Advances in Data Analysis and Classification
https://doi.org/10.1007/s11634-025-00631-y

REGULAR ARTICLE

Statistical jump model for mixed-type data with missing data imputation

Federico P. Cortese[1] · **Antonio Pievatolo**[1]

Received: 23 July 2024 / Revised: 15 January 2025 / Accepted: 1 March 2025

Abstract
In this paper, we address the challenge of clustering mixed-type data with temporal evolution by introducing the statistical jump model for mixed-type data. This novel framework incorporates regime persistence, enhancing interpretability and reducing the frequency of state switches, and efficiently handles missing data. The model is easily interpretable through its state-conditional medians and modes, making it accessible to practitioners and policymakers. We validate our approach through extensive simulation studies and an empirical application to air quality data, demonstrating its superiority in inferring persistent air quality regimes compared to the traditional air quality index. Our contributions include a robust method for mixed-type temporal clustering, effective missing data management, and practical insights for environmental monitoring.

Keywords Environmental monitoring · Mixed-type data · Missing data · Regime-switching models · Unsupervised learning

Mathematics Subject Classification 37M10 · 62D10 · 62H30

1 Introduction

Clustering mixed-type data is particularly challenging when accounting for its temporal evolution (Aghabozorgi et al. 2015). Existing literature on temporal clustering techniques addressed various aspects of this problem, with some exploring the inclusion of temporal persistence (Nystrup et al. 2020). In fact, persistence is crucial for interpretability and practical applications, as frequent state switches can complicate

✉ Federico P. Cortese
federico.cortese@mi.imati.cnr.it

[1] Institute for Applied Mathematics and Information Technologies "E. Magenes", National Research Council of Italy, via Alfonso Corti, 12, 20133 Milan, Italy

Published online: 25 March 2025

decision-making processes (Nystrup et al. 2019). Moreover, many real-world applications, especially in environmental and ecological studies (Kasam et al. 2014), face the issue of missing data, necessitating robust methods to handle such gaps effectively (Little and Rubin, 2019).

Regime-switching models (Hamilton, 1989), of which hidden Markov models are an example, (Bartolucci et al. 2013; Zucchini et al. 2017; Russo and Farcomeni, 2024) offer a valuable framework for analyzing time series data that exhibit different states or regimes over time. These models have been effectively applied in various domains, including financial markets (Cortese et al. 2023b; Bosancic et al. 2024; Shu et al. 2024b) and environmental studies (Kehagias, 2004; Cortese and Pievatolo, 2024). The integration of mixed-type data and the handling of missing values in these models still require further development.

The main purpose of this study is to provide a unified framework for clustering mixed-type temporal data, which accounts for regime persistence and handles missing values efficiently. Our proposed method, which we call *statistical jump model for mixed-type data* (JM-mix), enhances interpretability and practical applicability by reducing the frequency of state switches and providing a robust solution for missing data. Our model is inspired by the statistical jump model of Bemporad et al. (2018), which Nystrup et al. (2021) extended to address data sparsity, and Aydınhan et al. (2024) further adapted for soft clustering.

Our approach is motivated by the need for an advanced yet interpretable clustering method that outperforms existing models such as spectral clustering (Von Luxburg, 2007) and k-prototypes (Huang, 1998). Recent works on clustering mixed-type data with missing values explore various methodologies but do not consider time series data or their persistence. Szepannek et al. (2024) extend the k-prototypes algorithm by incorporating Gower's distance to handle ordinal data effectively, while Aschenbruck et al. (2023) investigate imputation strategies, comparing internal, external, and multiple imputation approaches for mixed-type data. Dinh et al. (2021) propose a method combining k-means with hierarchical clustering for mixed-type data, emphasizing scalability. Unlike more sophisticated methods like artificial neural networks (Liao and Wen, 2007), our model is straightforward to understand and interpret, making it accessible to practitioners and policymakers.

We make three main contributions to the existing literature. First, we introduce a new method for clustering mixed-type data over time, incorporating persistence in regimes to enhance interpretability and utility. Second, our approach includes a fast and efficient solution for managing missing data, ensuring robust clustering results. Third, we demonstrate the applicability of our method through an empirical study on air quality data. Our results show that our method can infer more persistent air quality regimes compared to the traditional air quality index (AQI, Rule 2005).

The rest of the paper is organized as follows. Section 2 introduces the model formulation, estimation procedures, and provides guidance on hyperparameter selection. Section 3 presents an extensive simulation study, demonstrating that our proposal outperforms competitive models in terms of clustering accuracy. Section 4 showcases an application to air quality data, where the level of air quality is treated as a latent process inferred from observable variables such as weather conditions and pollutant

levels. Section 5 concludes the paper with a discussion of the findings and implications for future research.

2 Methodology

Let us consider data $\boldsymbol{y}$, with T rows (times) and P columns (features). We denote the vector of observations at time $t = 1, \ldots, T$, as $\boldsymbol{y}_{t,\cdot} = (y_{t,1}, \ldots, y_{t,P})' \in \mathbb{R}^P$, and the vector of observations for feature p, $p = 1, \ldots, P$, as $\boldsymbol{y}_{\cdot,p} = (y_{1,p}, \ldots, y_{T,p}) \in \mathbb{R}^T$. Among these P features, some are continuous, while others are categorical.[1]

We further assume that some observations are missing. We denote a missing observation for feature p at time t as $y_{t,p}^{(miss)}$. When we need to specify an observed value at time t, or the vector of observed values for feature p, we denote them as $y_{t,p}^{(obs)}$ and $\boldsymbol{y}_{\cdot,p}^{(obs)}$, respectively. For simplicity, we may sometimes omit these superscripts.

A statistical jump model (Bemporad et al. 2018) transitions among a finite set of states or submodels to appropriately represent a sequence of data, while considering the order of the data points. It takes as input data $\boldsymbol{y}$, and produces two main outputs:

- a sequence of latent states $\boldsymbol{s} = (s_1, \ldots, s_T)'$ where each s_t takes a value from the discrete set $\{1, 2, \ldots, K\}$;
- a series of P-dimensional vectors of state-conditional prototypes $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$. Specifically, the prototypes are medians for continuous variables and modes for categorical variables (Huang, 1998).[2]

Similar to Nystrup et al. (2020), we propose fitting a JM-mix with K states by minimizing the following objective function

$$\sum_{t=1}^{T-1} \left[g(\boldsymbol{y}_{t,\cdot}, \boldsymbol{\mu}_{s_t}) + \lambda \mathbb{I}(s_t \neq s_{t+1}) \right] + g(\boldsymbol{y}_{T,\cdot}, \boldsymbol{\mu}_{s_T}), \tag{1}$$

over the latent state sequence $\boldsymbol{s} = (s_1, \ldots, s_T)'$ and the model parameters $\boldsymbol{\mu}_k = (\mu_{k,1}, \ldots, \mu_{k,P})'$. $\lambda \geq 0$ is a temporal jump penalty as in Nystrup et al. (2020) and Nystrup et al. (2021): it balances data fitting with state sequence stability, where a higher λ value makes the model more likely to stay in the same state, thus reducing the frequency of state changes. Finally, $g(\cdot, \cdot)$ is the Gower distance (Gower, 1971; Tuerhong and Kim, 2014), which is a metric used to measure dissimilarity between mixed-type variables. Given two vectors $\boldsymbol{x}_{i,\cdot}, \boldsymbol{x}_{j,\cdot} \in \mathbb{R}^P$, it is defined[3] as:

[1] This framework can be extended to ordinal variables, but we focus on continuous and categorical variables relevant to our application. For a similar approach including ordinal variables, see Szepannek et al. (2024).

[2] For ordinal variables, the state-specific median can be used as the prototype, as proposed in Szepannek et al. (2024).

[3] We refer the reader to Podani (1999) for a detailed definition of the Gower distance in the presence of ordinal variables.

$$g(\boldsymbol{x}_{i,\cdot}, \boldsymbol{x}_{j,\cdot}) = \frac{\sum_{p=1}^{P} w_{ijp} d_{ijp}}{\sum_{p=1}^{P} w_{ijp}} \tag{2}$$

where:

- d_{ijp} is the contribution of feature $p = 1, \ldots, P$ to the distance between $\boldsymbol{x}_{i,\cdot}$ and $\boldsymbol{x}_{j,\cdot}$.
- w_{ijp} is a weight assigned to feature p for the pair $(\boldsymbol{x}_{i\cdot}, \boldsymbol{x}_{j\cdot})$. We set all these weights to 1.

The specific contribution d_{ijp} depends on the type of feature p.

- For continuous features[4]:

$$d_{ijp} = \frac{|x_{ip} - x_{jp}|}{\max(\boldsymbol{x}_{\cdot,p}) - \min(\boldsymbol{x}_{\cdot,p})}$$

  where $\max(\boldsymbol{x}_{\cdot,p})$ and $\min(\boldsymbol{x}_{\cdot,p})$ are the maximum and minimum values of feature p across all data points $\boldsymbol{x}_{\cdot,p} = (x_{1,p}, \ldots, x_{T,p}) \in \mathbb{R}^T$.
- For categorical features:

$$d_{ijp} = \begin{cases} 0 \text{ if } x_{ip} = x_{jp} \\ 1 \text{ if } x_{ip} \neq x_{jp} \end{cases}$$

2.1 Model estimation

Following Nystrup et al. (2021), jump models can be efficiently fitted using a coordinate descent algorithm, which iteratively alternates between optimizing model parameters for a fixed state sequence, and updating the state sequence, based on these parameters, to minimize the objective function (1) (Bemporad et al. 2018). This process is repeated until the state sequence stabilizes. Specifically, we terminate either after a maximum of 10 iterations, based on empirical trials showing consistent convergence within this limit, or when the objective value changes by less than 10^{-6} between iterations. Achieving the global optimum is not guaranteed due to the dependency on the initial state sequence: to enhance solution quality, we follow Nystrup et al. (2020) and execute the algorithm multiple times with 10 initial state sequences in parallel, selecting the model with the lowest objective value (Arthur et al. 2007). The most likely sequence of states is found using dynamic programming: in particular, we use a reversed Viterbi algorithm (Viterbi, 1967).

We handle missing data by imputing the missing values at each iteration of the estimation process with the corresponding estimates of state-conditional medians or modes, depending on the type of data. This approach extends the k-pods method

[4] The use of Gower distance, incorporating the l_1-norm for continuous features, represents a methodological shift from the l_2-norm used in the original jump model (Nystrup et al. 2020). An additional simulation study, detailed in the supplementary material, demonstrates that our choice results in more robust and accurate clustering performance.

from Chi et al. (2016) to accommodate mixed-type data, whereas the original method focuses solely on continuous data. It is effective even under unknown missingness mechanisms, in the absence of external information, and when the data exhibits substantial missingness[5].

The estimation process can be summarized as follows.

(a) Initialize state sequence $s_1, \ldots, s_T$.
(b) Initialize missing values with median or mode of each feature, i.e. $y_{t,p}^{(miss)} = \text{Med}\left(\boldsymbol{y}_{\cdot,p}^{(obs)}\right)$ if $\boldsymbol{y}_{\cdot,p}$ is continuous and $y_{t,p}^{(miss)} = \text{Mode}\left(\boldsymbol{y}_{\cdot,p}^{(obs)}\right)$ if $\boldsymbol{y}_{\cdot,p}$ is categorical.
(c) Iterate the following steps for $j \in 1, \ldots,$ until $\boldsymbol{s}^j = \boldsymbol{s}^{j-1}$:
   (i) Fit model parameters $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$

$$\boldsymbol{\mu}^j = \underset{\boldsymbol{\mu}}{\text{argmin}} \sum_{t=1}^{T} g(\boldsymbol{y}_{t,\cdot}, \boldsymbol{\mu}_{s_t^{j-1}}).$$

   (ii) Update missing values

$$y_{t,p}^{(miss)} = \mu_{s_t,p}^{j}, \quad t = 1, \ldots, T, \quad p = 1, \ldots, P,$$

   where $\mu_{s_t,p}^{j}$ is the median (or mode) of feature p in cluster s_t. For legibility, the superscript $j-1$ on s_t is omitted.
   (iii) Fit state sequence

$$\boldsymbol{s}^j = \underset{\boldsymbol{s}}{\text{argmin}} \left\{ \sum_{t=1}^{T-1} \left[g(\boldsymbol{y}_{t,\cdot}, \boldsymbol{\mu}_{s_t}^{j}) + \lambda \mathbb{I}(s_t \neq s_{t+1}) \right] + g(\boldsymbol{y}_{T,\cdot}, \boldsymbol{\mu}_{s_T}^{j}) \right\}.$$

This algorithm drives the objective function downhill: step (i) minimizes the Gower distances without affecting the penalty term; in step (ii) the assignment of the updated prototype to the missing feature values zeroes the corresponding Gower distance terms, which were either zero or positive at the end of step (i); in step (iii) a further potential reduction of the objective function is achieved by optimizing over the states.

As suggested in Nystrup et al. (2020), finding the most likely state sequence requires $O(TK^2)$ operations, plus $O(TPK)$ operations to evaluate the loss function for each t and k, equivalent to one forward-backward pass of the expectation-maximization algorithm (Baum et al. 1970; Welch, 2003). Even so, the JM-mix estimator converges faster in practice, as hard clustering methods typically converge faster than soft clustering methods (Bottou and Bengio, 1994).

[5] Aschenbruck et al. (2023) evaluate alternative imputation methods for clustering mixed-type data, including internal imputation (iterative within clusters), external imputation (post-clustering), one-step imputation (single imputation post-clustering), and multiple imputation with pooling (aggregating results from multiple imputed datasets). We refer the reader to their work for a detailed comparison of these methods.

2.2 Hyperparameters selection

Witten and Tibshirani (2010) proposed selecting hyperparameters for a sparse k-means model with the value that maximizes the gap between the between-cluster sum of squares for the actual data and that for random permutations of the data. Alternatively, one can select hyperparameters based on the specific application, using methods such as backtesting (Shu et al. 2024a; Shu and Mulvey, 2024). For instance, if the estimated states are used in an investment strategy, λ can be chosen to maximize risk-adjusted returns, accounting for transaction costs (Nystrup et al. 2019; Nystrup et al. 2021). Additionally, if ground truth data is available, accuracy can be maximized against this benchmark.

In this work, we use a version of the generalized information criterion (GIC) from Cortese et al. (2024), shown to be effective in selecting the correct hyperparameter values through extensive simulation studies for continuous data. The criterion is given by

$$\text{GIC} := \frac{1}{T}\Big\{(\text{BCD}^S - \text{BCD}) + a_T M\Big\} + 2\Big(C(K^S) - C(K)\Big), \tag{3}$$

where $C(K) := -\log(K) - \frac{P}{2}\log(2\pi)$, and

$$\text{BCD} = \sum_{k=1}^{K} g(\boldsymbol{\mu}_k, \bar{\boldsymbol{\mu}}) n_k,$$

is the between cluster total deviance (Personn, 2023), essentially the JM-mix version of the between clusters sum of squares (Witten and Tibshirani, 2010). n_k is the number of points in cluster k, and $\bar{\boldsymbol{\mu}}$ is the vector of unconditional prototypes. a_T depends on the sample size T, and possibly the number of parameters and/or features considered; following Fan and Tang (2013), we set $a_T = \log(\log(T))\log(P)$. The superscript S denotes the saturated model, a JM-mix without any penalty on jumps (i.e., $\lambda^S = 0$) and with the maximum number of states[6] set to $K^S = 6$. The term $2(C(K^S) - C(K))$ adjusts for the difference in state complexity between the saturated and the current model.

The GIC essentially balances two key aspects.

Model fit is captured by the term $(\text{BCD}^S - \text{BCD})$, which compares the between-cluster deviance of the current model to that of a saturated model. A higher BCD indicates better clustering, as it reflects greater separation between clusters.

Model complexity is penalized through M, where $M = K(P + \Gamma_\lambda)$, and reflects the complexity of the model, considering both the number of states K and the number of jumps across states, $\Gamma_\lambda = \sum_t \mathbb{I}_{s_t \neq s_{t-1}}$. This penalty ensures the model does not overfit by excessively increasing the number of transitions. In our empirical application, we use the GIC solely to select λ. However, we do not apply the GIC in the simulation study, as its primary focus is on testing classification accuracy. In simulations, we

[6] The selection of the number of states for the saturated model, K^S, is less straightforward than that of λ^S. For details on this, please refer to Cortese et al. (2023a) or Cortese et al. (2024).

prefer to compare the ground truth of the simulated data directly with the estimation results.

3 Simulation study

In this section, we show a series of simulation studies assessing the JM-mix performance under various scenarios. After outlining the simulation setup in Subsection 3.1, Subsection 3.2 investigates performance of the model with complete data, while Subsection 3.3 focuses on incomplete data.

3.1 Simulation setup

We simulate features from a multivariate Gaussian HMM with 3 latent states,

$$\boldsymbol{y}_{t,\cdot} \mid s_t \sim N_P(\boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_P), \tag{4}$$

where the three centroids are defined as $(\mu, \ldots, \mu)'$, $(0, \ldots, 0)'$, and $(-\mu, \ldots, -\mu)'$. The state-conditional covariance matrix $\boldsymbol{\Sigma}_P$ has diagonal elements equal to 1 and off-diagonal elements $\rho_{ij} = \rho$, $i, j = 1, \ldots, P$, $i \neq j$. We consider three possible simulation setups for the parameters of the data generating process in (4), namely

(1) $\mu = 1, \rho = 0$;
(2) $\mu = 1, \rho = 0.20$;
(3) $\mu = 0.50, \rho = 0$.

Setups 1 and 2 correspond to high state separation with no correlation and mild correlation among features, respectively. Setup 3 represents lower state separation with no correlation among features.

The latent process $\{s_t\}$ is a Markov chain of first order with 3 states, with initial probabilities $\pi_i = 1/3$, $\forall i = 1, \ldots, 3$, and transition probability matrix denoted by $\boldsymbol{\Pi} \in \mathbb{R}^{3\times 3}$, with elements π_{ij}, $i, j = 1, \ldots, 3$. We simulate latent states considering self-transition probabilities $\pi_{kk} = 0.95$, $k = 1, 2, 3$, and probabilities of transitioning to a different state $\pi_{ij} = (1 - \pi_{kk})/2$, $i \neq j$.

Next, we transform half of these features $(P/2)$ into categorical variables with 3 levels $l = 1, 2, 3$, with probabilities

$$\mathbb{P}(l = \tilde{k} \mid s_t = k) = \begin{cases} 0.80 \text{ if } \tilde{k} = k, \\ 0.10 \text{ if } \tilde{k} \neq k, \end{cases}$$

for $\tilde{k} = 1, 2, 3$.

We vary $T \in \{50, 100, 500\}$ and $P \in \{4, 20, 50\}$, and simulate 100 datasets as described above for each combination of T and P.

For the JM-mix, for each combination and each simulated dataset, we vary λ in a grid between 0 and 1 with a step size of 0.05, and select the optimal λ as the value that maximizes the adjusted rand index (ARI, Hubert and Arabie 1985) computed between

true and estimated states sequences[7]. The ARI measures the similarity between two clusterings, adjusting for chance, with higher ARI values indicating better agreement between the estimated and true sequences. See Appendix A for details on how to compute this index.

In the following simulation studies and the empirical application in Sect. 4, we observe that upper bounds for λ greater than approximately 0.9 yield stable results for the decoded state sequence and estimated parameters; consequently, we search for λ in [0, 1].

We compare our proposal to two clustering alternatives: spectral clustering (spClust, Jia et al. 2014; Von Luxburg 2007) and a modified version of the k-means method that incorporates continuous and categorical variables by Huang (1998), known as k-prototypes (k-prot).

spClust involves constructing a similarity graph from the data points based on Euclidean distance for continuous and Hamming distance for categorical variables (Mbuga and Tortora, 2021), with a scaling parameter ζ adjusting the range of connections between points. The process continues with computing the graph Laplacian, performing dimensionality reduction, and applying a standard clustering algorithm like k-means. This approach effectively captures complex relationships that k-means alone might miss. Specifically, the algorithm is initialized 10 times with random seeds, clustering is performed for up to 300 iterations, and ζ is set equal to 20. These are the default settings for the `mspec` function from the `SpectralClMixed` package in `R` (Tortora et al. 2023).

The k-prot algorithm corresponds to our proposed method when no penalty on jumps is applied[8] (i.e., JM-mix with $\lambda = 0$). The initialization process follows the same approach used for JM-mix.

We highlight that, unlike JM-mix, both spClust and k-prot are not designed to account for temporal structures with jumps. Consequently, their performance in such settings is limited and aligns with what could be expected.

3.2 Performance evaluation with complete data

Table 1 illustrates the ARI computed between true and estimated sequences of states, for each simulated setup, using the three clustering models, and across different sample sizes.

The JM-mix consistently outperforms k-prot and spClust, demonstrating its robustness in various contexts. When $\mu = 1$ and $\rho = 0$ (setup 1), JM-mix notably excels, achieving ARIs up to 0.99. As correlation is introduced in setup 2 ($\rho = 0.2$), JM-mix continues to show superior performance, particularly at larger sample sizes, indicating its effectiveness in handling both correlation and larger datasets. For instance, at $T = 50$ and $P = 4$, JM-mix achieves ARIs close to 0.86, clearly outperforming the

[7] We also considered the cluster purity index, which can be thought of as the average of the proportions of data points in each cluster belonging to the majority class (weighted by the cluster size). The results obtained with this index, along with a brief discussion on other cluster quality measures, are provided in the supplementary material.

[8] For an alternative implementation of k-prot, please refer to the `R` package `ClustMixType` (Szepannek, 2018).

other two models. Setup 3 explores a smaller mean ($\mu = 0.5$) without correlation. Here, all models show a decline in their performance compared to setups 1 and 2, with JM-mix still leading.

We attribute the superior performance of JM-mix to its consideration of the temporal nature of the data, which other methods do not account for. An interesting observation is that JM-mix performs particularly well in scenarios with a low number of features ($P = 4$), where the other two methods barely surpass a performance level of 0.5, corresponding to roughly half-correct predictions. This result is notable because, in scenarios with an increasing number of features, the performance gap is reduced (but not uniformly over T) and accuracy improves for all methods, likely due to the increased information about the latent state provided by a larger number of observations that share the same conditional distribution. However, this property may not hold in real-world applications, where irrelevant or noisy features could negatively impact clustering performance. This highlights an important challenge of *feature selection* when working with real data. A promising direction to address this issue could involve extending the current framework to incorporate sparsity, as discussed in Nystrup et al. (2021), considering a sparse version of the JM-mix.

3.3 Performance evaluation with incomplete data

We consider three missing data mechanisms (Little and Rubin, 2019): MCAR (Missing Completely at Random), where the probability of missingness is independent of both observed and unobserved data; MAR (Missing at Random), where the probability of missingness depends only on observed data; and MNAR (Missing Not at Random), where the probability of missingness is related to unobserved variables.

To evaluate the impact of these mechanisms on classification performance, we introduce missingness into the datasets simulated under setup 1. To ensure variability and robustness in assessing the ability of the method to handle missing data, we generate simulations using different random seeds. Missing data is introduced with the `R` package **missMethods** (Rockel, 2022), targeting 10% and 20% of the total observations. MAR observations are created by censoring variables in even positions based on the immediately preceding variable, while MNAR observations are introduced using a censoring mechanism dependent on the variable itself. It is noteworthy that spectral clustering was not considered in this analysis as it does not handle missing values.

Results from Table 2 indicate that the classification performance of the JM-mix model remains robust in the presence of missing data, except under the MNAR mechanism, where all methods face challenges. This is likely because MNAR missingness depends on unobserved values, leading to a loss of critical information that affects model accuracy. In the empirical application of Sect. 4, involving air quality monitoring data, missingness is caused by sensor malfunctions, transmission errors, or equipment failures (Kasam et al. 2014). Consequently, the missingness mechanism in such data is typically assumed to be MAR or MCAR (Yozgatligil et al. 2013).

Notably, JM-mix performs particularly well with a lower number of features ($P = 4$). Additionally, JM-mix is less impacted by a higher percentage of missing data. The ARI decreases by approximately 0.05 points when increasing the missingness level

Table 1 Adjusted Rand Index (ARI) computed between true and estimated latent sequences for the three models: k-prototypes (k-prot), spectral clustering (spClust), and jump model for mixed data (JM-mix) for the three setups. Bold values indicate the highest ARI among the three methods for each setup and each combination of T and P. Monte Carlo standard deviations are in parentheses

Setup 1				
T	Method	$P = 4$	$P = 20$	$P = 50$
50	k-prot	0.55 (0.22)	0.82 (0.33)	0.93 (0.32)
	spClust	0.54 (0.23)	0.81 (0.32)	0.93 (0.32)
	JM-mix	**0.86** (0.18)	**0.97** (0.11)	**0.99** (0.14)
100	k-prot	0.65 (0.15)	**1.00** (0.17)	**1.00** (0.15)
	spClust	0.65 (0.12)	**1.00** (0.17)	**1.00** (0.15)
	JM-mix	**0.91** (0.08)	0.98 (0.07)	0.98 (0.04)
500	k-prot	0.73 (0.05)	**1.00** (0.03)	**1.00** (0.05)
	spClust	0.72 (0.06)	**1.00** (0.03)	**1.00** (0.05)
	JM-mix	**0.93** (0.05)	0.98 (0.01)	0.98 (0.02)
Setup 2				
T	Method	$P = 4$	$P = 20$	$P = 50$
50	k-prot	0.52 (0.25)	0.77 (0.31)	0.86 (0.31)
	spClust	0.51 (0.25)	0.76 (0.31)	0.86 (0.31)
	JM-mix	**0.86** (0.15)	**0.97** (0.13)	**0.98** (0.09)
100	k-prot	0.67 (0.12)	0.95 (0.16)	0.97 (0.18)
	spClust	0.66 (0.13)	0.94 (0.16)	0.97 (0.17)
	JM-mix	**0.90** (0.11)	**0.98** (0.06)	**0.98** (0.06)
500	k-prot	0.74 (0.04)	0.95 (0.03)	**0.98** (0.04)
	spClust	0.73 (0.04)	0.95 (0.03)	**0.98** (0.04)
	JM-mix	**0.92** (0.03)	**0.98** (0.01)	**0.98** (0.01)
Setup 3				
T	Method	$P = 4$	$P = 20$	$P = 50$
50	k-prot	0.46 (0.20)	0.82 (0.31)	0.87 (0.32)
	spClust	0.46 (0.20)	0.81 (0.32)	0.87 (0.32)
	JM-mix	**0.84** (0.20)	**0.97** (0.10)	**0.97** (0.12)
100	k-prot	0.56 (0.11)	**0.98** (0.17)	**1.00** (0.17)
	spClust	0.56 (0.11)	**0.98** (0.17)	**1.00** (0.17)
	JM-mix	**0.89** (0.10)	**0.98** (0.08)	0.98 (0.08)
500	k-prot	0.63 (0.05)	**0.99** (0.01)	**1.00** (0.03)
	spClust	0.61 (0.05)	**0.99** (0.01)	**1.00** (0.03)
	JM-mix	**0.91** (0.02)	0.98 (0.00)	0.98 (0.00)

from 10% to 20%, whereas the k-prot model shows a higher decline, with drops of up to 0.10 points.

The average imputation error, measured by the Gower distance between true and imputed data, decreases with increasing T in both approaches. However, the differences in imputation accuracy are not significant, with JM-mix showing a slight advantage, reducing imputation errors by approximately 0.01.

3.4 Convergence

To evaluate local convergence of the estimation algorithm, we followed an approach similar to Szepannek et al. (2024), estimating JM-mix under setup 1 with parameters $T = 200$, $P = 20$, using 30 different initializations and allowing a maximum of 50 iterations per run. We analyze the convergence of the objective function defined in equation (1) and monitored the number of observations moved.

Fig. 1 illustrates that all initializations converge within less than four iterations.[9] Only three runs of the algorithm became trapped in local maxima. Additionally, the number of observations moved after each iteration of the algorithm decreases rapidly after the first iteration across all cases. To evaluate the validity of the proposed algorithm in handling missing data, we repeat the procedure under different missing data mechanisms (MAR, MCAR, and MNAR) and examined the convergence of the objective function with 20% missingness. The results demonstrate that the algorithm successfully drives the objective function downhill, achieving convergence during the estimation process.[10]

4 Application to air quality data

We use the JM-mix to track air quality levels, treating it as a latent process that influences observable continuous and categorical variables. Our analysis focuses on air quality data from ARPA Lombardia[11] for Milan, spanning June 5, 2021, to June 5, 2024.

4.1 Data description

Our dataset includes daily weather data on average, maximum, and minimum air temperature (Temp, $°C$), relative humidity (RH, %), rainfall (RF, mm), global radiation (GR, W/m^2), and wind speed (WS, m/s). These measurements are taken from the weather station at Piazza Zavattari, Milan, which was selected because it provides data for all the aforementioned variables.

Additionally, we collect air pollutant data on daily average, maximum, and minimum levels of nitrogen dioxide (NO_2) and ozone (O_3), as well as daily average levels

[9] Results for the objective function, as presented in Szepannek et al. (2024), are available in the supplementary material.

[10] Detailed results are provided as supplementary material.

[11] Data available here https://www.arpalombardia.it/dati-e-indicatori/.

Table 2 Adjusted Rand Index (ARI) for k-prototypes (k-prot) and jump model for mixed data (JM-mix) under MCAR, MAR, and MNAR mechanisms with 10% and 20% missing data (setup 1). Bold values indicate the highest ARI among the two methods for each setup and each combination of T and P. Monte Carlo standard deviations are in parentheses

	T	Method	$P = 4$	$P = 20$	$P = 50$
10% Missing Data					
MCAR	50	k-prot	0.52 (0.22)	0.78 (0.32)	0.90 (0.33)
		JM-mix	**0.84** (0.18)	**0.98** (0.14)	**0.99** (0.13)
	100	k-prot	0.58 (0.11)	**1.00** (0.15)	**1.00** (0.13)
		JM-mix	**0.89** (0.09)	0.98 (0.06)	**0.98** (0.09)
	500	k-prot	0.64 (0.04)	0.98 (0.08)	**1.00** (0.04)
		JM-mix	**0.90** (0.03)	**0.98** (0.03)	0.98 (0.02)
MAR	50	k-prot	0.44 (0.20)	0.79 (0.31)	0.92 (0.32)
		JM-mix	**0.82** (0.19)	**0.97** (0.13)	**0.98** (0.11)
	100	k-prot	0.55 (0.13)	**0.97** (0.17)	**1.00** (0.20)
		JM-mix	**0.86** (0.09)	**0.97** (0.09)	0.98 (0.07)
	500	k-prot	0.64 (0.07)	**0.97** (0.02)	**1.00** (0.08)
		JM-mix	**0.89** (0.05)	**0.97** (0.00)	0.98 (0.02)
MNAR	50	k-prot	0.20 (0.20)	0.44 (0.27)	0.45 (0.29)
		JM-mix	**0.38** (0.20)	**0.60** (0.27)	**0.58** (0.29)
	100	k-prot	0.36 (0.15)	0.52 (0.15)	0.58 (0.18)
		JM-mix	**0.50** (0.15)	**0.59** (0.15)	**0.58** (0.18)
	500	k-prot	0.36 (0.06)	0.55 (0.07)	0.56 (0.08)
		JM-mix	**0.51** (0.06)	**0.55** (0.07)	**0.55** (0.08)
20% Missing Data					
MCAR	50	k-prot	0.43 (0.20)	0.79 (0.30)	0.92 (0.31)
		JM-mix	**0.80** (0.20)	**0.95** (0.13)	**0.97** (0.10)
	100	k-prot	0.52 (0.11)	0.93 (0.17)	**1.00** (0.11)
		JM-mix	**0.83** (0.11)	**0.97** (0.05)	0.98 (0.06)
	500	k-prot	0.56 (0.05)	0.93 (0.05)	**0.99** (0.04)
		JM-mix	**0.85** (0.03)	**0.97** (0.01)	0.98 (0.01)
MAR	50	k-prot	0.44 (0.20)	0.77 (0.31)	0.87 (0.34)
		JM-mix	**0.79** (0.18)	**0.95** (0.16)	**0.95** (0.14)
	100	k-prot	0.49 (0.12)	0.94 (0.16)	**1.00** (0.16)
		JM-mix	**0.80** (0.12)	**0.95** (0.08)	0.96 (0.10)
	500	k-prot	0.52 (0.04)	0.94 (0.04)	**1.00** (0.06)
		JM-mix	**0.84** (0.04)	**0.95** (0.02)	0.96 (0.04)
MNAR	50	k-prot	0.20 (0.20)	0.44 (0.27)	0.45 (0.29)
		JM-mix	**0.38** (0.20)	**0.60** (0.27)	**0.58** (0.29)
	100	k-prot	0.36 (0.15)	0.52 (0.15)	0.58 (0.18)

Table 2 (continued)

	T	Method	$P = 4$	$P = 20$	$P = 50$
		JM-mix	**0.50** (0.15)	**0.59** (0.15)	**0.58** (0.18)
	500	k-prot	0.36 (0.06)	0.55 (0.07)	0.56 (0.08)
		JM-mix	**0.51** (0.06)	**0.55** (0.07)	**0.55** (0.08)

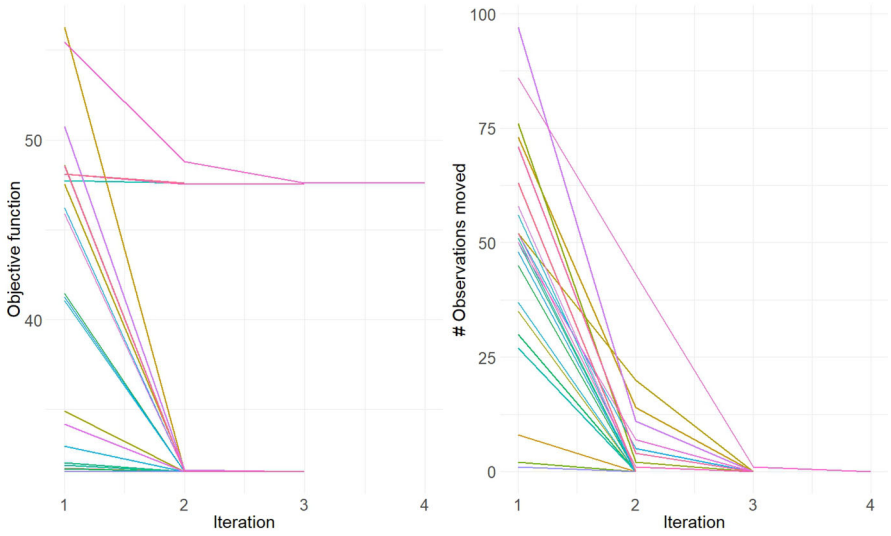

Fig. 1 The left panel shows the convergence of the objective function over iterations, while the right panel depicts the number of observations moved after each iteration

of particulate matter with diameter around 2.5 μm ($PM_{2.5}$), and with diameter around 10 μm (PM_{10}). The ARPA Lombardia air quality monitoring network consists of fixed stations that provide continuous data through automatic analyzers at regular intervals (ARPA Lombardia, 2023; Maranzano and Algieri, 2024). This data includes aggregated municipal values from modeling simulations integrated with the monitoring network data.

Figures 2 and 3 show the time series of pollutant and weather data. O_3 trends similarly to Temp and GR, while NO_2, $PM_{2.5}$, and PM_{10} levels are higher in winter.

We incorporate categorical features such as wind speed greater[12] than 0.7 m/s, which is the observed median value (windy feature: "Yes" for windy days); similarly, we include a rainy feature ("Yes" for rainy days) and consider a day to be rainy when rainfall is above the observed median value of 0 mm. Additional categorical features are month (1 to 12), weekend (yes or no), and holiday (yes or no). The latter two variables are included to account for changes in human activity that can affect pollution levels. Furthermore, we include 7-day moving averages of NO_2, O_3, $PM_{2.5}$,

[12] This threshold corresponds to "light air" on the Beaufort scale.

PM_{10}, Temp, WS, RF, RH, and GR. Additionally, similar as in Cortese et al. (2023b), we calculate 7-day moving correlations between each weather variable and each air pollutant. The moving correlation is computed by calculating the Pearson correlation coefficient over a rolling 7-day window, providing a dynamic view of the relationship between variables over time. The final dataset comprises 1,097 time observations and 56 features.

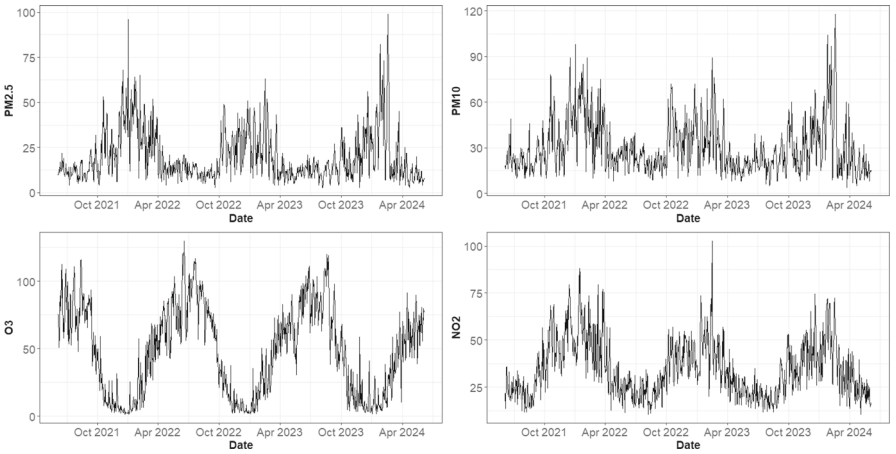

Fig. 2 Daily time series of pollutants $PM_{2.5}$, PM_{10}, O_3, and NO_2 observed for the period June 2021 to June 2024

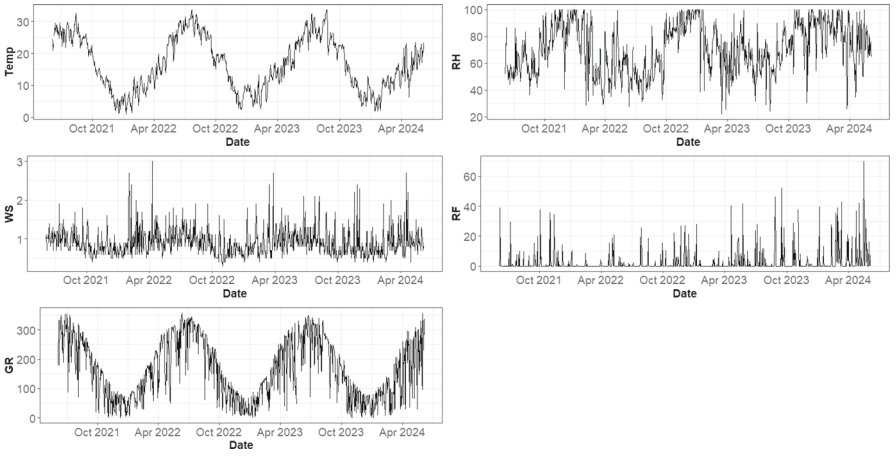

Fig. 3 Daily time series temperature (Temp), relative humidity (RH), wind speed (WS), rainfall (RF), and global radiation (GB) observed for the period June 2021 to June 2024

4.2 Summary statistics

Table 3 presents the summary statistics and percentage of missing values for continuous variables in the dataset. The table includes the mean, standard deviation (SD), minimum, and maximum values for each variable. The percentage of missing values (NA %) indicates the proportion of missing data for each variable. The values reveal that Temp, RH, WS, RF and GR have a small percentage of missing data, while other variables have no missing values.

The mean values indicate that the average levels of $PM_{2.5}$ and PM_{10} are below the standard safety limits (WHO, 2021), but their maximum values show that there are instances of very high pollution. O_3 and NO_2 also show high variability, with maximum values indicating significant pollution episodes. Temp and RH exhibit substantial variability, reflecting seasonal changes. WS and RF have relatively low average values but show occasional extreme events. GR shows a wide range, indicating variability in solar radiation received.

Table 4 displays the correlation matrix for continuous variables. Correlations greater than 0.5 or lower than -0.5 are highlighted in bold to emphasize strong relationships between variables.

We observe strong positive relationships between $PM_{2.5}$ and PM_{10}, and strong negative relationships between O_3 and NO_2. Temperature is positively correlated with O_3 and GR, indicating higher ozone levels and global radiation during summer time. In contrast, temperature is negatively correlated with NO_2 and RH, suggesting lower nitrogen dioxide levels and humidity with higher temperatures.

Table 5 shows the percentage of observations for each category and the percentage of missing values for each categorical variable.

The percentages reveal that windy days are almost equally distributed, while there is a substantial prevalence of non-rainy days. Weekends account for about 29% of the days, while holidays are rare, making up only about 3% of the observations.

Table 3 Summary statistics and percentage of missing values for continuous variables

Variable	Mean	SD	Min	Max	NA %
$PM_{2.5}$	19.88	13.79	3.00	99.00	0.00
PM_{10}	30.28	17.30	4.00	118.00	0.00
O_3	48.18	32.60	1.40	129.80	0.00
NO_2	34.73	15.07	10.50	102.60	0.00
Temp	16.70	8.28	1.00	33.70	0.36
RH	70.91	18.42	22.00	100.00	0.36
WS	0.90	0.34	0.30	3.00	0.36
RF	2.36	6.95	0.00	70.00	0.36
GR	162.85	100.56	0.40	357.20	0.27

Table 4 Correlation (upper right) and partial correlation (lower left) matrix for continuous variables with correlations greater than 0.5 or lower than −0.5 highlighted in bold. Partial correlation coefficients are in italics

	$PM_{2.5}$	PM_{10}	O_3	NO_2	Temp	RH	WS	RF	GR
$PM_{2.5}$	-	**0.97**	**−0.58**	**0.76**	**−0.53**	0.40	−0.44	−0.16	−0.47
PM_{10}	***0.93***	-	**−0.51**	**0.75**	−0.43	0.31	−0.43	−0.20	−0.39
O_3	*0.05*	*0.01*	-	**−0.73**	**0.87**	**−0.74**	0.39	−0.06	**0.86**
NO_2	*−0.10*	*0.33*	*−0.39*	-	**−0.68**	0.33	−0.47	−0.11	**−0.56**
Temp	*−0.37*	*0.39*	*0.45*	*−0.24*	-	**−0.59**	0.20	−0.07	**0.80**
RH	*0.16*	*−0.04*	*−0.42*	***− 0.50***	*0.11*	-	**−0.33**	0.33	**−0.79**
WS	*−0.09*	*0.10*	*0.11*	*−0.35*	*−0.29*	*−0.34*	-	0.20	0.23
RF	*0.03*	*−0.12*	*0.23*	*0.22*	*0.07*	*0.37*	*0.25*	-	−0.25
GR	*0.11*	*−0.10*	*0.29*	*−0.06*	*0.25*	*−0.41*	*−0.17*	*−0.15*	-

Table 5 Frequency table for the categorical variables

Variable	No (%)	Yes (%)	NAs (%)
Windy	51.64	48.00	0.36
Rainy	69.64	30.00	0.36
Weekend	71.37	28.63	0
Holiday	96.72	3.28	0

4.3 Results

We select the model based on the modified version of the GIC by Cortese et al. (2023a) and Cortese et al. (2024), described in Sect. 2.2. According to this, we select $\lambda = 0.55$. The GIC selects $K = 2$, but we set $K = 4$ following the air quality index (AQI, Rule 2005) levels computed from observed pollutants (O_3, NO_2, $PM_{2.5}$, and PM_{10}), which correspond to four categories: "Good", "Moderate", "Unhealthy for Sensitive Groups" (US), and "Unhealthy". This choice of $K = 4$ is a compromise between model fit and interpretability, aligning with the AQI categories for better practical relevance and ease of understanding.

Table 6 presents the conditional medians and modes for the variables in each air quality state. This table allows us to comprehensively describe each state in terms of air quality, providing insights into the typical environmental conditions associated with each category. Based on these conditional medians and modes, we assign meaningful labels to each state, using terminology from the AQI. For example, the "Unhealthy" state is characterized by higher levels of $PM_{2.5}$, PM_{10}, and NO_2, while the "Good" state corresponds to lower values of these pollutants. Additionally, "Good" and "Moderate" air quality states are often associated with windy days, contrasting with the "Unhealthy" and "US" states, which tend to occur calmer days. December and October emerge as the most polluted months, suggesting seasonal influences on

Table 6 State conditional medians and modes

Variable	Good	Moderate	US	Unhealthy
$PM_{2.5}$	12.10	14.52	23.10	33.20
PM_{10}	21.66	23.83	34.94	44.30
O_3	82.35	58.88	25.91	10.76
NO_2	22.58	31.49	41.01	49.28
Temp	26.29	15.21	14.11	6.46
RH	57.38	63.94	82.59	85.14
WS	1.00	1.01	0.76	0.77
RF	1.68	3.26	3.49	1.34
GR	258.51	194.09	99.59	58.74
Windy	Yes	Yes	No	No
Rainy	No	No	No	No
Month	7	4	10	12
Weekend	No	No	No	No
Holiday	No	No	No	No

air quality. The absence of a significant difference in the rainy, weekend, and holiday features across states is likely due to the overall prevalence of non-rainy days and working days, as shown in Table 5.

Table 7 shows the percentage of visits in each state, reflecting how often the air quality falls into each of the four categories. The "Good" category has the highest percentage (33.27), followed by "US" (24.07), "Unhealthy" (21.97), and "Moderate" (20.69).

Table 8 details the correlation and partial correlation coefficients conditional on each state. This table helps to understand the relationships between different environmental factors under varying air quality conditions. For example, in the "Good" state, there is a positive correlation between $PM_{2.5}$ or PM_{10} and Temp, while in the other states, this correlation is negative. Additionally, the partial correlation between RF and PM_{10} is negative in the US and Unhealthy states, while for $PM_{2.5}$ it is low. This observation indicates that rainfall might be more effective in removing larger particulate matter from the atmosphere compared to smaller particles.

Table 7 Percentage of visits in each state

State	Percentage (%)
Good	33.27
Moderate	20.69
US	24.07
Unhealthy	21.97

Our model provides deeper insights into these correlation patterns, highlighting how they vary between different air quality states.

Figures 4 and 5 illustrate the temporal variations of air pollutants and meteorological variables including background color bands representing air quality categories: "Good" (green), "Moderate" (yellow), "US" (orange), and "Unhealthy" (red).

We observe significant seasonal variations in $PM_{2.5}$ and PM_{10}, with higher concentrations during colder months, likely due to increased heating activities and stable atmospheric conditions (Sharma et al. 2021). These seasonal patterns are evident across multiple pollutants, with colder months generally associated with poorer air quality. Peaks frequently enter the "Unhealthy" range, occasionally persisting for extended periods. O_3 levels exhibit a distinct seasonal cycle, with higher concentrations during warmer months due to increased photochemical activity, frequently fluctuating between "Good" and "Moderate" categories. NO_2 levels display seasonal fluctuations with higher values during colder months, possibly due to increased emissions from heating and reduced atmospheric dispersion, sometimes reaching the US category. These seasonal cycles align with the traffic limitation measures implemented when PM_{10} levels exceed set thresholds. For instance, in Milan, the threshold was exceeded 84 times[13] in 2022 and 49 times[14] in 2023 (source: ARPA Lombardia).

Figure 6 shows the time evolution of the state sequences fitted with JM-mix and the AQI computed with the standard formula. The AQI for pollutant q, denoted as AQI_q, is given by

$$\text{AQI}_q = \frac{(I_{high} - I_{low})}{(C_{high} - C_{low})} \times (C_q - C_{low}) + I_{low},$$

where AQI_q is the AQI for pollutant q, C_q is the concentration of pollutant q, C_{low} and C_{high} are the breakpoints closest to C_q, and I_{low} and I_{high} are the AQI values corresponding to C_{low} and C_{high}. The general AQI is just the maximum AQI_q among all pollutants.

The AQI sequence appears highly volatile, indicating rapid shifts in air quality due to its sensitivity to short-term fluctuations in pollutant concentrations. In contrast, the JM-mix sequence shows more stable transitions between states, suggesting it captures more consistent and prolonged periods of air quality. Essentially, the JM-mix acts as a smoothing mechanism for the categorical AQI variable, providing a clearer and more stable representation of air quality over time.

The differences between the AQI and JM-mix sequences arise from their distinct nature. The AQI is computed using a standard formula that primarily considers the immediate concentrations of key pollutants (O_3, NO_2, $PM_{2.5}$, and PM_{10}). It reacts quickly to changes in these pollutants, hence the high volatility. The JM-mix incorporates a broader set of variables: this comprehensive approach allows it to provide a more nuanced representation of air quality states, considering the underlying factors that influence pollutant levels.

[13] https://www.arpalombardia.it/agenda/notizie/2024/qualita-dell-aria-2023-l-anno-migliore-di-sempre/

[14] https://www.arpalombardia.it/media/iphjbgtu/analisi-anno-2023_def.pdf

Table 8 State conditional correlation and partial correlation coefficients. Correlations with absolute value greater than 0.5 are highlighted in bold.

	Conditional Correlation				Conditional Partial Correlation			
Variable	Good	Moderate	US	Unhealthy	Good	Moderate	US	Unhealthy
$PM_{2.5}$, Temp	0.34	−0.29	−0.12	−0.28	*0.00*	*−0.17*	***−0.58***	*−0.38*
$PM_{2.5}$, RH	0.02	−0.22	−0.08	0.35	*0.35*	*0.06*	*0.21*	*−0.12*
$PM_{2.5}$, WS	−0.26	−0.24	−0.43	−0.42	*−0.10*	*−0.10*	*−0.02*	*0.01*
$PM_{2.5}$, RF	−0.07	−0.19	−0.28	−0.16	*−0.04*	*−0.05*	*0.05*	*0.19*
$PM_{2.5}$, GR	−0.04	−0.19	0.09	0.04	*0.03*	*−0.09*	*0.13*	*0.04*
PM_{10}, Temp	0.37	−0.23	0.03	0.42	*0.19*	*0.17*	***0.59***	*0.42*
PM_{10}, RH	−0.06	−0.25	−0.15	0.24	*−0.15*	*0.02*	*−0.15*	*0.24*
PM_{10}, WS	−0.20	−0.22	−0.47	−0.39	*0.12*	*0.15*	*0.00*	*0.04*
PM_{10}, RF	−0.12	−0.21	−0.33	−0.25	*−0.07*	*−0.04*	*−0.11*	*−0.25*
PM_{10}, GR	−0.03	−0.13	0.17	0.02	*−0.14*	*0.04*	*−0.14*	*0.02*
O_3, Temp	**0.73**	**0.54**	0.49	0.44	*0.38*	*0.44*	*0.28*	*0.04*
O_3, RH	**−0.55**	−0.28	−0.40	**−0.75**	*−0.36*	***− 0.58***	*−0.29*	***− 0.51***
O_3, WS	0.02	0.20	0.26	**0.74**	*0.07*	*−0.06*	*0.25*	*0.44*
O_3, RF	−0.24	−0.04	0.00	−0.11	*0.11*	*0.28*	*0.20*	*0.09*
O_3, GR	**0.55**	**0.51**	0.49	0.37	*0.16*	*−0.04*	*0.30*	*0.37*
NO_2, Temp	−0.12	−0.38	−0.24	−0.28	*−0.21*	*0.10*	*−0.22*	*−0.44*
NO_2, RH	−0.03	−0.32	−0.25	−0.15	***−0.46***	***− 0.58***	*−0.37*	***− 0.63***
NO_2, WS	−0.32	−0.34	−0.49	−0.36	*−0.28*	*−0.39*	*−0.25*	*−0.21*
NO_2, RF	0.03	−0.21	−0.25	− −0.17	*0.17*	*0.25*	*0.24*	*0.26*
NO_2, GR	−0.18	−0.15	0.12	0.30	*−0.06*	*−0.16*	*0.03*	*0.18*

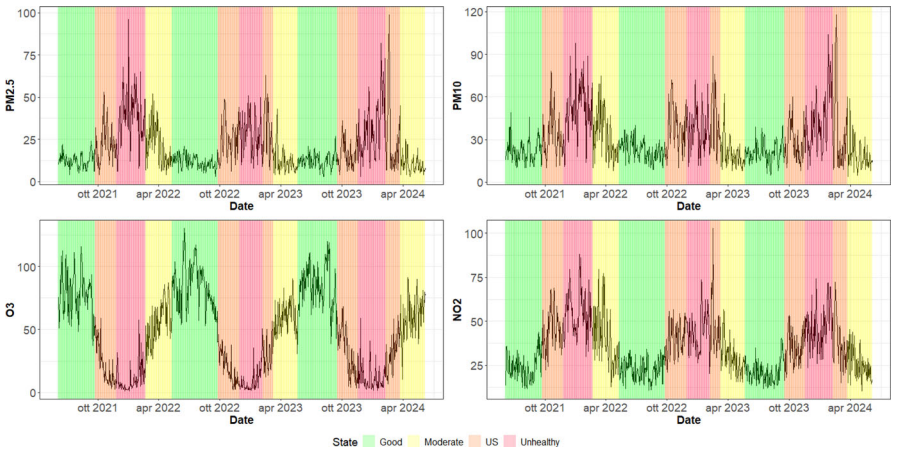

Fig. 4 Daily time series of pollutants $PM_{2.5}$, PM_{10}, O_3, and NO_2 observed for the period June 2021 to June 2024, with the decoded state sequence fitted by the JM-mix. Colors represent different air quality states: green is "Good", yellow is "Moderate", orange is "Unhealthy for Sensitive Groups (US)", and red is "Unhealthy"

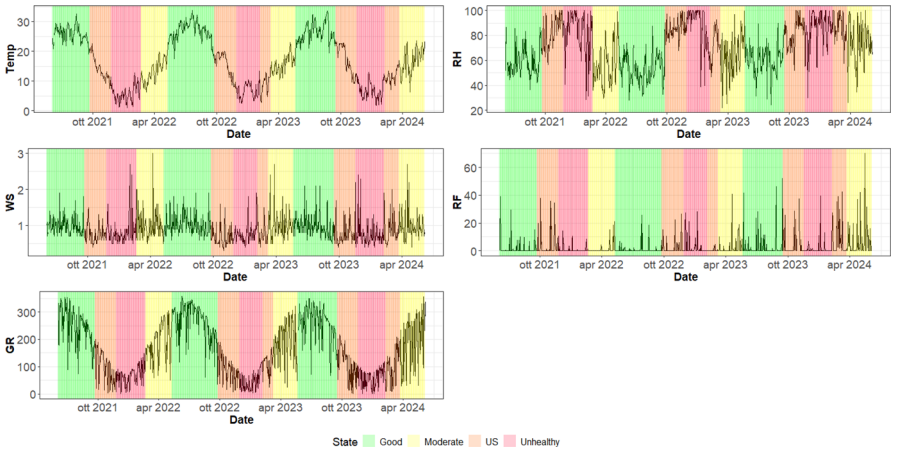

Fig. 5 Daily time series temperature (Temp), relative humidity (RH), wind speed (WS), rainfall (RF), and global radiation (GB) observed for the period June 2021 to June 2024, with the decoded state sequence fitted by the JM-mix. Colors represent different air quality states: green is "Good", yellow is "Moderate", orange is "Unhealthy for Sensitive Groups (US)", and red is "Unhealthy"

The reduced volatility in the JM-mix sequence might be beneficial in contexts where stability and consistency are crucial. For instance, in public health and policy-making, having a stable representation of air quality can lead to better short to medium term planning and responses.

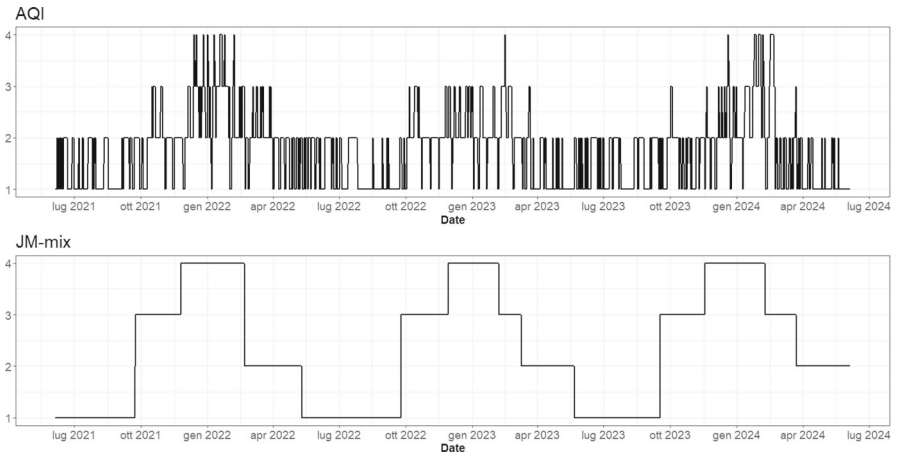

Fig. 6 Visual comparison between the Air Quality Index (AQI) and latent states of the JM-mix. The states are ordered as follows: from 1 to 4, "Good", "Moderate", "Unhealthy for Sensitive Groups", and "Unhealthy"

5 Discussion

In this work, we make three main contributions. First, we introduce the statistical jump model for mixed-type data, a novel method for clustering mixed-type data over time, incorporating regime persistence to enhance interpretability and utility. Second, our approach provides a fast and efficient solution for managing missing data, ensuring robust clustering results. Third, we demonstrate the applicability of our method through an empirical study on air quality data, showing that the proposal can infer more persistent air quality regimes compared to the traditional air quality index.

Simulation studies show that our model outperforms existing models like spectral clustering and k-prototypes in terms of clustering accuracy and robustness against missing data.

In future work, we aim to investigate the issue raised by Hennig and Liao (2013), who argue that categorical variables may dominate distance-based clustering unless weighted down. This is significant for practical applications, and our proposed algorithm allows for setting variable-specific weights for clustering. A potential variable selection scheme could be based on Hennig and Murphy (2023), who introduced variable importance to assess the influence of each variable on clustering results. Another approach could involve a sparse version of our jump model with mixed-type data, as in Nystrup et al. (2021), using Gower distance weights as defined in Sect. 2 and applying soft thresholding for weights estimation.

Appendix A

The adjusted Rand index (ARI, Hubert and Arabie 1985) measures the degree of overlap between two groupings of partitions: $G^m = \{G_1^m, \ldots, G_h^m\}$, obtained with a clustering algorithm on a set of n elements, and $G^r = \{G_1^r, \ldots, G_q^r\}$, interpreted as the real clustering.

Given the following contingency table:

$$\begin{array}{c|cccc|c} & G_1^m & G_2^m & \cdots & G_h^m & \sum_{j=1}^h n_{ij} \\ \hline G_1^r & n_{11} & n_{12} & \cdots & n_{1h} & a_1 \\ G_2^r & n_{21} & n_{22} & \cdots & n_{2h} & a_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ G_q^r & n_{q1} & n_{q2} & \cdots & n_{qh} & a_q \\ \hline \sum_{i=1}^q n_{ij} & b_1 & b_2 & \cdots & b_h & n \end{array} \tag{A1}$$

$\mathrm{ARI}(G^r, G^m)$ can be calculated using the methodology proposed by Das and Biswas (2023) as

$$\mathrm{ARI}\left(G^r, G^m\right) := \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}}. \tag{A2}$$

Here, n_{ij} represents the count of shared elements within clusters G_i^m and G_j^r, while a_i aggregates all n_{ij} values linked to any G_j^r in G^r, and all G_i^m in G^m. Similarly, b_j sums all n_{ij} values related to any G_i^m in G^m, and all G_j^r in G^r.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11634-025-00631-y.

Author contributions Corresponding author contributed to the study conception and design. Second author supervised the entire project. All authors contributed to material preparation, data collection and analysis. The first draft of the manuscript was written by corresponding author and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by Consiglio Nazionale Delle Ricerche (CNR) within the CRUI-CARE Agreement. This work has been supported by the European Commission, NextGenerationEU, Mission 4 Component 2, "Dalla ricerca all'impresa", Innovation Ecosystem RAISE "Robotics and AI for Socio-economic Empowerment", ECS00000035.

Data availability Data is available upon request from the corresponding author.

Code availability The code used for estimation and analysis is publicly available at the following repository: https://github.com/FedericoCortese/JM-mix.git.

Declarations

Conflict of interest The authors declare that there are no Conflict of interest.

References

Aghabozorgi S, Shirkhorshidi AS, Wah TY (2015) Time-series clustering: a decade review. Inf Syst 53:16–38

ARPA Lombardia (2023) Elenco e posizione delle stazioni di monitoraggio qualità dell'aria e dei sensori. https://www.dati.lombardia.it/Ambiente/Stazioni-qualit-dell-aria/ib47-atvt

Arthur D, Vassilvitskii S, et al. (2007) k-means++: The advantages of careful seeding. In: proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms (vol. 7, pp. 1027-1035)

Aschenbruck R, Szepannek G, Wilhelm AF (2023) Imputation strategies for clustering mixed-type data with missing values. J. Classif. 40(1):2–24

Aydınhan AO, Kolm PN, Mulvey JM, Shu Y (2024) Identifying patterns in financial markets: extending the statistical jump model for regime identification. Ann Oper Res. https://doi.org/10.1007/s10479-024-06035-z

Bartolucci F, Farcomeni A, Pennoni F (2013) Latent Markov models for longitudinal data. Chapman & Hall/CRC Press, Boca Raton, FL

Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Stat 41(1):164–171

Bemporad A, Breschi V, Piga D, Boyd SP (2018) Fitting jump models. Automatica 96:11–21

Bosancic T, Nie Y, Mulvey J (2024) Regime-aware factor allocation with optimal feature selection. J Financ Data Sci 6(3):10–37

Bottou L, Bengio Y (1994) Convergence properties of the *k*-means algorithms. Adv Neural Inf Process Syst 7:585–592

Chi JT, Chi EC, Baraniuk RG (2016) *k*-POD: a method for *k*-means clustering of missing data. Am Stat 70(1):91–99

Cortese FP, Kolm PN, Lindström E (2023a) Generalized information criteria for sparse statistical jump models. P. Linde (Ed.), Symposium i anvendt statistik (Vol. 44, p. 69-78). Copenhagen: Copenhagen Business School. Retrieved from http://www.statistiksymposium.dk/

Cortese FP, Kolm PN, Lindström E (2023) What drives cryptocurrency returns? A sparse statistical jump model approach. Digit Financ 5:1–36

Cortese FP, Kolm PN, Lindstrom E (2024) Generalized information criteria for high-dimensional sparse statistical jump models. Available at SSRN 4774429

Cortese FP, Pievatolo A (2024) Spatio-temporal jump model for urban thermal comfort monitoring. arXiv preprint arXiv:2411.09726

Das S, Biswas A (2023) Analyzing correlation between quality and accuracy of graph clustering. Patgiri R, Deka GC, Biswas A (Eds.), Principles of big graph: in-depth insight (Vol. 128, p. 135-163). Elsevier

Dinh D-T, Huynh V-N, Sriboonchitta S (2021) Clustering mixed numerical and categorical data with missing values. Inf Sci 571:418–442

Fan Y, Tang CY (2013) Tuning parameter selection in high dimensional penalized likelihood. J R Stat Soc Ser B (Stat Methodol) 75(3):531–552
Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857–871
Hamilton J (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57:357–384
Hennig C, Liao TF (2013) How to find an appropriate clustering for mixed type variables with application to socio-economic stratification. J R Stat Soc Ser C Appl Stat 62(3):309–369
Hennig C, Murphy K (2023) Quantifying variable importance in cluster analysis. Antwerp Belgium, p. 85
Huang Z (1998) Extensions to the *k*-means algorithm for clustering large data sets with categorical values. Data Min Knowl Discov 2(3):283–304
Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218
Jia H, Ding S, Xu X, Nie R (2014) The latest research progress on spectral clustering. Neural Comput Appl 24:1477–1486
Kasam AA, Lee BD, Paredis CJ (2014) Statistical methods for interpolating missing meteorological data for use in building simulation. Build Simul 7:455–465
Kehagias A (2004) A hidden Markov model segmentation procedure for hydrological and environmental time series. Stoch Environ Res Risk Assess 18:117–130
Liao S-H, Wen C-H (2007) Artificial neural networks classification and clustering of methodologies and applications-literature analysis from 1995 to 2005. Expert Syst Appl 32(1):1–11
Little RJ, Rubin DB (2019) Statistical analysis with missing data, vol 793. John Wiley & Sons, Hoboken
Maranzano P, Algieri A (2024) ARPALData: an R package for retrieving and analyzing air quality and weather data from ARPA Lombardia (Italy). Environ Ecol Stat 31:1–32
Mbuga F, Tortora C (2021) Spectral clustering of mixed-type data. Stats 5(1):1–11
Nystrup P, Boyd S, Lindström E, Madsen H (2019) Multi-period portfolio selection with drawdown control. Ann Oper Res 282(1–2):245–271
Nystrup P, Kolm PN, Lindström E (2021) Feature selection in jump models. Expert Syst Appl 184:115558
Nystrup P, Lindström E, Madsen H (2020) Learning hidden Markov models with persistent states by penalizing jumps. Expert Syst Appl 150:113307
Personn J (2023) Robust statistical jump model with feature selection. Lund University, Faculty of Science, Centre for Mathematical Sciences
Podani J (1999) Extending Gower's general coefficient of similarity to ordinal characters. Taxon 48(2):331–340
Rockel T (2022) missMethods: Methods for missing data. (R package version 0.4.0)
Rule CAI (2005) Environmental protection agency 40 cfr part 52. Environ Prot 804:2
Russo A, Farcomeni A (2024) A copula formulation for multivariate latent Markov models. TEST 33:1–21
Sharma SK, Mukherjee S, Choudhary N, Rai A, Ghosh A, Chatterjee A, Vijayan N, Mandal TK (2021) Seasonal variation and sources of carbonaceous species and elements in PM2.5 and PM10 over the eastern Himalaya. Environ Sci Pollut Res 28(37):51642–51656
Shu Y, Mulvey JM (2024) Dynamic factor allocation leveraging regime-switching signals. arXiv preprint arXiv:2410.14841
Shu Y, Yu C, Mulvey JM (2024) Downside risk reduction using regime-switching signals: a statistical jump model approach. J Asset Manag 25:1–15
Shu Y, Yu C, Mulvey JM (2024) Dynamic asset allocation with asset-specific regime forecasts. Ann Oper Res 346:1–34
Szepannek G (2018) Clustmixtype: User-friendly clustering of mixed-type data in R. R J., 10(2), 200
Szepannek G, Aschenbruck R, Wilhelm A (2024) Clustering large mixed-type data with ordinal variables. Adv Data Anal Classif. https://doi.org/10.1007/s11634-024-00595-5
Tortora C, Mbuga F, Bonnet Z (2023) SpectralClMixed: spectral clustering for mixed type data. (R package version 1.0.1)
Tuerhong G, Kim SB (2014) Gower distance-based multivariate control charts for a mixture of continuous and categorical variables. Expert Syst Appl 41(4):1701–1707
Viterbi A (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inf Theory 13(2):260–269
Von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17:395–416
Welch LR (2003) Hidden Markov models and the Baum-Welch algorithm. IEEE Inf Theory Soc Newsl 53(4):10–13

WHO (2021). World Health Organization global air quality guidelines: Particulate matter (PM2. 5 and PM10). Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide, p 267

Witten DM, Tibshirani R (2010) A framework for feature selection in clustering. J Am Stat Assoc 105(490):713–726

Yozgatligil C, Aslan S, Iyigun C, Batmaz I (2013) Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. Theor Appl Climatol 112:143–167

Zucchini W, MacDonald IL, Langrock R (2017) Hidden Markov models for time series: an introduction using R. CRC Press, Boca Raton