POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Clustering Nations by Educational Performance

Course: Bayesian Statistics
Professor: Alessandra Guglielmi

Authors: **Gaia Caringi, Ignacio Cunado, Alice Flamigni, Sara Fregnan, Francesca Gamba, Marco Ronchetti**

Tutors: Alessandro Carminati, Alessandra Ragni
Academic Year: 2024-25

Abstract

Educational attainment plays a crucial role in national development, influencing economic growth, social mobility, and overall societal well-being. However, significant disparities in student performance persist both within and across countries.
This study aims to identify data-driven clusters of nations based on the proportion of low-achieving students, defined as those with math proficiency below Level 3 in the OECD PISA 2022 assessment using Bayesian models.

To achieve this, we first employ a Generalized Linear Mixed Model (GLMM) combined with a Mixture of Dirichlet Processes (MDP). The GLMM accounts for hierarchical structures and random effects, while the MDP enables flexible clustering by allowing the number of clusters to be inferred from the data rather than being pre-specified.
This approach successfully identifies 13 clusters among 42 countries, revealing clear geographic and socio-economic patterns, with European countries typically exhibiting better educational outcomes compared to South American nations.

Building on this, we extend our analysis using a Nested Dirichlet Process (NDP), which introduces an additional level of clustering to capture hierarchical dependencies between schools and countries. This model enables simultaneous clustering at multiple levels, distinguishing groups of countries that share similar distributions of student performance rather than just similar average scores. Moreover, the NDP formulation adds a grouping structure of the schools inside each country, providing a more refined classification. Through this approach, 9 clusters were identified, overall sharing consistent features with the previous ones.

Our results highlight the impact of key factors such as student-teacher ratios, school size, and socio-economic status on educational performance. By leveraging Bayesian nonparametric methods, our study presents a flexible and interpretable framework for analyzing global disparities in education, offering insights that can support data-driven policy interventions aimed at reducing educational inequality.

Contents

1 | Introduction

Educational attainment is a fundamental pillar of national development, influencing economic growth, social mobility, and overall quality of life. However, disparities in student performance within and across countries pose significant challenges.
This project employs Bayesian statistical methods to cluster nations all around the world based on the percentage of low-achieving students in their schools, where a low-achieving student is defined as having a math proficiency below Level 3.

1.1. Dataset

From the OECD Program for International Student Assessment (PISA), we selected a dataset that includes information from both the student questionnaire (administered to students) and the school questionnaire (administered to school principals).
In particular, PISA tests the skills and knowledge of 15-year-old students in mathematics, reading and science. Eighty-one countries took part in the 2022 assessment, which focused on mathematics and the data were released by the OECD on 5 December 2023.
Variables explanation:

- CNT = country name
- STRATIO = (# of students) / (# of teachers)
- PV1MATH = plausible value 1 of students in math
- mean-ESCS = mean of the socio-economic status of students
- PRIVATESCH = explains whether a school is public or private
- SCHSIZE = describes the size of a school
- SCHTESTED = number of students tested in that school
- Y-MATH1-rate = rate of low-achieving students related to plausible value 1
- Y-MATH1 = number of low-achieving students related to plausible value 1
- PV1-MATH = average school score in math test

2 | Bayesian semi-parametric approach

We adopted in our analysis a methodological framework that captures both structured dependencies and distributional flexibility. Specifically, we employed a Generalized Linear Mixed Model (GLMM), which effectively accounts for hierarchical structures and random effects, alongside a Mixture of Dirichlet Processes, which provides a flexible, non-parametric way to model complex distributions.

This combination allowed us to refine inference and enhance predictive performance, resulting in a Bayesian semi-parametric approach that balances interpretability with adaptability.

2.1. Generalized Linear Mixed Models

A Generalized Linear Mixed Model (GLMM) is an extension of Generalized Linear Models that incorporates random effects, making it suitable for analyzing hierarchical or correlated data.

Specifically, for LME models, we assume:

$$\boldsymbol{y_j}|\boldsymbol{\beta}, \boldsymbol{b_j} \sim \mathcal{N}_{n_j}\left(\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{b_j}, \sigma^2\boldsymbol{I}_{n_j}\right)$$

If we now consider a sampling distribution for y_{ji} from an exponential family, we obtain:

$$p(y_{ji}|\theta_{ji}, \tau) \propto \exp\left(\tau\left[y_{ji}\theta_{ji} - a(\theta_{ji})\right] + c(y_{ji}, \tau)\right) \quad (*)$$

In the GLMM, the canonical parameter θ_{ji} is linked to the covariates through a link function $h(\cdot)$, which defines the relationship:

$$h(\theta_{ji}) = \eta_{ji} = \boldsymbol{x_{ji}}^\top\boldsymbol{\beta} + \boldsymbol{z_{ji}}^\top\boldsymbol{b_j}$$

where: $E[y_{ji}|\boldsymbol{\beta}, \boldsymbol{b_j}] = \eta_{ji}$.
For GLMM logistic regression, the probability function takes the form:

$$p(y_{ji}|\boldsymbol{\beta}, \boldsymbol{b_j}, \tau) = \exp\left(y_{ji}\left(\boldsymbol{x}_{\boldsymbol{ji}}^\top \boldsymbol{\beta} + \boldsymbol{z}_{\boldsymbol{ji}}^\top \boldsymbol{b_j}\right) - \log\left(1 + \exp\left(\boldsymbol{x}_{\boldsymbol{ji}}^\top \boldsymbol{\beta} + \boldsymbol{z}_{\boldsymbol{ji}}^\top \boldsymbol{b_j}\right)\right)\right)$$

where:

$$\theta_{ji} = \eta_{ji} = \boldsymbol{x}_{\boldsymbol{ji}}^\top \boldsymbol{\beta} + \boldsymbol{z}_{\boldsymbol{ji}}^\top \boldsymbol{b_j}, \quad \tau = 1$$

Moreover we will assume

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$$
$$\boldsymbol{b_j} \sim G$$

2.2. Mixture of Dirichlet Process models

A Mixture of Dirichlet Processes (MDP) is a non-parametric probabilistic model used in clustering problems where the number of clusters is unknown and inferred from the data. It extends the concept of a Dirichlet Process (DP), which defines a distribution over distributions, allowing for an infinite mixture model.

In an MDP, clusters are not fixed but are generated dynamically as new data is observed. Each observation belongs to a probabilistic cluster, influenced by both the data distribution and the probability of forming new clusters. The semi-parametric model can be summarized in:

$$\begin{aligned} &(1) \quad x_j|b_j \sim D_{nj}(g(b_j)) \\ &(2) \quad b_j \mid G \overset{\text{iid}}{\sim} G \\ &(3) \quad G \mid M \sim DP(MG_0) \end{aligned}$$

where

- (1) represents the fully parametric part, while (2) and (3) the non-parametric one.
- D_s is the s-dimensional parametric distribution
- G is a general distribution
- M is a positive scalar called precision parameter: if M is large G is highly concentrated around G_0.

An important property of the Dirichlet process is the discrete nature of G, that can be written as

$$\sum_{h=1}^{\text{inf}} w_h \delta_{m_h}()$$

where w_h are probability weights and δ_x denotes the Dirac measure at x.
Under this hierarchical model, the posterior distribution on G is a mixture of DP's, i.e., $p(G|y_1, ..., y_n)$ is a mixture of DP models, mixing with respect to the latent b_j.

2.3. Our model

The innovation of the Bayesian semi-parametric approach consists of replacing the normal prior on the random effects with a non-parametric prior, followed by a Dirichlet process prior on the general distribution.
We decided to model our problem through a Poisson regression:

$$\begin{aligned} y_{ji} \mid \boldsymbol{\beta}, b_j &\overset{\text{ind}}{\sim} \mathcal{P}\left(\exp\left(\boldsymbol{X}_{ji}^T\boldsymbol{\beta} + b_j + \log(T_{ji})\right)\right) \\ \boldsymbol{\beta} &\sim \mathcal{N}_p(0, \sigma_\beta^2 \boldsymbol{I}_p) \\ b_j \mid G &\overset{\text{iid}}{\sim} G \\ G &\sim DP(M, G_0) \\ G_0 &\sim \mathcal{N}(\mu_0, \sigma_0^2) \end{aligned}$$

where:

- j : country $j = 1, \dots, J$
- i : school $i = 1, \dots, n_j$
- J : the number of countries in the dataset
- n_j : number of schools in country j
- $\boldsymbol{X_{ji}}$: the vector of p covariates for school i in country j
- y_{ji} : the rate of low-achieving students that did the test in school i in country j
- $\boldsymbol{I_p}$: identity matrix of dimension p
- M is the precision parameter
- T_{ji}: number of students in school i in country j
- $log(T_{ji})$: offset
- μ_0, σ_0^2, σ_β^2 are fixed values

It's clear to see that, considering a random variable $Y \sim P(\lambda)$, then

$$\mathbb{P}(Y = k|\lambda) = (\lambda^k e^{-\lambda})/(k)! = exp(klog(\lambda) - \lambda - log(k!))$$

In our specific case:

$$P(y_{ji} = k \mid \boldsymbol{\beta}, b_j) = \exp\{k(\mathbf{X}_{ji}^T\boldsymbol{\beta} + b_j + \log(T_{ji})) - e^{\mathbf{X}_{ji}^T\boldsymbol{\beta}+b_j+\log(T_{ji})} - \log(k!)\}$$

and comparing this formula with (*) it's easy to see that, in our model,

$$\tau = 1$$

$$\theta_{ji} = \mathbf{X}_{ji}^T\boldsymbol{\beta} + b_j + log(T_{ji})$$

$$a(\theta_{ji}) = e^{\theta_{ji}}$$

$$c(y_{ji}, \tau) = -log(y_{ji}!)$$

In our model the offset is necessary since the target variable y_{ji} is the rate of low-achieving students, so it is needed to normalize the counting based on the number of students.

2.4. Sampling procedure and algorithm

We implemented our model in Julia using the Turing package for probabilistic modeling. In fact, although Julia and Python share many similarities, Julia has some key advantages:

- The Turing package in Julia provides an equally powerful alternative to probabilistic programming frameworks in Python but with **greater flexibility**.
- Julia offers **superior computational performance**, especially in terms of speed, which is crucial for Bayesian modeling and large-scale simulations.

Dirichlet Process Construction via Stick-Breaking

In order to implement our model in Julia, we required a practical approach to construct the Dirichlet Process (DP). To achieve this, we explicitly formulated the stick-breaking process, an iterative method that sequentially divides a unit-length "stick" into smaller segments, each representing a probability mass.

Stick-Breaking Construction

- **Initialization:** Start with a stick of length 1, representing the total probability mass.
- **First Component Assignment:** Sample a fraction v_1 from a Beta$(1, \alpha)$ distribu-

tion:

$$v_1 \sim \text{Beta}(1, \alpha)$$

The probability assigned to the first component is: $p_1 = v_1$

- **Subsequent Component Assignments:** For each new component k, sample another fraction v_k from the same Beta distribution:

$$v_k \sim \text{Beta}(1, \alpha)$$

The probability of each new component is computed as:

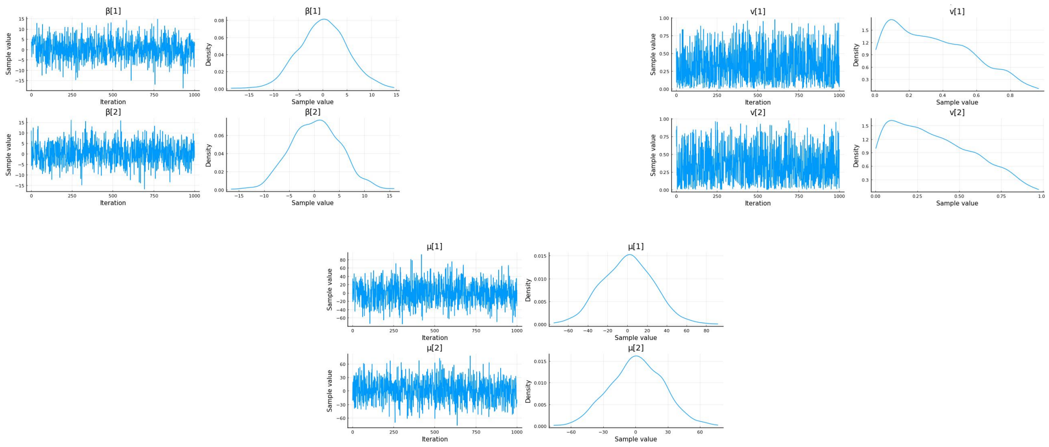
$$p_k = v_k \prod_{i=1}^{k-1} (1 - v_i)$$

- **Iteration:** This process continues until the total probability mass is fully assigned.

This construction was necessary to implement our model in Julia, ensuring a flexible and computationally efficient approach to represent the Dirichlet Process.

Convergence of the chains

The MCMC trace plots show a converging dynamics, for all covariate coefficients β, all the components of the stick breaking process v and the random effect μ, as shown in the plot below.



2.5. Results

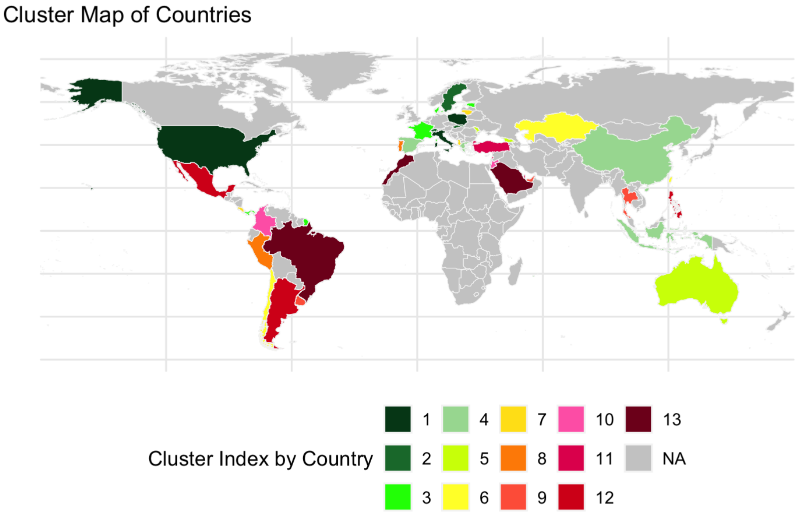


Figure 2.1: Clusters obtained through the Dirichlet Process (Poisson model)

We identified 13 clusters among our 42 nations and ranked them based on their average response variable, which is the rate of low achieving students.

Specifically, we calculated the mean of the response for all countries within each cluster, providing a representative response value for each group. To enhance interpretability, we assigned colors to the clusters, using a gradient from dark green,indicating the lowest proportion of low-achieving students, to burgundy, representing the highest. From the figure, we observe that European countries tend to belong to clusters of similar performance levels. Likewise, all South American countries are grouped into clusters that are closely aligned with each other.

3 | Nested Dirichlet process

3.1. What is a Nested Dirichlet Process

The NDP is a DP in which the baseline measure is itself a Dirichlet process generating probability distributions.

Let us consider y_{ji} the average score of students in school i and country j. We assume that $y_{ji} \overset{ind}{\sim}$ F$_j$ with j = 1,...J. The final aim will be clustering F$_j, j = 1, ...J$.

Consider a collection of distributions $\{G_1, \ldots, G_J\}$. The NDP assumes

$$G_j \mid Q \sim Q, \quad j = 1, \ldots, J$$
$$Q \sim \text{DP}(\alpha\text{DP}(\gamma G_0))$$

where DP denotes a Dirichlet process, G_0 is a non-atomic baseline probability measure, and $\alpha, \gamma > 0$ are the total mass parameters.
The discrete nature of Q is highlighted in

$$Q = \sum_h w_h \delta_{G_h^*}$$

with $G_h^* \overset{iid}{\sim} \text{DP}(\gamma G_0)$.

Similarly,

$$G_h^* = \sum_f v_f \delta_{\theta_f^*}$$

with $\theta_f^* \overset{iid}{\sim} G_0$.
The weights w_h in Q are generated with the total mass parameter α of the outer DP, and the weights v_f in G_h^* are generated with the total mass parameter γ of the nested, inner DP.

Since we have to work with a continuous distribution, we use an additional convolution with a continuous kernel $p(\cdot \mid \theta)$,

$$F_j(\cdot) = \int_\theta p(\cdot \mid \theta) G_j(d\theta), \tag{1}$$

Here $p(\cdot \mid \theta)$ is a sampling model for y_{ji} and θ is the finite dimensional parameter associated with the sampling model. The collection $\{F_1, \dots, F_J\}$ is said to follow a Nested Dirichlet process (NDP) mixture first introduced in Rodriguez et al. (2008).
The general model is

$$\begin{aligned} y_{ji} \mid \boldsymbol{\theta_{ji}} &\overset{\text{ind}}{\sim} p(y_{ji} \mid \boldsymbol{\theta_{ji}}) \\ \boldsymbol{\theta_{ji}} \mid G_j &\overset{\text{ind}}{\sim} G_j \\ G_j &\overset{\text{iid}}{\sim} Q \\ Q &\sim \text{DP}(\alpha \text{DP}(\gamma G_0)) \end{aligned}$$

This model gives rise to two levels of clustering:

- The discrete nature of Q induces ties among G_j. Denoting with $\{G_1^*, \dots, G_K^*\}$ the unique elements among the G_j, it is possible to identify clusters defined by the configuration of these ties:
$$S_k = \{j \mid G_j = G_k^*\}$$
We will refer to S_k, $k = 1, \dots, K$ as *distributional clusters.*
- The discrete nature of G_k^* induces ties among:
$$\{\boldsymbol{\theta_{ji}} \mid j \in S_k, \quad i = 1, \dots, I_j\}$$
Denoting with $\{\boldsymbol{\theta^*_{k1}}, \dots, \boldsymbol{\theta^*_{kL_k}}\}$ the L_k unique elements, the following clusters can be identified:
$$R_{k,l} = \{(j,i) \mid s_j = k, \quad \boldsymbol{\theta_{ji}} = \boldsymbol{\theta^*_{kl}}\}$$
The sets $R_{k,l}$ are referred to as *observational clusters.*

It is useful to introduce the following alternative representation: the vectors

$$\boldsymbol{s} = (s_1, \dots, s_J) \quad \text{and} \quad \boldsymbol{r_j} = (r_{j,1}, \dots, r_{j,I_j})$$

denote membership indicators for the distributional and observational clusters, respectively.

It is possible to show that:

$$\begin{aligned} \boldsymbol{r_j} \mid \boldsymbol{s} &\overset{\text{iid}}{\sim} \text{PU}(\gamma) \\ \boldsymbol{s} &\sim \text{PU}(\alpha) \end{aligned}$$

where PU denotes the Polya Urn representation.

3.2. Our model

With this second approach we aim aim to model the average students' score of school i, in country j, here denoted as y_{ji}.

The final model becomes:

$$\begin{aligned} y_{ji} \mid \boldsymbol{\theta_{ji}}, \boldsymbol{\beta_j} &\overset{\text{ind}}{\sim} \mathcal{N}\big(\mu_{ji} + \boldsymbol{X_{ji}^T}\boldsymbol{\beta_j}, \sigma^2_{ji}\big), & j = 1, \ldots, J; \quad i = 1, \ldots, I_j \\ \boldsymbol{\beta_j} &\overset{\text{iid}}{\sim} \mathcal{N}_p(0, \text{diag}(\sigma^2_\beta)), & j = 1, \ldots, J \\ \boldsymbol{\theta_{ji}} \mid G_j &\overset{\text{ind}}{\sim} G_j, & j = 1, \ldots, J \\ G_j &\overset{\text{iid}}{\sim} Q, & j = 1, \ldots, J \\ Q &\sim \text{DP}(\alpha \text{DP}(\gamma G_0)) \\ G_0 &\sim \text{NIG}(\mu_0, \lambda, a, b) \end{aligned}$$

where $\boldsymbol{\theta_{ji}} = (\mu_{ji}, \sigma^2_{ji})$.

In our case the conjugate baseline measure is

$$G_0(\boldsymbol{\theta}^*_{k\ell}) = \text{NIG}(\mu_0, \lambda, a, b).$$

We choose a normal sampling model:

$$y_{ji} \mid \boldsymbol{\theta_{ji}}, \beta_j \overset{\text{ind}}{\sim} \mathcal{N}(\mu_{ji} + \boldsymbol{X_{ji}^T}\boldsymbol{\beta_j}, \sigma^2_{ji}), \quad \text{where} \quad \boldsymbol{\theta_{ji}} = (\mu_{ji}, \sigma^2_{ji}).$$

We have introduced 4 covariates $\boldsymbol{X_{ji}}$, which are STRATIO, SCHSIZE, SCHTESTED, meanESCS, and the corresponding coefficients $\boldsymbol{\beta_j}$.

3.3. Sampling Procedure and Algorithm

Using the algorithm described in the paper by Zuanetti,Müller, Zhu,Yang, Ji (2017) we implement MCMC simulation for the model:

$$\begin{aligned} y_{ji} \mid s_j = k, r_{ji} = \ell &\overset{\text{ind}}{\sim} \mathcal{N}(\mu_{ji} + \boldsymbol{X}_{\boldsymbol{ji}}^T \boldsymbol{\beta_j}, \sigma_{ji}^2) \\ \boldsymbol{\theta}_{k\ell}^* &\sim \text{NIG}(\mu_0, \lambda, a, b) \\ \boldsymbol{r_j} \mid \boldsymbol{s} &\overset{\text{iid}}{\sim} \text{PU}(\gamma) \\ \boldsymbol{s} &\sim \text{PU}(\alpha) \end{aligned}$$

From now on we will denote $y_{ji} - \boldsymbol{X_{ji}}^T \boldsymbol{\beta}$ as y_{ji}.

Updating $\theta_{k\ell}^*$

As the NIG base measure $G_0(\boldsymbol{\theta}_{k\ell}^*)$ is conjugate to the normal kernel $p(y_{ji} \mid \boldsymbol{\theta}_{k\ell}^*)$, the marginal distribution of y_{ji} and the complete conditional posterior distribution of $(\boldsymbol{\theta}_{k\ell}^*)$ and r_{ji} are available in closed form. We can therefore define Gibbs sampling transition probabilities to update $\boldsymbol{\theta}_{k\ell}^*$ by drawing from the complete conditional posterior distribution. For given $\boldsymbol{s}$ and $\boldsymbol{r_j}$, the complete conditional posterior distribution of $\boldsymbol{\theta}_{k\ell}^*$ is given by:

$$p(\boldsymbol{\theta}_{k\ell}^* \mid \dots) = \text{NIG}(\mu_{k\ell}, \lambda_{k\ell}, a_{k\ell}, b_{k\ell}) \tag{6}$$

with

$$\mu_{k\ell} = \frac{\bar{y}_{k\ell} m_{k\ell} + \lambda \mu_0}{m_{k\ell} + \lambda}, \quad \lambda_{k\ell} = \lambda + m_{k\ell}, \quad a_{k\ell} = a + \frac{m_{k\ell}}{2},$$
$$b_{k\ell} = b + \frac{s_{k\ell}^2 m_{k\ell} + \frac{m_{k\ell}\lambda}{m_{k\ell}+\lambda}(\bar{y}_{k\ell} - \mu_0)^2}{2},$$

where

$$\bar{y}_{k\ell} = \frac{\sum_{(j,i) \in R_{k\ell}} y_{ji}}{m_{k\ell}}, \quad s_{k\ell}^2 = \frac{\sum_{(j,i) \in R_{k\ell}} (y_{ji} - \bar{y}_{k\ell})^2}{m_{k\ell}}$$

are the cluster-specific sample means and (scaled) variances.

Updating r_{ji}

The observational cluster indicator $\boldsymbol{r_{ji}}$, for $j = 1, \dots, J$ and $i = 1, \dots, I_j$, is drawn using its complete conditional posterior distribution:

$$p(r_{ji} = \ell \mid r_{-ji}, s_j = k, \boldsymbol{y}, \boldsymbol{\theta}^*) \propto \begin{cases} m_{k\ell}^- \mathcal{N}(y_{ji} \mid \boldsymbol{\theta}_{k\ell}^*), & \text{for } \ell = 1, \dots, L_k^- \\ \gamma h_0(y_{ji}), & \text{for } \ell = L_k^- + 1. \end{cases}$$

where $h_0(y_{ji})$ is a Student's t distribution evaluated in y_{ji} with $2a$ degrees of freedom:

$$h_0(y_{ji}) = t\left(y_{ji} \mid 2a, \mu_0, \frac{b(1+\lambda)}{a\lambda}\right)^{\frac{1}{2}}$$

At each step of the algorithm, a new configuration of the observational clusters is computed. In particular, the second line allows to generate a new observational cluster in the k-th distributional cluster; in that case, $\boldsymbol{\theta}^*_{k\ell}$ will be updated as shown in the previous step. At each iteration, empty observational clusters in the new configuration of $\mathbf{r}$ and their respective parameters are excluded from the model. The remaining observational clusters are relabeled, and the values of L_k are recalculated.

Updating s_j

Updating s_j requires a more intricate procedure, as reassigning $\boldsymbol{y_j}$ to a different distributional cluster necessitates the simultaneous reassignment of the observational clustering r_{ji}. Consequently, a joint update of s_j and r_{ji} is required. To update the cluster indicators, the Metropolis-Hastings algorithm is employed: given the current configuration $\boldsymbol{z} = (s_j, \mathbf{r}_j)$, the new proposed configuration is accepted with probability:

$$\psi(\boldsymbol{z}' \mid \boldsymbol{z}) = \min(1, A')$$

with:

$$A' = \frac{p(\boldsymbol{z}' \mid y, \mathbf{r}_{-j}, s_{-j})q(\boldsymbol{z} \mid \boldsymbol{z}')}{p(\boldsymbol{z} \mid y, \mathbf{r}_{-j}, s_{-j})q(\boldsymbol{z}' \mid \boldsymbol{z})}$$

where:

- $p(\boldsymbol{z} \mid y, \mathbf{r}_{-j}, \boldsymbol{s}_{-j})$ represents the marginal posterior probability of the cluster indicators:

  $$p(s_j = k, \mathbf{r}_j \mid \mathbf{y}, \mathbf{r}_{-j}, \boldsymbol{s}_{-j}) \propto p(s_j = k \mid \boldsymbol{s}_{-j})p(\mathbf{r}_j \mid s_j = k, \mathbf{r}_{-j}, \boldsymbol{s}_{-j})p(y_j \mid \mathbf{r}_j, s_j = k, y_{-j}, \mathbf{r}_{-j}, \boldsymbol{s}_{-j})$$

  where all the three distributions are known in closed form, in particular:

  $$p(s_j = k \mid \boldsymbol{s}_{-j}) \propto \begin{cases} n_k^-, & \text{if } k \in \{1, \dots, K^-\} \\ \alpha, & \text{if } k = K^- + 1 \end{cases}$$

- $q(\cdot \mid \cdot)$ is the proposal distribution used to generate a proposal in the Metropolis–Hastings transition probability. The proposal distribution $q(\boldsymbol{z}'|\boldsymbol{z})$ is constructed in two steps. First, we sample s'_j from its conditional prior distribution given s_{-j}

(the current configuration excluding j) $p(s_j = k \mid \boldsymbol{s}_{-j})$.

As a second step, the new observational cluster r'_{ji} is sampled from:

$$p(r'_{ji} = \ell \mid \boldsymbol{r_{-j}}, r'_{j1}, \dots, r'_{ji-1}, s'_j = k, y_{ji}, \boldsymbol{\theta^*}) \propto \begin{cases} m_{k\ell}^{-(r_{ji}\dots r_{jI_j})} \mathcal{N}(y_{ji} \mid \boldsymbol{\theta^*_{k\ell}}), & \text{for } \ell = 1, \dots, L_k^- \\ \gamma t \left(y_{ji} \mid 2a, \mu_0, \frac{b(1+\lambda)}{a\lambda} \right)^{\frac{1}{2}}, & \text{for } \ell = L_k^- + 1. \end{cases}$$

Unlike the step "Updating r_{ji}", in this case, when sampling r'_{ji} we need to take into account that the last $I_j - i + 1$ observations have not been reassigned yet.
Overall, the proposal distribution takes the following form:

$$q(\boldsymbol{z}' \mid \boldsymbol{z}) = p(s'_j = k \mid \boldsymbol{s_{-j}}) \prod_{i=1}^{I_j} p(r'_{ji} = \ell \mid r_{-j}, r_{j1}, \dots, r_{ji-1}, s_j = k, y_{ji}, \boldsymbol{\theta}^*)$$

Once the acceptance of the new configuration is evaluated, empty observational and distributional clusters are removed, the remaining clusters are relabeled, K and the values of L_k are recalculated, and $\boldsymbol{\theta}^*_{k\ell}$ is updated as previously described.

Updating β_j

For $j = 1, \dots, J$, conditioning on the clustering, we have the following model:

$$\boldsymbol{Y_j} \sim \mathcal{N}_{I_j}(\boldsymbol{\mu_j} + \boldsymbol{X_j}\boldsymbol{\beta_j}, \mathrm{diag}(\boldsymbol{\sigma}_j^2))$$

$$\boldsymbol{\beta_j} \sim \mathcal{N}_p(0, \mathrm{diag}(\sigma_\beta^2))$$

We introduce the Moore-Penrose left inverse, supposing that $p \leq I_j$ and that the rank of X_j is I_j:

$$\boldsymbol{X}_j^+ = (\boldsymbol{X}_j^{+T}\boldsymbol{X}_j^+)^{-1}\boldsymbol{X}_j^{+T}$$

This particular $p \times I_j$ matrix is such that $X_j^+ X_j = I_p$. We can use an affine transformation for the first row of our model. The model becomes:

$$\boldsymbol{X}_j^+(\boldsymbol{Y}_j - \boldsymbol{\mu}_j) \sim \mathcal{N}_p\left(\boldsymbol{\beta}_j, \boldsymbol{X}_j^+ \,\mathrm{diag}(\sigma_j^2)\boldsymbol{X}_j^+\right)$$

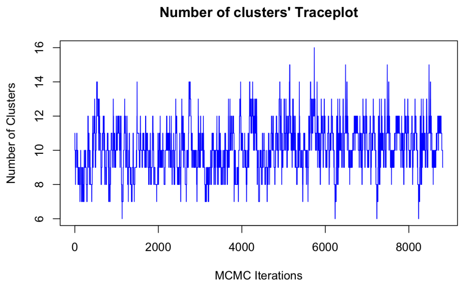
$$\boldsymbol{\beta}_j \sim \mathcal{N}_p(\boldsymbol{0}, \mathrm{diag}(\sigma_\beta^2))$$

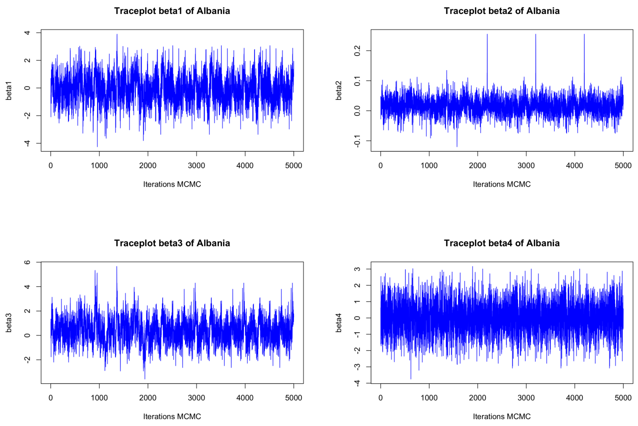
This model is a conjugate multivariate normal - multivariate normal model. Therefore, we can compute the posterior distribution of β_j in closed form:

$$\boldsymbol{\beta}_j \mid \text{rest} \sim \mathcal{N}_p\left(\left(\text{diag}(\sigma_\beta^{-2}) + \boldsymbol{X}_j^+ \,\text{diag}(\sigma_j^2)\boldsymbol{X}_j^+\right)^{-1} \boldsymbol{X}_j^+ \,\text{diag}(\sigma_j^2)(\boldsymbol{X}_j^+)^T \boldsymbol{X}_j^+ (\mathbf{Y}_j - \boldsymbol{\mu}_j),\right.$$

$$\left.\left(\text{diag}(\sigma_\beta^{-2}) + \boldsymbol{X}_j^+ \,\text{diag}(\sigma_j^2)(\boldsymbol{X}_j^+)^T\right)^{-1}\right)$$

The MCMC trace plots show a converging dynamics, both for the number of clusters and for the $\boldsymbol{\beta}$, as shown in the plot below.





3.4. Interpretation of the results

Through the use of this method, we were able to obtain 9 clusters for the 38 countries. To provide a clear and informative representation of the performance of the different clusters, we computed the mean math test score for each cluster. The clusters were then labeled and ranked in descending order based on their performance, with labels assigned according to the quality of the scores achieved. In Figure 3.1, countries are displayed according to their cluster membership, using a color palette that represents cluster performance, ranging from the highest-performing clusters (green) to the lowest-performing ones (red).

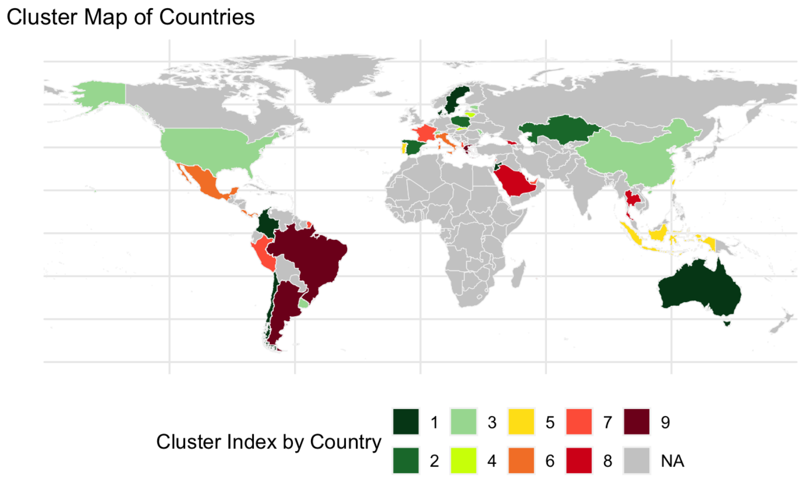
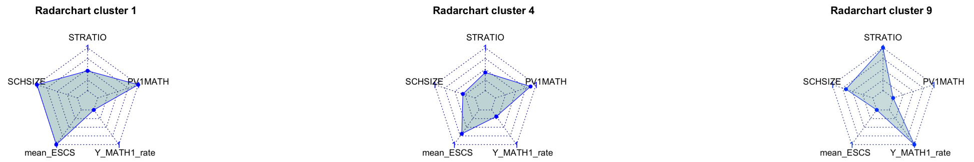


Figure 3.1: Cluster obtained through the Nested Dirichlet Process

To understand which are the characteristics of the countries inside each single cluster, it can be useful to analyze the radarchart of our covariates. Below we show the radarchart of the "best", "intermediate" and "worst" clusters.

The results derived from the radar chart appear reasonable and consistent with expectations. For example, the student-to-teacher ratio, which we would expect to be low in a well-performing school, is indeed low in Cluster 1, while it is significantly higher in Cluster 9. Similarly, the socioeconomic index covariate is highest in what we identify as the "best-performing cluster" and shows a marked decline in the lowest-performing cluster. Another interesting point is the consistency between these radar chart results and those obtained from the first model, as evidenced by the analysis of the YMATH1rate variable in the radar chart.



(a) Radarchart of the "best" cluster

(b) Radarchart of the "intermediate" cluster

(c) Radarchart of the "worst" cluster

Figure 3.2: Radarcharts for different clusters

As an additional analysis, confidence intervals for the $\boldsymbol{\beta}$ were computed. The only significant intervals were found for β_1, which is the coefficient of the variable STRATIO, reported in Figure 3.3 and colored accordingly to the belonging cluster.

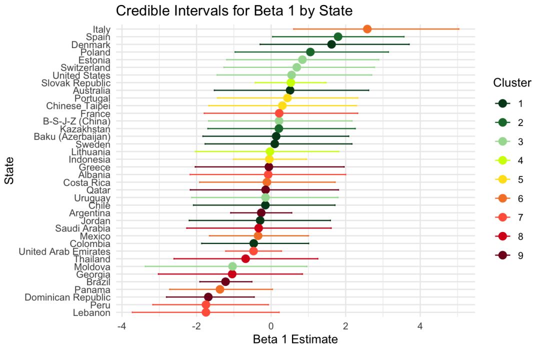


Figure 3.3: Confidence interval for the coefficient of the variable STRATIO

These confidence intervals suggest that, in the higher-performing clusters, the impact of this variable is positive, whereas for the lower-performing clusters, the effect is negative. This finding further supports the importance of the student-to-teacher ratio as a significant factor influencing student performance.

4 | Comparison of different approaches

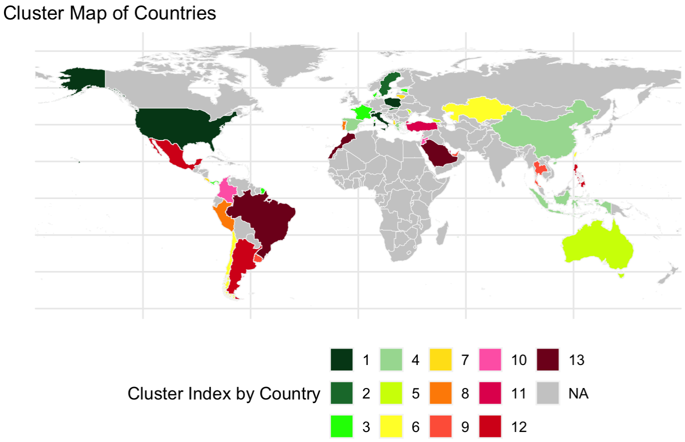


Figure 4.1: Clusters obtained through the Dirichlet Process (Poisson model)

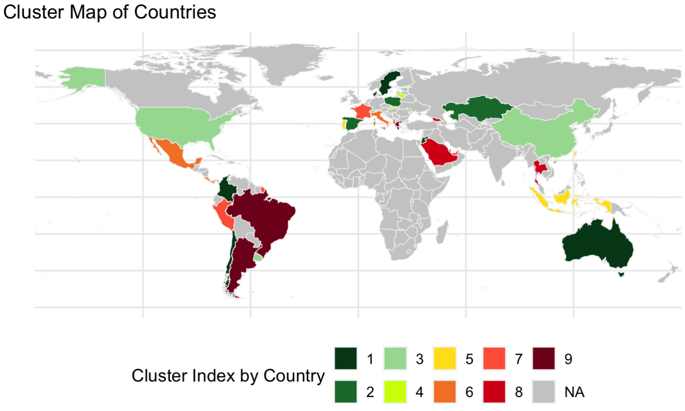


Figure 4.2: Clusters obtained through the Nested Dirichlet Process

The comparison between the cluster maps generated by the Mixture of Dirichlet Processes (MDP) model and the Nested Dirichlet Process (NDP) model highlights key differences in how each approach captures the structure of educational disparities across nations.

The MDP-based clustering provides a data-driven partitioning of countries, grouping them based on their overall proportion of low-achieving students. This method effectively identifies broad performance trends, distinguishing high-performing countries (mostly in Europe) from lower-performing ones (notably in South America and parts of Asia). However, since the MDP treats each nation as an independent unit, it does not account for internal heterogeneity within countries, potentially overlooking variations in school-level distributions.

The NDP-based clustering introduces an additional level of aggregation by allowing within-country variation to play a role in the clustering process. Instead of only grouping nations based on aggregate performance, the NDP captures differences in the distributional structure of schools' performance. As a result, some trends observed in the first model are still present, e.g. the low performance in South America and the medium-high in Europe. Otherwise, some countries that appeared similar under the MDP approach are now assigned to distinct clusters due to their internal diversity in school-level achievement, suggesting that some nations with comparable average performance levels may, in fact, have different underlying school distributions.

The key takeaway from this comparison is that while the MDP approach is effective for identifying macro-level patterns, the NDP approach offers a more detailed, hierarchical perspective, making it particularly useful for understanding both between-country and within-country disparities.

Bibliography

Ken P. Kleinman, Joseph G. Ibrahim (1998). “A semi-parametric Bayesian approach to generalized linear mixed models”. In: *Statist. Med.*, pp. 2579–2596.

Muller, Quintana, Jara and Hanson (2015). "Bayesian Non Parametric Data Analysis". *Springer.*

Abel Rodríguez, David B Dunson and Alan E Gelfand (2008). “The Nested Dirichlet Process”. In: *Journal of the American Statistical Association*, pp. 1131–1154.

Daiane Aparecida Zuanetti, Peter Muller, Yitan Zhu, Shengjie Yang and Yuan Ji (2017). “Clustering Distributions with the Marginalized Nested Dirichlet Process”. In: *The International Biometric Society.*

Neal, Radford M. (2000). “Markov Chain Sampling Methods for Dirichlet Process Mixture Models.” In: *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–65. JSTOR

Chopin, Nicolas, and Omiros Papaspiliopoulos (2020). “An introduction to sequential Monte Carlo.” *Springer*, vol. 4, Berlin.