STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, Month 2026, pp 0–15.
Published online in International Academic Press (www.IAPress.org)

Intelligent and Bayesian Models for Estimating the Transition Probability Matrix of Markov chains: Applications to the Analysis of Sustainable Development Indicators

Rasoul Ali Hussein, Muthanna Subhi Sulaiman*

Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

Abstract In discrete-time Markov chains, correct estimation of the transition probability matrix is required to explain the dynamic behaviour of the system. Nonetheless, classical estimators can be unstable in many cases when the transition data used are sparse or irregular. This paper looks at four methods of estimating the transition probability matrix, and they include the Maximum Likelihood Estimator (MLE), the Bayesian–Dirichlet estimator, the shrinkage estimator, and an intelligent optimization-based estimator based on the Artificial Bee Colony (ABC) algorithm. To test the effectiveness of these estimators, a simulation study was conducted under three transition-density conditions, i.e., dense, moderately sparse, and sparse. In each case, 100 synthetic frequency matrices were generated, and the competing estimators were compared on the basis of the mean squared error (MSE) as the main measure of estimation accuracy. Moreover, to determine the stability and dynamic behaviour of the estimated chains, the variance of the estimated transition matrix elements, the spectral gap, as well as the mixing time were also investigated. The proposed methodology was also applied to real-world data associated with Sustainable Development Goal 4 (Quality Education), which were represented by annual adult literacy rates (percentage of the population aged 15 years and above) over a long-term period. The simulation findings indicate that the ABC-based estimator achieves the lowest estimation error across all transition-density settings, and also has competitive stability and mixing behaviour. The results suggest that intelligent optimization may offer a suitable and scalable alternative for estimating transition probability matrices, particularly in cases where the transition structure is irregular or sparse.

Keywords Artificial bee colony, Bayesian–Dirichlet estimator, maximum likelihood estimator, shrinkage estimator, transition probability matrix.

DOI: 10.19139/soic-2310-5070-3445

1. Introduction

Markov chains have become essential in the modeling of stochastic systems where states change in successive time steps as a result of transition processes that are probabilistic in nature. The core of any serious study of discrete-time Markov chains is the transition probability matrix that formally describes the probability of transitioning from one state to another. Proper estimation of such a matrix is thus critical in explaining the dynamics of the system as well as in making reliable inference in an array of applied fields [1] and [2]. Such matrices are usually reconstructed in practice using empirical transition frequency tables. The most common paradigm in this respect is the Maximum Likelihood Estimator (MLE), which is obtained by normalization of the observed transition counts in each row of the frequency matrix. Although this estimator is simple and has a high level of statistical efficiency when transition data are abundant, it may develop instability due to sparsity or when many transition counts are small or zero [1] and [3].

In order to alleviate this instability, Bayesian procedures based on the Dirichlet prior have been suggested in order to incorporate prior information. The addition of prior information introduces a smoothing mechanism that reduces

*Correspondence to: Muthanna Subhi Sulaiman (Email: muthanna.sulaiman@uomosul.edu.iq). Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

the effects of sparsity and stabilizes the resulting probability estimates. Under this scheme, Trendelkamp-Schroer and Noé showed that Bayesian estimation may significantly enhance the stability of transition-matrix inference when data are scarce. However, such estimators are very sensitive to the selection of prior hyperparameters, which determines the degree of regularization applied to the estimated transition probabilities [4], [5] and [6]. In addition to Bayesian smoothing, regularization and shrinkage methods have been considered in order to increase the stability of estimators. These methods attempt to decrease estimator variance by combining empirical counts with more organized inferential objectives. As an example, the notions of shrinkage were explored in the context of hidden Markov models by Fiecas et al., and later literature directly applied penalized-likelihood concepts to Markov-chain models. Although these achievements were made, the explicit application of linear shrinkage estimators to the recovery of transition probability matrices based on observed counts remains relatively limited in the literature [7], [8], [9] and [3] .

Intelligent optimization methods have become more popular in recent years as versatile tools for addressing complex estimation problems. In particular, meta-heuristic algorithms have the appeal of being able to search large solution spaces as well as address multi-objective requirements. Among these, one of the most popular is the Artificial Bee Colony (ABC) algorithm, which was proposed by Karaboga due to its simplicity and a good exploration-exploitation balance. In spite of the successful application of the ABC algorithm to diverse optimization problems, relatively little attention has been paid to its direct application to the estimation of transition probability matrices in discrete-time Markov chains [10], [11], [12] and [13] .

In a nutshell, the literature identifies four major methodological paths that can be used to estimate transition matrices: classical likelihood-based methods, Bayesian smoothing methods, regularization-based methods, and intelligent optimization methods. However, comparisons between these streams on a systematic basis have been uncommon, particularly across different levels of transition-data density. This observation leads to the construction of a single comparative framework that compares the relative performance of classical, Bayesian, shrinkage-based, and optimization-driven estimators of Markov transition probability matrices.

1.1. Gap in research and objective of the study

Although much effort has been devoted to estimating transition probability matrices in discrete-time Markov chains, it is clear that most recent research focuses on classical or Bayesian estimators, while systematic comparisons between regularization-based estimators and intelligent optimization methods remain comparatively underinvestigated. Besides, the performance of these estimators also varies significantly with transition-data density, especially in cases where observed frequencies are sparse or uneven [9] and [3].

As a result, the current research investigates the estimation of transition probability matrices through four different methodologies, namely the Maximum Likelihood Estimator (MLE), the Bayesian–Dirichlet estimator, the shrinkage-based estimator, and an optimization-based estimator built on the Artificial Bee Colony (ABC) algorithm. These estimators are also empirically tested by running simulation experiments over a range of transition densities, and the measure used to assess estimation performance is defined as the mean squared error (MSE).

2. Markov model

In order to build a Markov model, an observations sequence is gathered first such that each observation represents the state of a system at one time. Based on these counts, the number of transitions between each state and all other states may be counted to form a frequency (transition counts) matrix denoted by the symbol F. Suppose the finite state space $s = \{1, 2, \ldots, K\}$, then K is the number of states in the Markov chain. Let f_{ij} the number of transitions that were observed between states i and j, then the frequency matrix is:

$$F = [f_{ij}]. \tag{1}$$

Given that the state space is discrete and finite, it is possible to express the probabilities of transitions from one state to another state in one transition in the form of a matrix called the Transition Matrix, and denoted by the symbol P, the element (i, j) of this matrix is represented by the probability of transitions from state i to state j

with a probability

$$p_{ij} = Pr(X_{t+1} = j | X_t = i). \tag{2}$$

The transition matrix is stochastic and has to comply with two requirements: it has to have nonnegative entries, (i.e., $0 \leq p_{ij} \leq 1$), and the sum of any row equals one [1] and [2].

3. Estimating the Transition Probability Matrix

After $F = [f_{ij}]$ has been constructed we estimate the transition probability matrix $P = [p_{ij}]$. Given that the state space is finite and discrete, the probability transition matrix can be directly estimated from observed transitions [1] and [2]. In this study, four estimation methods are considered: MLE, Bayesian–Dirichlet, global data-driven shrinkage, and ABC.

3.1. Maximum Likelihood Estimator (MLE)

This method estimates the transition probabilities using relative frequencies only. Accordingly, the probability of transition between state i and state j is estimated from observed transition count. The multinomial likelihood under the Markov assumption, the likelihood function of the transition matrix can be written as:

$$L(P) = \prod_{i=1}^{K} \prod_{j=1}^{K} p_{ij}^{f_{ij}}, \tag{3}$$

By taking the logarithm:

$$\ln L(P) = \sum_{i=1}^{K} \sum_{j=1}^{K} f_{ij} \ln p_{ij}, \tag{4}$$

In order to make sure that each row is a valid probability distribution imposing the constraint

$$\sum_{j=1}^{K} p_{ij} = 1, \quad p_{ij} \geq 0, \quad \forall i, \tag{5}$$

$$\hat{p}_{ij}^{(MLE)} = \frac{f_{ij}}{\sum_j f_{ij}} = \frac{f_{ij}}{n_i}, \tag{6}$$

where $\hat{p}_{ij}^{(MLE)}$ is maximum likelihood estimator of the transition probability from state i to state j. $n_i = \sum_{j=1}^{K} f_{ij}$: the total number of transitions out of state i.
All symbols used above follow the definitions introduced earlier [1], [14] and [2]. This estimator is consistent and asymptotically normal under standard regularity conditions, but it can become unstable under sparse or highly imbalanced transition counts, where some transitions are unobserved and the estimated matrix may be poorly conditioned [1] and [3].
It is for this reason that we have had to resort, in further sections, to some estimation procedure such as Bayesian Dirichlet estimation, empirical Bayes learning of prior parameters, shrinkage procedures and generic optimization methods all aimed at arriving at more stable (and less variable) estimates.

3.2. Bayesian Dirichlet Prior Estimation

The MLE may not be ideal, especially when the data are sparse or unbalanced. The Dirichlet distribution is the conjugate prior for each row of a transition matrix, which makes it convenient for application in discrete-time Markov models [4], [5] and [6].
Assume that each row of P follows a Dirichlet distribution. $P_i \sim Dir(\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})$, with density:

$$\pi(P_i) = \frac{\Gamma(\alpha_{i0})}{\prod_{j=1}^{K} \Gamma(\alpha_{ij})} \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1}, \tag{7}$$

$$\alpha_{i0} = \sum_{j=1}^{K} \alpha_{ij}. \tag{8}$$

$\alpha_{ij} > 0$ are the initial strength or pseudo counts of transitions.
Given the observed frequencies f_{ij}, the posterior is:

$$\alpha_{ij}^{post} = \alpha_{ij} + f_{ij}. \tag{9}$$

So the Bayesian estimator of the transition probabilities is:

$$\hat{p}_{ij}^{(Bayes)} = \frac{\alpha_{ij} + f_{ij}}{\sum_{j=1}^{K}(\alpha_{ij} + f_{ij})}. \tag{10}$$

In this study, the prior parameters are specified using an empirical Bayes scheme. First, the global column-sum distribution is defined as:

$$g_j = \left(\frac{\sum_{i=1}^{K} f_{ij}}{\sum_{i=1}^{K}\sum_{j=1}^{K} f_{ij}} \right), \tag{11}$$

where g_j denotes the normalized global column-sum distribution. Then, for each row i, a data-driven base distribution is constructed as

$$q_{ij} = w\hat{p}_{ij}^{(MLE)} + (1-w)g_j, \qquad j = 1, \ldots, K. \tag{12}$$

The choice $w = 0.5$ was used as a balanced implementation choice, giving equal importance to the local row-wise empirical structure and the global column-based transition profile, where g_j is the normalized global column-sum component. The Dirichlet prior is then defined by

$$\alpha_{ij} = q_{ij}\hat{c}, \tag{13}$$

$$\hat{c} = \arg \max_{c \in [10^{-3}, 5\times 10^{3}]} \sum_{i=1}^{K} \log p(F_i | cq_i), \tag{14}$$

where $\hat{c}$ is the estimated global concentration parameter representing the overall prior strength assigned to the base distribution q_{ij}, obtained by maximizing the total Dirichlet-multinomial marginal likelihood over all rows [15] and [5].
Hence, the resulting empirical Bayes estimator can be written as

$$\hat{p}_{ij}^{(Bayes)} = \frac{q_{ij}\hat{c} + f_{ij}}{n_i + \hat{c}}, \tag{15}$$

where $n_i = \sum_{j=1}^{K} f_{ij}$.
This construction allows the Bayesian estimator to adapt to both row-specific empirical information and the overall transition structure in the data, thereby improving stability relative to fixed-prior Dirichlet models [4] and [6].

3.3. Shrinkage Estimation

Using the notation introduced in the previous subsections, let $\hat{P}^{(MLE)}$ denote the empirical maximum likelihood estimator of the transition probability matrix, and let n_i denote the total number of observed transitions leaving state i. When the observed transition counts are sparse or uneven across states, the empirical transition matrix may become highly variable and unstable. To reduce this variability while preserving the main data-driven transition structure, a global data-driven shrinkage estimator is employed. This is consistent with the general principle of shrinkage estimation, in which an empirical estimator is regularized toward a structured target in order to improve stability under limited or noisy data [7] and [8]. In the context of Markov chains, recent work on regularized and penalized estimation also supports the need for structured estimation when the available transition information is limited or irregular [9] and [3].
In the implemented code, the local target is taken directly as the empirical MLE matrix itself. Thus,

$$T^{local} = \hat{P}^{(MLE)}. \tag{16}$$

Hence, the shrinkage construction does not use additive smoothing, pseudo-count correction, numerical floors, or uniform filling at this stage. This is important because the implemented estimator is intentionally based on the clean empirical matrix rather than on a pre-smoothed local target.
To represent the global transition tendency in the observed data, the column totals are computed as

$$G_j = \sum_{i=1}^{K} f_{ij}, \tag{17}$$

and then normalized to obtain the global transition profile

$$g_j = \frac{G_j}{\sum_{j=1}^{K} G_j}. \tag{18}$$

Thus,

$$g = (g_1, g_2, \ldots, g_K), \tag{19}$$

is a probability vector summarizing the overall destination tendency across all observed transitions. This vector is repeated across all rows to form the global target matrix $T^{(global)}$, where every row equals g:

$$T^{(global)} = \begin{pmatrix} g_1 & \cdots & g_K \\ \vdots & \ddots & \vdots \\ g_1 & \cdots & g_K \end{pmatrix}. \tag{20}$$

so that the final target combines row-specific empirical information and global destination structure. The use of a structured target to trade bias for variance reduction is fully aligned with the classical shrinkage rationale of [7] and [8].
The final target matrix is defined by an equal-weight mixture of the local and global targets:

$$T = wT^{local} + (1 - w)T^{(global)}, \tag{21}$$

where $w = 0.5$
Therefore, at the element level,

$$T_{ij} = 0.5\,\hat{p}_{ij}^{(MLE)} + 0.5\,g_j. \tag{22}$$

This elementwise expression is the direct implementation used in the code, since $T^{(local)} = \hat{P}^{(MLE)}$ and each row of $T^{(global)}$ is equal to g.
After constructing the target matrix, the shrinkage intensity is estimated directly from the data through a single global coefficient rather than by row-wise tuning or grid search. Specifically, the code uses

$$\lambda_{global} = \frac{\hat{A}}{\hat{A} + \hat{B}}, \tag{23}$$

where $\hat{A}$ is an estimated variability component and $\hat{B}$ is a discrepancy component measuring the distance between the empirical matrix and the target. The variability component is defined as

$$\hat{A} = \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{\hat{p}_{ij}^{(MLE)}(1 - \hat{p}_{ij}^{(MLE)})}{n_i}, \tag{24}$$

and the discrepancy component is defined as

$$\hat{B} = \sum_{i=1}^{K} \sum_{j=1}^{K} (\hat{p}_{ij}^{(MLE)} - T_{ij})^2. \tag{25}$$

To guarantee a valid shrinkage intensity, the implementation constrains the coefficient to the interval $[0, 1]$. Hence, the practical estimator uses the clamped form

$$\lambda_{global}^{c} = \min\left(1, \max\left(0, \frac{\hat{A}}{\hat{A} + \hat{B}}\right)\right). \tag{26}$$

There is a direct interpretation of this formulation. When the data are sparse, the totals of the rows in n_i are small, so $\hat{A}$ is larger and the resulting shrinkage strength is higher, creating a greater pulling force toward the target matrix.

In case of denser data, $\hat{A}$ is smaller and the degree of shrinking is less, and therefore the estimate can remain closer to the empirical transition matrix. Simultaneously $\hat{B}$ discourages over-shrinkage in case the target is far from the empirical estimate. This is in line with the overall bias-variance explanation of the shrinkage estimators in the statistical literature [7] and [8].
The final shrinkage estimator is therefore constructed as

$$\hat{p}_{ij}^{(SH)} = \left(1 - \lambda_{global}^{c}\right) \hat{p}_{ij}^{(MLE)} + \lambda_{global}^{c} T_{ij}. \tag{27}$$

or, in matrix form,

$$\hat{P}^{(SH)} = \left(1 - \lambda_{global}^{c}\right) \hat{P}^{(MLE)} + \lambda_{global}^{c} T, \tag{28}$$

Where $0 \leq \lambda_{global}^{c} \leq 1$.
Finally, the resulting matrix is row-normalized to preserve the row-stochastic property:

$$\hat{P}^{(SH)} = \text{RowNormalize}\left(\hat{P}^{(SH)}\right). \tag{29}$$

3.4. Artificial Bee Colony Optimization for Transition-Matrix Estimation

The current study uses an Artificial Bee Colony (ABC) optimization method to estimate a complete transition probability matrix with improved statistical and dynamical properties, in addition to the described empirical, Bayesian-Dirichlet, and shrinkage estimators. The ABC procedure does not rely on a single global coefficient to modify the empirical matrix; rather, admissible transition matrices are directly searched and evaluated using a composite objective function. This role is fully consistent with the original purpose of the ABC algorithm as a derivative-free optimizer for continuous numerical problems [10], [11] and [12]. ABC was adopted because the problem involves constrained matrix-valued optimization with a composite objective, which is not naturally reduced to a simple closed-form or standard convex optimization problem.
Let $P^{(data)}$ denote the empirical transition matrix obtained from the observed transition counts, and let n_i denote the row totals already defined in the previous subsections. In the present formulation, each food source is represented by a real-valued matrix $X \in \mathbb{R}^{K \times K}$. Since such a matrix is not necessarily stochastic, it is transformed into a valid transition matrix through a row-wise softmax map, followed by row normalization and a Frobenius-norm trust-region projection around an anchor matrix.
The row-wise softmax transformation is defined elementwise as

$$[\text{Softmax}_{row}(X)]_{ij} = \frac{\exp(x_{ij})}{\sum_{r=1}^{K} \exp(x_{ir})}. \tag{30}$$

Accordingly, the feasible candidate matrix is defined as

$$P = \Pi_{\delta}\left(\text{RowNormalize}\left(\text{Softmax}_{row}(X)\right)\right), \tag{31}$$

Thus, P denotes the feasible stochastic candidate matrix corresponding to the source matrix X.
Here, $\Pi_{\delta}(\cdot)$ denotes a Frobenius-norm trust-region projection around the anchor matrix P_{anchor}.

$$Q = \text{RowNormalize}\left(\text{Softmax}_{row}(X)\right), \tag{32}$$

Let Q denote the row-normalized softmax candidate before projection.
This construction ensures that every candidate matrix remains nonnegative and row-stochastic throughout the search process, while preserving the continuous-search spirit of the standard ABC framework [11] and [12].
The anchor matrix is taken as the empirical transition matrix itself,

$$P_{anchor} = P^{(data)}. \tag{33}$$

The trust-region radius is defined by

$$\delta = c \frac{\|P^{(data)} - T\|_F}{\sqrt{\text{mean}(n_i)}}, \tag{34}$$

where $c = 3$, and the final radius used in the implementation is truncated according to

$$\delta = \min(\delta, 0.6). \tag{35}$$

Then the projected candidate is computed as:

$$\Pi_\delta(Q) = \begin{cases} P_{anchor} + \dfrac{\delta}{\|Q - P_{anchor}\|_F}(Q - P_{anchor}), & \text{if } \|Q - P_{anchor}\|_F > \delta, \\ Q, & \text{otherwise.} \end{cases} \tag{36}$$

Otherwise, the candidate is kept unchanged. If projection is applied, row normalization is performed again to preserve row-stochasticity.

Thus, the search is centered around the empirical transition structure while still allowing sufficient flexibility for optimization. In practical terms, larger discrepancy between the empirical matrix and the target matrix permits broader exploration, whereas denser data reduce the search radius through the denominator involving mean n_i. This use of constrained neighborhood search is compatible with the exploratory–exploitative balance emphasized in the ABC literature [10] and [12].

The ABC estimator minimizes a composite objective function defined on the full candidate matrix P. The first component measures weighted discrepancy from the empirical transition matrix:

$$J_d(P) = \sum_{i=1}^{K}\sum_{j=1}^{K} n_i \left(p_{ij} - p_{ij}^{(data)}\right)^2. \tag{37}$$

The row totals are used as weights so that rows supported by larger observed transition counts contribute proportionally more to the discrepancy measure. The second component is the variance of the candidate matrix entries:

$$J_v(P) = \text{Var}\left(\text{vec}(P)\right). \tag{38}$$

The third component measures dynamical convergence speed through the mixing-time criterion:

$$J_s(P) = \text{MixingTime}\left(P, \varepsilon_{mix}\right), \tag{39}$$

With: $\varepsilon_{mix} = 10^{-3}$.

Before constructing the final fitness function, the three raw components J_d, J_v and J_s are evaluated on a preliminary set of $Samp = 40$ randomly generated feasible candidate matrices.

Here, $Samp = 40$ denotes the number of preliminary random feasible matrices used to estimate the sample minima and maxima required for normalization. In the implementation, the sample maxima used for normalization are slightly inflated by a numerical constant 10^{-12}. Thus, the normalization bounds are taken as

$$J_{d,max}^* = J_{d,max} + 10^{-12}, \quad J_{v,max}^* = J_{v,max} + 10^{-12} \quad \text{and} \quad J_{s,max}^* = J_{s,max} + 10^{-12},$$

$$J_{d,norm}(P) = \frac{J_d(P) - J_{d,min}}{J_{d,max}^* - J_{d,min} + 10^{-12}}. \tag{40}$$

$$J_{v,norm}(P) = \frac{J_v(P) - J_{v,min}}{J_{v,max}^* - J_{v,min} + 10^{-12}}. \tag{41}$$

$$J_{s,norm}(P) = \frac{J_s(P) - J_{s,min}}{J_{s,max}^* - J_{s,min} + 10^{-12}}. \tag{42}$$

In the implementation, if the mixing-time component becomes infinite, it is replaced by a large finite constant 10^9 before normalization.

These three terms are combined after normalization, so that the final fitness function is

$$\text{Fit}(P) = w_{data}J_{d,norm}(P) + w_{var}J_{v,norm}(P) + w_{spd}J_{s,norm}(P), \tag{43}$$

The ABC search is formulated as a minimization problem, and the final estimate is the feasible transition matrix with the smallest fitness value.

With equal weights $w_{data} = w_{var} = w_{spd} = \frac{1}{3}$.

Hence, the optimization simultaneously seeks empirical fidelity, reduced elementwise variability, and improved dynamical mixing behavior. The use of ABC for such multi-component continuous objectives is consistent with its well-established use in complex numerical optimization [11], [12] and [13].

The colony is initialized with $SN = 10$ food sources, where SN denotes the colony size. The trial counters are initialized to zero for all food sources. The algorithm is run for a maximum of 50 iterations. Each source is associated with a trial counter, and the abandonment threshold is fixed at $limit = 100$.
where $limit$ is the abandonment threshold used to identify non-improving food sources.
The neighborhood perturbation scale is $\tau = 0.5$.
At initialization, each source is generated as a real-valued random matrix from a standard normal distribution, then mapped to a feasible stochastic matrix using equation (31), and finally evaluated by the fitness function in equation (43). This produces the initial population of admissible transition matrices, in line with the standard population-based architecture of ABC [10] and [11].
During the employed-bee phase, each source X_i is updated by selecting another source X_k, where $k \neq i$, and generating an elementwise perturbation matrix according to

$$\phi = \tau(2\,\mathrm{rand} - 1), \tag{44}$$

where ϕ is an elementwise perturbation matrix having the same dimension as X_i.
The new candidate source is then formed as

$$X_i^{(new)} = X_i + \phi \odot (X_i - X_k), \tag{45}$$

where $\odot$ denotes elementwise multiplication. The resulting matrix is mapped again through equation (31), evaluated by the fitness function, and accepted if it improves the objective value. Otherwise, the previous source is retained and its trial counter is incremented. This neighborhood-update rule is the standard local search mechanism of the ABC algorithm in continuous domains [11].
During the onlooker-bee phase, food sources are selected probabilistically according to their relative quality. In the present formulation, the quality score of source i is defined as

$$qual_i = \frac{1}{1 + S_i - \min(S) + 10^{-12}}, \tag{46}$$

where S_i denotes the current fitness value of source i. The corresponding selection probability is

$$prob_i = \frac{qual_i}{\sum_{r=1}^{SN} qual_r}. \tag{47}$$

Onlooker bees then apply the same neighborhood update rule used in the employed-bee phase. This stage intensifies the search around promising regions and reflects the exploitation mechanism emphasized in the original and later ABC studies [11] and [12].

During the scout-bee phase, any source whose trial counter exceeds the abandonment threshold is discarded and replaced by a new random source. In the present formulation, this occurs whenever $trial(i) > limit$.
Here, $trial(i)$ denotes the number of consecutive unsuccessful updates associated with source i.
The replacement source is generated anew as a real-valued random matrix from a standard normal distribution and then mapped to a feasible stochastic matrix using equation (31). This step preserves exploration and reduces the risk of stagnation in poor local regions, which is one of the defining features of the ABC framework [11] and [12].
After completing the prescribed number of iterations, the source with the minimum fitness value is selected as the final estimate. If b denotes the index of the best food source, then the ABC estimator is

$$\hat{P}^{(ABC)} = P_b. \tag{48}$$

Based on this, the obtained estimation solution is a full-matrix, data-anchored, trust-region optimization solution. It is in stark contrast to closed-form regularization since, as a method, it does not optimize a single scalar parameter; instead, it is able to search through the space of admissible transition matrices directly and indeed measures each candidate by its empirical fit, variance control, and dynamic convergence behavior. This formulation is fully consistent with the established ABC literature and with more recent reviews of ABC variants in broader optimization settings [12] and [13].

3.4.1. Main steps of the proposed ABC estimator

Algorithm 1 Main steps of the proposed ABC estimator for transition-matrix estimation

Input: observed transition frequency matrix F, row totals n_i, tolerance ε_{mix}, colony size SN, maximum number of iterations $maxIter$, abandonment limit $limit$, perturbation scale τ, and trust-region radius δ.
Output: estimated transition probability matrix $\hat{P}^{(ABC)}$.

1: **Initialization**
Compute the empirical transition matrix $P^{(data)} = \hat{P}^{(MLE)}$, set the anchor matrix $P_{anchor} = P^{(data)}$, and generate an initial population of SN real-valued source matrices. Each source is mapped into a feasible candidate transition matrix by applying the row-wise softmax transformation, row normalization, and trust-region projection. Then, evaluate all feasible candidates using the composite fitness function.

2: **Employed-bee phase**
Update each source through the neighborhood rule, map the updated source into a feasible transition matrix, and accept it whenever the fitness value is improved.

3: **Onlooker-bee phase**
Compute the quality scores and selection probabilities of all sources, then update the selected sources using the same neighborhood rule and acceptance criterion.

4: **Scout-bee phase**
For any source whose trial counter exceeds $limit$, replace it with a newly generated random source and map it again into a feasible transition matrix.

5: **Termination and final estimation**
Repeat the employed-bee, onlooker-bee, and scout-bee phases until the maximum number of iterations $maxIter$ is reached. Finally, select the feasible candidate matrix with the minimum fitness value and return it as the final estimator $\hat{P}^{(ABC)} = P_b$.

4. Evaluation Metrics of Transition Matrix Estimation

The performance of the estimated transition probability matrices is evaluated using four criteria: mean squared error (MSE), element-wise variance, spectral gap, and mixing time. In the present study, MSE is used as the primary measure of estimation accuracy in the simulation study because the true transition probability matrix is known. The remaining criteria are used as complementary measures to assess the statistical stability and dynamic behaviour of the estimated Markov chains [16], [1] and [17].

4.1. Mean Squared Error (MSE)

In the simulation study, the direct statistical accuracy of an estimated transition matrix is assessed using the mean squared error (MSE). Let

$$P^{true} = (p_{ij}^{true})$$

denote the true transition probability matrix used in the data-generating process, and let

$$\hat{P} = (\hat{p}_{ij})$$

represent the estimated transition probability matrix obtained from a certain estimation method.
Then, the MSE is defined as:

$$MSE(\hat{P}) = \frac{1}{K^2} \sum_{i=1}^{K} \sum_{j=1}^{K} (\hat{p}_{ij} - p_{ij}^{true})^2. \tag{49}$$

This criterion measures the overall elementwise deviation between the estimated matrix and the true matrix. Smaller values of MSE indicate more accurate recovery of the underlying transition probabilities. In the present study, MSE is the main criterion used to compare the competing estimators under controlled simulation settings [16].

4.2. Element-wise Variance

In order to gauge the overall dispersion of the estimated transition probabilities, the element-wise variance of the estimated transition matrix is calculated as:

$$\text{VarElem}(\hat{P}) = \frac{1}{K^2 - 1} \sum_{i=1}^{K} \sum_{j=1}^{K} (\hat{p}_{ij} - \bar{p})^2, \tag{50}$$

where

$$\bar{p} = \frac{1}{K^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \hat{p}_{ij}. \tag{51}$$

is the mean of all entries of the estimated transition matrix. This measure reflects the general spread of the estimated transition probabilities across all matrix entries. Smaller values indicate lower variability and greater stability of the estimator, particularly when the observed transition counts are sparse or irregular. This variance-based interpretation is consistent with the broader literature on shrinkage and regularized estimation [7] and [8].

4.3. Spectral Gap

In order to assess the dynamic behaviour of the estimated transition matrix, the spectral gap is determined by the second largest absolute eigenvalue of . Let be the second largest absolute eigenvalue of the estimated transition matrix. Then the spectral gap is given by:

$$\text{Gap}(\hat{P}) = \begin{cases} 1 - |\lambda_2|, & 0 < |\lambda_2| < 1 \\ 0, & |\lambda_2| \geq 1 \end{cases}. \tag{52}$$

The spectral gap is a measure of the forgetting rate of the Markov chain on its starting distribution. The larger the values used, the quicker the convergence to equilibrium and vice versa. In the present implementation, when the second eigenvalue is greater than or equal to one in absolute value, the chain is considered to be non-mixing and the gap is defined as zero. The spectral gap is a diagnostic of convergence related to Markov-chain theory and mixing-time analysis [1] and [17].

4.4. Mixing Time

The mixing time is a complementary measure of the convergence rate, which is estimated with the aid of the second largest absolute eigenvalue using the tolerance level

$$\varepsilon_{mix} = 10^{-3}.$$

The mixing time is defined as:

$$t_{mix}(\varepsilon_{mix}) = \begin{cases} \left\lceil \dfrac{\log(\varepsilon_{mix})}{\log(|\lambda_2|)} \right\rceil, & 0 < |\lambda_2| < 1 \\ \infty, & |\lambda_2| \geq 1. \end{cases} \tag{53}$$

The second eigenvalue is used to make this approximation because of the geometric decay. Smaller values of mixing time mean that the chain converges much faster and infinite values mean that the estimated matrix does not meet the needed mixing condition under this spectral criterion. The relationship between the spectral gap and the second eigenvalue and mixing behaviour is well known in the literature on Markov chains [17] and [18].
In summary, MSE is the primary measurement of estimation error in the simulation experiment, but the element-wise variance, spectral gap, and mixing time are complementary measures regarding the stability and dynamic behaviour of the estimated transition matrices.

5. Simulation Study

5.1. Design of the simulation experiment

To evaluate the performance of the competing estimators under controlled conditions, a Monte Carlo simulation study was conducted. The number of states was fixed at $K = 5$, and 100 transition frequency matrices were generated for each transition-density scenario. In each simulation replication, an independent true transition matrix P^{true} was first generated as a strictly positive row-stochastic matrix. More specifically, a random 5×5 matrix was generated, a small positive constant was added to all entries, and each row was then normalized so that its sum equals one. This construction ensures that the true transition matrix is probabilistically valid and free from structural zeros.
After generating P^{true}, a transition frequency matrix $F = (f_{ij})$ was sampled row by row from a multinomial distribution, with the row totals varying according to the transition-density setting. Three density levels were considered:
Sparse: $n_i \in [2, 8]$
Moderate: $n_i \in [20, 50]$
Dense: $n_i \in [100, 500]$
Rows with zero total counts were excluded by construction, so that all generated frequency matrices were valid for subsequent estimation. For each generated matrix F, four estimators were computed:

- the Maximum Likelihood Estimator (MLE),
- the empirical Bayes Dirichlet estimator,
- the global data-driven shrinkage estimator,
- the proposed Artificial Bee Colony (ABC) estimator.

To ensure reproducibility, the random generation of the simulated data was controlled independently from the stochastic search stage of the ABC algorithm.
The estimators were evaluated using four criteria. The mean squared error (MSE) was used as the primary measure of estimation accuracy, since the true transition matrix was known in the simulation study. In addition, element-wise variance, spectral gap, and mixing time were recorded as complementary indicators of statistical stability and dynamic behaviour.

5.2. Simulation results

The average results over the 100 replications for each density scenario are summarized in Tables 1-3.

Table 1. Performance measures under dense transition counts

Method	Mean VarElem	Mean Gap	Mean MixingTime	Mean MSE
MLE	0.012793	0.745649	5.640000	0.000653
Dirichlet	0.012749	0.746214	5.630000	0.000650
Shrinkage	0.010974	0.771426	5.220000	0.000656
ABC	0.010955	0.766090	5.220000	0.000627

5.3. Discussion of simulation results

The simulation results indicate that the proposed ABC estimator achieved the smallest MSE across the three transition-density scenarios. This suggests that, under the current simulation design, it was the most effective estimator for recovering the true transition matrix.
In the dense case, all four estimators performed well and the differences in MSE were small. Even so, the ABC estimator again achieved the highest accuracy (0.000627), followed closely by the Dirichlet estimator (0.000650),

Table 2. Performance measures under moderate transition counts

Method	Mean VarElem	Mean Gap	Mean MixingTime	Mean MSE
MLE	0.016478	0.723620	5.970000	0.004549
Dirichlet	0.016091	0.727818	5.880000	0.004440
Shrinkage	0.010297	0.803983	4.800000	0.003601
ABC	0.010052	0.786624	4.950000	0.003415

Table 3. Performance measures under sparse transition counts

Method	Mean VarElem	Mean Gap	Mean MixingTime	Mean MSE
MLE	0.052791	0.480902	12.979798	0.038616
Dirichlet	0.043713	0.539295	10.510000	0.031688
Shrinkage	0.028912	0.661226	7.390000	0.021868
ABC	0.013058	0.752504	5.430000	0.012487

MLE (0.000653), and Shrinkage (0.000656). This suggests that when transition counts are sufficiently large, the competing estimators become closer in performance. Still, the proposed ABC procedure keeps a slight advantage. In the moderate case, the advantage of the ABC estimator became clearer. It produced the best MSE (0.003415), followed by Shrinkage (0.003601), while the Dirichlet and MLE estimators were noticeably less accurate. It should also be noted that, in this setting, Shrinkage achieved a slightly larger spectral gap and a slightly smaller mixing time than ABC, which indicates stronger dynamic regularization. However, its MSE was still higher than that of ABC. This is important because it shows that improved dynamic behaviour does not necessarily lead to better statistical recovery of the true transition matrix. Under the evaluation criterion adopted in this study, MSE remains the main criterion. For that reason, ABC remains the preferred estimator in the moderate-density case.

The largest differences appeared in the sparse scenario. Here, ABC clearly outperformed all competing approaches in MSE as well as in the complementary measures of stability and dynamics. It achieved a mean MSE of (0.012487) compared with Shrinkage (0.021868), Dirichlet (0.031688), and MLE (0.038616). In addition, ABC recorded the smallest element-wise variance, the largest spectral gap, and the smallest mixing time. These findings show that the proposed ABC estimator works especially well when transition data are sparse or irregular. This is exactly the setting in which classical frequency-based estimation is most vulnerable to instability.

Another pattern in the results is that the Dirichlet estimator performed better than MLE at all density levels. This supports the idea that empirical Bayes smoothing provides useful stabilization when the observed data are limited. Still, the gains achieved by Dirichlet remained smaller than those achieved by ABC. The shrinkage estimator also offered a strong balance between stability and accuracy, especially in the moderate and sparse cases, but it did not outperform ABC in MSE under any of the three density settings.

Overall, the simulation study provides consistent evidence that the proposed ABC-based estimator combines high estimation accuracy with favorable stability and convergence behavior. Its advantage is small under dense data, becomes clearer under moderate data, and grows larger under sparse data. This pattern is methodologically meaningful because sparse and irregular transition structures are exactly the situations in which transition-matrix estimation becomes most difficult in practice.

5.4. Main conclusion from the simulation study

From the standpoint of estimation accuracy, the simulation results support the following ordering:

- ABC: best overall performance across all density settings,
- Shrinkage: strongest competitor to ABC, especially in moderate and sparse settings,
- Dirichlet: consistently better than MLE, but weaker than ABC and Shrinkage,
- MLE: most sensitive to sparsity and generally the least accurate estimator.

Therefore, the simulation evidence supports the use of the proposed ABC estimator as the most reliable method among the competing procedures, particularly when the transition data are sparse or moderately dense, while remaining competitive even under dense transition structures.

6. Real Data Application

The raw data used in this study consist of annual adult literacy rates (percentage values) reported by the UNESCO Institute for Statistics (UIS) for the period of 1970 – 2024. The data were obtained from the (UIS) Data Browser, under SDG 4 Education: Global and Thematic Indicators (last update: September 2025). The series was extracted for the Arab Mashreq countries and used for the period 1970 – 2024. The data collected represent the adult literacy rate (%) per year t, where $t = \{1970, 1971, \ldots, 2024\}$ [19].

6.1. State Classification

The data were classified into five main states (transitions occur between them) as shown in the following table:

Table 4. State Description

State	Description	Range
S_1	Low	Less than 55
S_2	Medium	55 – 64
S_3	Good	65 – 74
S_4	Very Good	75 – 84
S_5	Excellent	More than 84

When this classification was applied to the 55-year series, the result was a time series discretized S into the state space. The time index is $t = 1, 2, \ldots, n$, where n is the number of observed time points (years).

6.2. Construction of the Frequency (Transition Counts) Matrix

The empirical frequency matrix was constructed by counting the number of transitions from state i to state j. This gives the following (5×5) matrix:

$$F = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 8 & 2 & 0 & 0 \\ 0 & 1 & 6 & 1 & 0 \\ 0 & 0 & 0 & 32 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

For instance, it means that system changed from S_2 to S_2 eight times and from S_4 to S_4 thirty-two times.

6.3. Estimation of the Transition Matrix

The four estimation methods developed in this study were applied to the observed annual literacy transition frequency matrix. The resulting estimated transition matrices are presented below.

1. **Maximum Likelihood Estimation (MLE)**

$$P^{(MLE)} = \begin{bmatrix} 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0909 & 0.7273 & 0.1818 & 0.0000 & 0.0000 \\ 0.0000 & 0.1250 & 0.7500 & 0.1250 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.9697 & 0.0303 \\ 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 \end{bmatrix}$$

2. **Bayesian-Dirichlet Estimation**

$$P^{(Dir)} = \begin{bmatrix} 0.0046 & 0.7963 & 0.0370 & 0.1574 & 0.0046 \\ 0.0879 & 0.7047 & 0.1804 & 0.0262 & 0.0008 \\ 0.0010 & 0.1283 & 0.7166 & 0.1530 & 0.0010 \\ 0.0003 & 0.0027 & 0.0022 & 0.9647 & 0.0301 \\ 0.0046 & 0.0463 & 0.0370 & 0.9074 & 0.0046 \end{bmatrix}$$

3. **Shrinkage Estimation**

$$P^{(SH)} = \begin{bmatrix} 0.0011 & 0.9524 & 0.0087 & 0.0368 & 0.0011 \\ 0.0867 & 0.6956 & 0.1798 & 0.0368 & 0.0011 \\ 0.0011 & 0.1285 & 0.7148 & 0.1545 & 0.0011 \\ 0.0011 & 0.0108 & 0.0087 & 0.9498 & 0.0296 \\ 0.0011 & 0.0108 & 0.0087 & 0.9783 & 0.0011 \end{bmatrix}$$

4. **Artificial Bee Colony (ABC) Estimation**

$$P^{(ABC)} = \begin{bmatrix} 0.0630 & 0.7272 & 0.0659 & 0.0668 & 0.0772 \\ 0.1246 & 0.5444 & 0.1934 & 0.0707 & 0.0669 \\ 0.0725 & 0.1502 & 0.5623 & 0.1484 & 0.0666 \\ 0.0652 & 0.0662 & 0.0665 & 0.7113 & 0.0908 \\ 0.0670 & 0.0689 & 0.0717 & 0.7272 & 0.0652 \end{bmatrix}$$

6.4. Evaluation Metrics for the Estimated Transition Matrices

Table 5. Comparison of the estimated transition matrices for the real-data application

Method	Variance	Spectral Gap	Mixing Time
MLE	0.12938	0.0579	116
Dirichlet	0.10457	0.1008	66
Shrinkage	0.11856	0.113	58
ABC	**0.05630**	**0.37209**	**15**

7. Discussion of Results

When the four estimation methods were applied to the real transition frequency matrix, the mean squared error (MSE) was no longer available, since the true transition probability matrix is unknown in a real-data setting. Therefore, unlike the simulation study, where MSE was used as the primary criterion for estimation accuracy, the real-data analysis relied on structural and dynamical measures, namely element-wise variance, spectral gap, and mixing time. This connection is significant in the sense that the simulation findings already indicated that the Artificial Bee Colony (ABC) estimator performed optimally in terms of MSE under all transition-density conditions especially under sparse and moderate settings, thus strongly demonstrating its capability to recover the true transition matrix when the ground truth is known.

The Maximum Likelihood Estimator (MLE) in the real-data application resulted in a relatively sharp transition matrix, some entries of which were almost deterministic. This was reflected in weaker dynamical properties, since it achieved the highest element-wise variance (0.12938), the smallest spectral gap (0.057912), and the longest mixing time (116). By contrast the Dirichlet estimator as well as the shrinkage estimator, improved the transition structure by reducing the sharpness of certain unstable entries, which decreased the element-wise variance and improved the spectral gap and mixing time relative to MLE.

However, the estimator based on ABC once again excelled in overall performance in the real-data application with respect to the considered criteria. It had the smallest element-wise variance (0.056308), the largest spectral gap

(0.37209), and the shortest mixing time (15) when compared with all competing methods. These results suggest that the simulation study did not only demonstrate the superiority of ABC in an artificial setting, but that the real-data application also confirmed the results reported in the simulation study. Although the simulation study had established its superiority in terms of direct estimation accuracy through MSE, the real-data results indicated that it also yield a transition matrix exhibiting more regular balanced, and dynamically favorable characteristics. The combination of the simulation and real-data outcomes provides strong methodological support for the effectiveness of the proposed ABC estimator.

8. Conclusion

This paper examined the estimation of transition probability matrices in discrete-time Markov chains by four competing methods: the Maximum Likelihood Estimator (MLE), the Bayesian-Dirichlet estimator, a global data-driven shrinkage estimator, and an estimation method based on the Artificial Bee Colony (ABC) algorithm. A Monte Carlo simulation study was conducted under three transition-density conditions, i.e., dense, moderate, and sparse, so as to compare the estimators under controlled conditions. The mean squared error (MSE) was taken as the main measure of estimation accuracy, and element-wise variance, spectral gap, and mixing time were discussed as auxiliary measures of structural stability and dynamic behaviour in the simulation study where the true transition matrix was known.

Results of the simulation indicated that the proposed ABC estimator performed best in general in terms of MSE under all density settings. Its advantage was relatively small in the dense scenario, became clearer in the moderate scenario, and was most distinct in the sparse scenario. These results suggest that the proposed procedure using ABC is particularly efficient when the available transition information is limited or irregular, and it is also competitive in a dense transition structure. The shrinkage estimator developed in this paper was found to be the strongest competitor to ABC, especially in the moderate and sparse conditions, whereas the Bayesian-Dirichlet estimator consistently outperformed the MLE yet was not as accurate as either ABC or shrinkage. The MLE was also found to be the most sensitive to sparsity and usually the poorest estimator with respect to estimation error. In the real-data example, the actual transition matrix was unknown, hence the unavailability of MSE. This is why the comparison was based on structural and dynamical measures, i.e., element-wise variance, spectral gap, and mixing time. The findings were that the ABC estimator once again produced the most regular and dynamically favourable transition matrix in terms of the discussed criteria, with the smallest variance, the largest spectral gap, and the shortest mixing time among all competing estimators. The fact that the simulation study is consistent with the actual data application provides further evidence in favor of the proposed ABC estimator. The major aspect of the current research is that the final ABC configuration was not chosen arbitrarily. To test the sensitivity of the algorithm to its major settings, which include the trust-region scaling constant, colony size, and maximum number of iterations, several preliminary numerical experiments were conducted. According to these experiments, the chosen setup gave the best overall performance under the current simulation design. Besides, the shrinkage estimator of the present study was not merely a benchmarking method, but was created here as a proposed global data-driven regularization scheme based on a mixed structural target and an analytically estimated shrinkage intensity. In general, the findings of this work indicate that the given ABC framework is a useful and adaptable method for estimating transition matrices, especially under sparse and irregular data conditions when the use of classical frequency-based estimation can be unstable. Simultaneously, the proposed formulation of shrinkage provides a significant regularized option and forms a significant contribution of the research in conjunction with the ABC procedure. The study under consideration has its limitations. To start with, the simulation model was confined to a five-state environment and three transition-density conditions. Second, the empirical application to real data was premised on one empirical transition frequency matrix, and thus more comprehensive empirical validation would be welcome. Third, despite the fact that the final ABC settings were justified by the initial numerical experiments, their applicability may still depend upon the arrangement of the data and the shape of the objective function. Further studies can build on this study by exploring higher-dimensional state spaces, alternative transition models, more real-world applications, and hybrid optimization methods that combine ABC with other regularization methods or metaheuristic approaches.

Adaptive versions of the objective weights, trust-region control, and shrinkage targets, as well as larger comparative studies of other intelligent optimization methods for Markov transition-matrix estimation, could also be of interest.

REFERENCES

1. J. R. Norris, *Markov chains*, no. 2, Cambridge university press, 1998.
2. S. M. Ross, *Introduction to probability models*, Academic press, 2014.
3. Y. Zhou, M. Gao, Y. Chen, and X. Shi, *Adaptive penalized likelihood method for markov chains*, arXiv Preprint arXiv:2406.00322, 2024.
4. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed., Boca Raton, FL, USA: CRC press, 2013.
5. T. Minka, *Estimating a Dirichlet distribution*, Technical report, MIT, 2000.
6. B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé, *Estimation and uncertainty of reversible Markov models*, J. Chem. Phys., vol. 143, no. 17, 2015.
7. O. Ledoit and M. Wolf, *A well-conditioned estimator for large-dimensional covariance matrices*, J. Multivar. Anal., vol. 88, no. 2, pp. 365–411, 2004.
8. J. Schäfer and K. Strimmer, *A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics*, Stat. Appl. Genet. Mol. Biol., vol. 4, no. 1, 2005.
9. M. Fiecas, J. Franke, R. von Sachs, and J. Tadjuidje Kamgaing, *Shrinkage estimation for multivariate hidden Markov models*, J. Am. Stat. Assoc., vol. 112, no. 517, pp. 424–435, 2017.
10. D. Karaboga, *An idea based on honey bee swarm for numerical optimization*, Erciyes university, Engineering Faculty, Computer Engineering Department, 2005.
11. D. Karaboga and B. Basturk, *A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm*, J. Glob. Optim., vol. 39, no. 3, pp. 459–471, 2007.
12. D. Karaboga, B. Gorkemli, C. Ozturk, and N. Karaboga, *A comprehensive survey: artificial bee colony (ABC) algorithm and applications*, Artif. Intell. Rev., vol. 42, no. 1, pp. 21–57, 2014.
13. B. Akay, D. Karaboga, B. Gorkemli, and E. Kaya, *A survey on the artificial bee colony algorithm variants for binary, integer and mixed integer programming problems*, Appl. Soft Comput., vol. 106, p. 107351, 2021.
14. V. G. Kulkarni, *Modeling and analysis of stochastic systems*, Chapman and Hall/CRC, 2016.
15. C. N. Morris, *Parametric empirical Bayes inference: theory and applications*, J. Am. Stat. Assoc., vol. 78, no. 381, pp. 47–55, 1983.
16. G. Casella and R. L. Berger, *Statistical inference*, 2nd ed., Pacific Grove, CA, USA: Duxbury, 2002.
17. D. A. Levin and Y. Peres, *Markov chains and mixing times*, vol. 107, American Mathematical Soc., 2017.
18. R. R. Montenegro and P. Tetali, *Mathematical aspects of mixing times in Markov chains*, Now Publishers Inc, 2006.
19. UNESCO Institute for Statistics, *SDG 4 – Education Global and Thematic Indicators Data Browser (Bulk Download Resources)*, UNESCO Institute for Statistics. [Online]. Available: https://databrowser.uis.unesco.org/resources/bulk