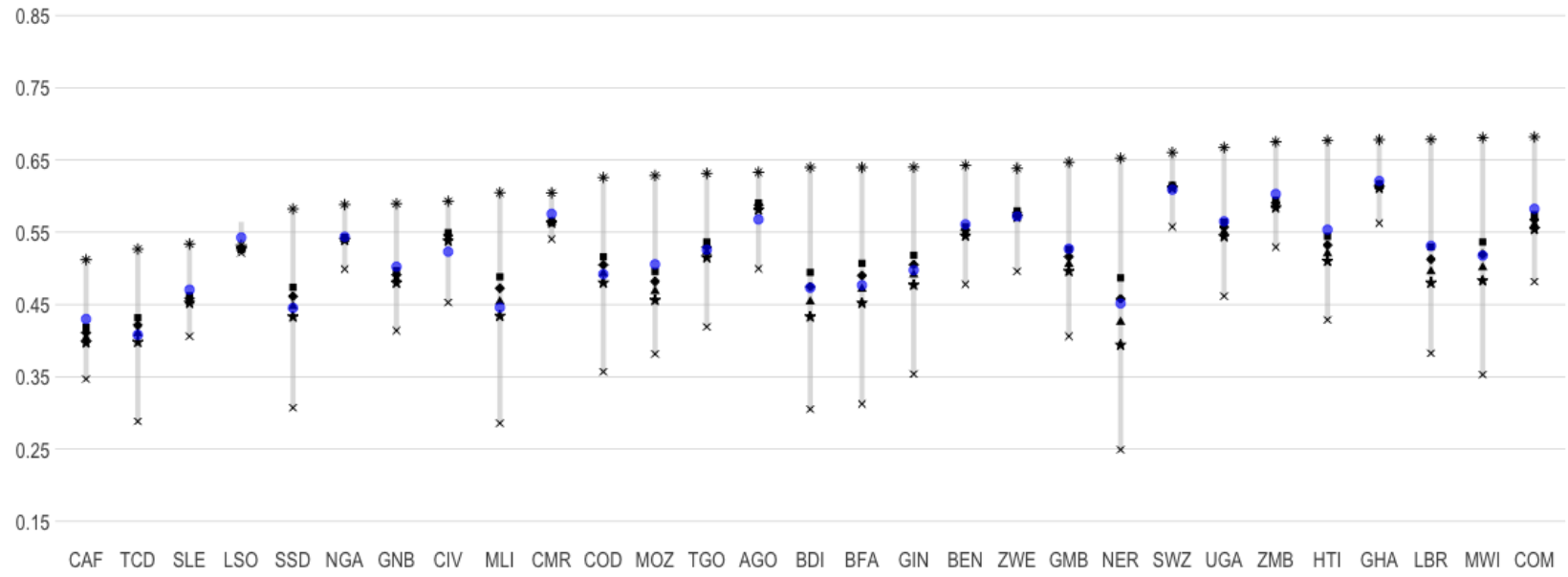


Figure 5: Countries of cluster 1-A according to dimensional indices of HDI: — $BoD - Min$ interval of synthesis; • mid aggregation point - *map*; × minimum; ★ geometric mean (HDI of UNDP); ▲ arithmetic mean; ♦ quadratic mean; ■ cubic mean; ∗ maximum. Year 2019.

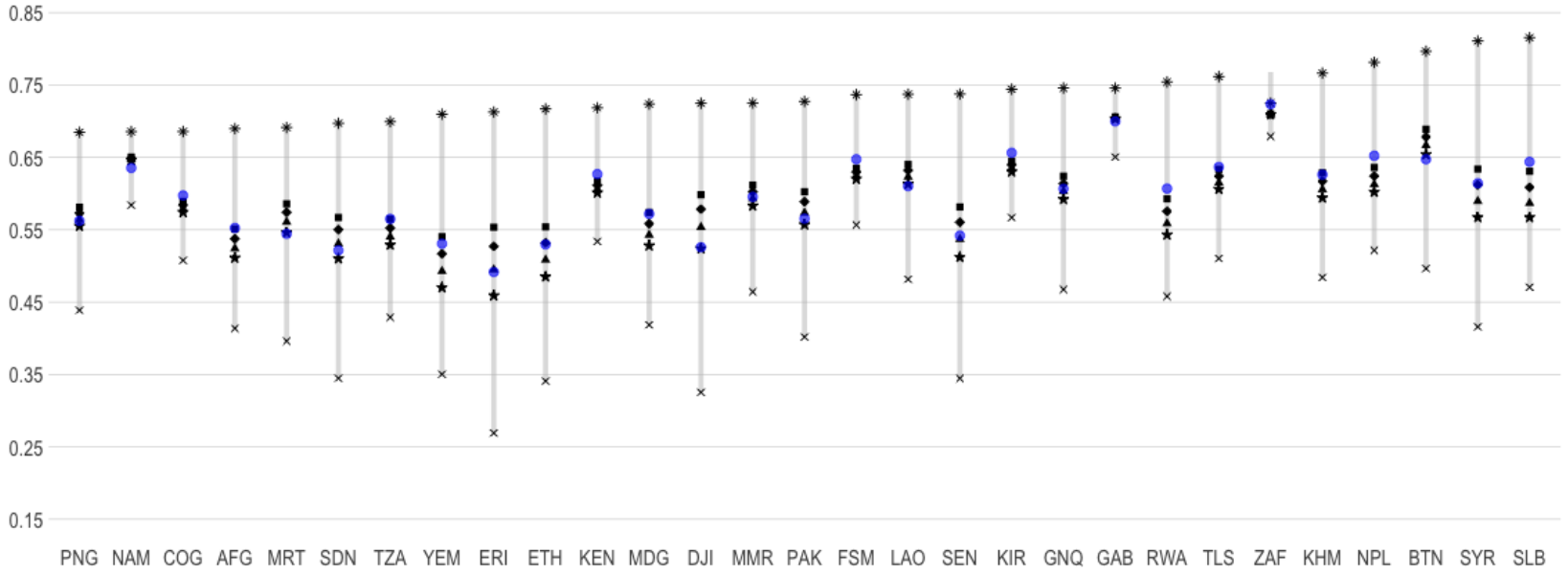


Figure 6: Countries of cluster 1-B according to dimensional indices of HDI: — $BoD - Min$ interval of synthesis; • mid aggregation point - *map*; × minimum; ★ geometric mean (HDI of UNDP); ▲ arithmetic mean; ♦ quadratic mean; ■ cubic mean; ∗ maximum. Year 2019.

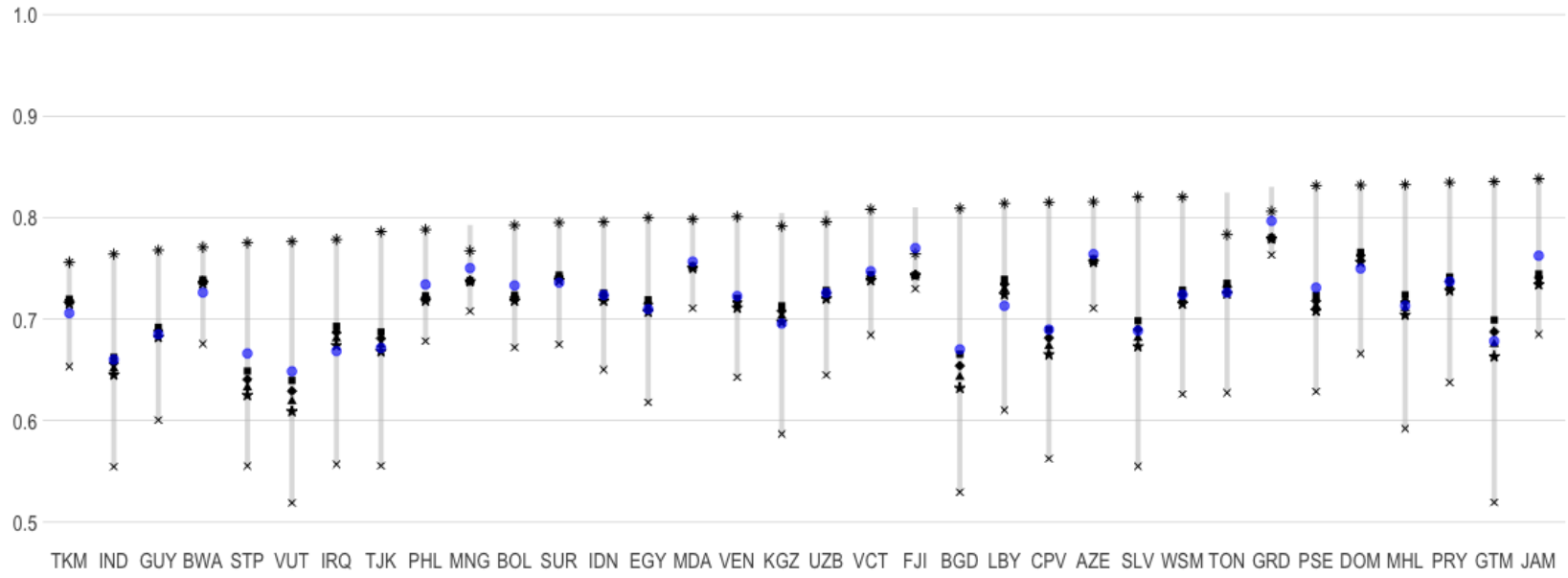


Figure 7: Countries of cluster 2-A according to dimensional indices of HDI: — $BoD - Min$ interval of synthesis; • mid aggregation point - *map*; × minimum; ★ geometric mean (HDI of UNDP); ▲ arithmetic mean; ♦ quadratic mean; ■ cubic mean; ∗ maximum. Year 2019.

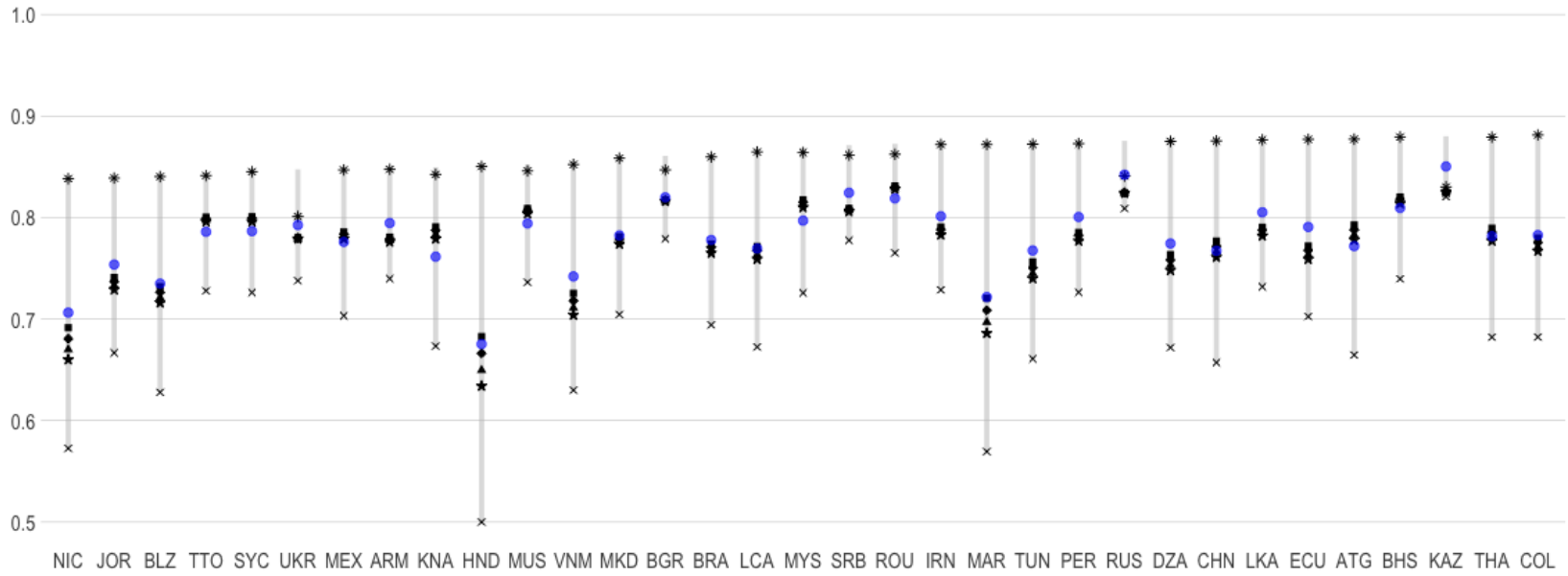


Figure 8: Countries of cluster 2-B according to dimensional indices of HDI: — $BoD - Min$ interval of synthesis; • mid aggregation point - *map*; × minimum; ★ geometric mean (HDI of UNDP); ▲ arithmetic mean; ♦ quadratic mean; ■ cubic mean; ∗ maximum. Year 2019.

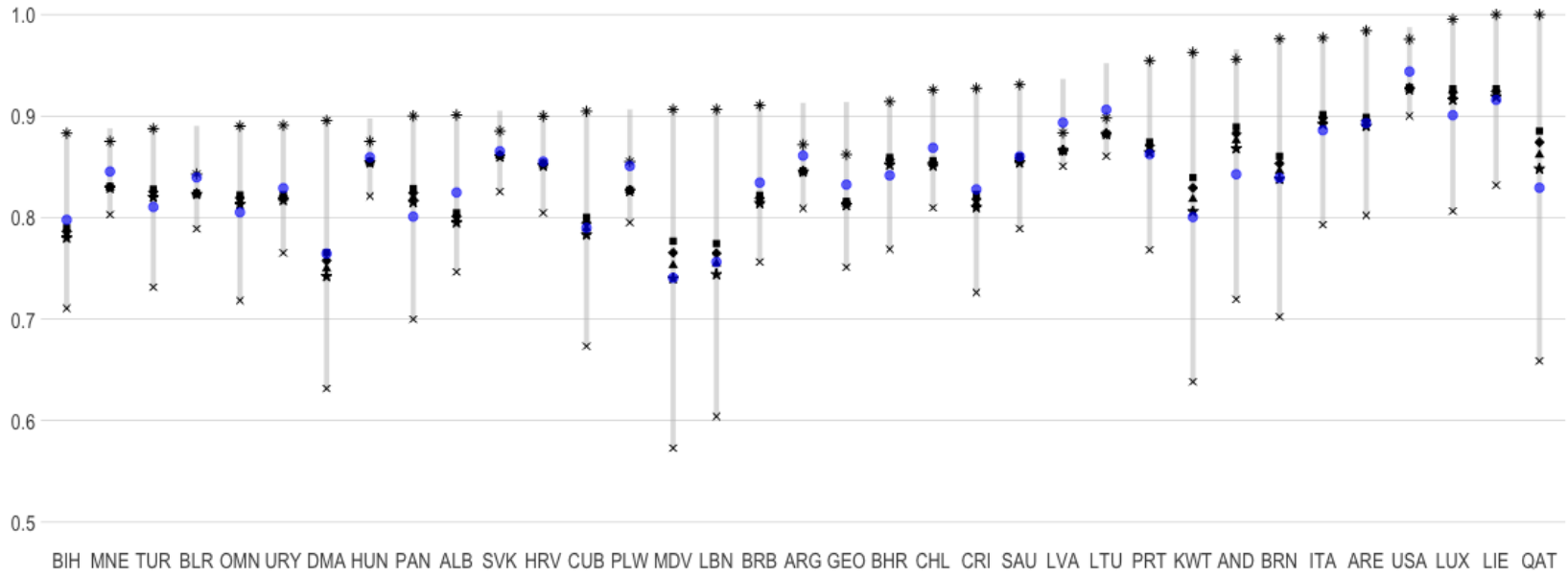


Figure 9: Countries of cluster 2-C according to dimensional indices of HDI: — $BoD - Min$ interval of synthesis; • mid aggregation point - *map*; × minimum; ★ geometric mean (HDI of UNDP); ▲ arithmetic mean; ♦ quadratic mean; ■ cubic mean; ∗ maximum. Year 2019.

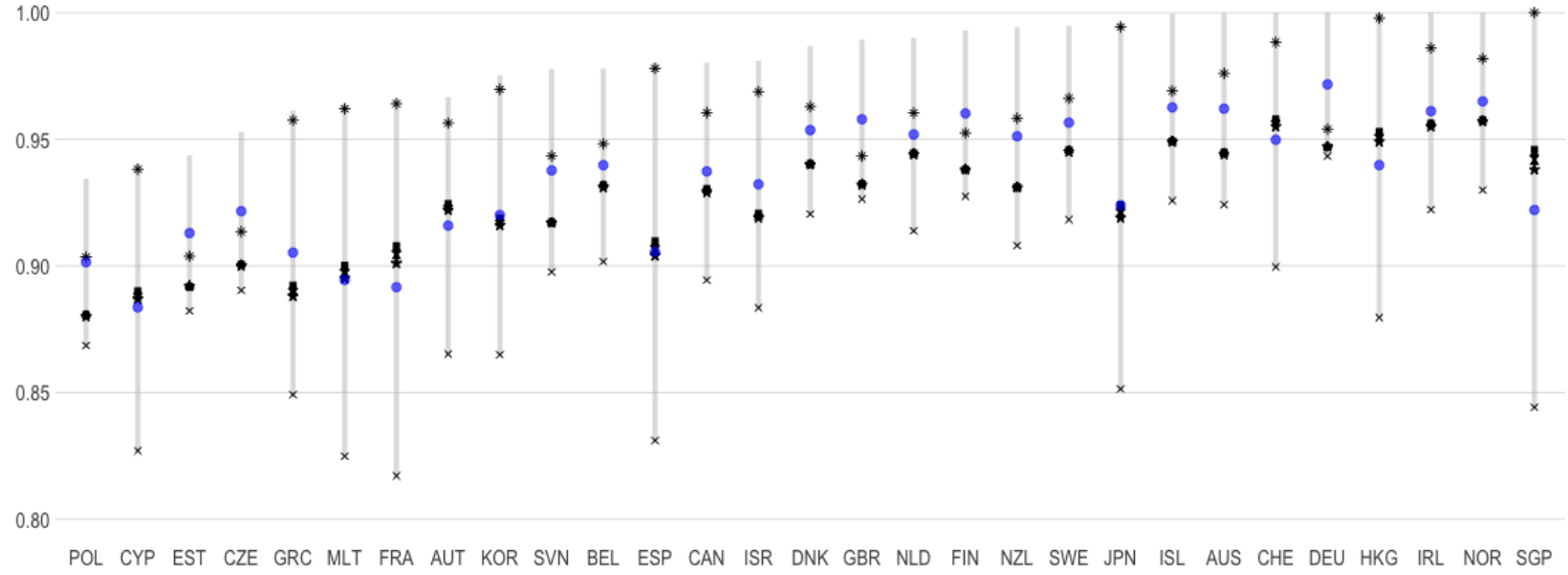


Figure 10: Countries of cluster 3 according to dimensional indices of HDI: — $BoD - Min$ interval of synthesis; • mid aggregation point - *map*; × minimum; ★ geometric mean (HDI of UNDP); ▲ arithmetic mean; ♦ quadratic mean; ■ cubic mean; ∗ maximum. Year 2019.