Received: 22 April 2016 | Revised: 21 December 2016 | Accepted: 22 December 2016

DOI 10.1002/env.2437

WILEY

RESEARCH ARTICLE

Model-based clustering for spatiotemporal data on air quality monitoring

A. S. M. Cheam | M. Marbac | P. D. McNicholas

Department of Mathematics and Statistics, McMaster University, Ontario, Canada

Correspondence
Paul McNicholas, Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada.
Email: paul@math.mcmaster.ca

Data extracted from air quality monitoring can require spatiotemporal clustering techniques. Of late, many clustering techniques are based on mixture models; however, there is a shortage of model-based approaches for spatiotemporal data. A new mixture to cluster spatiotemporal data, named STM, is introduced, and generic identifiability is proved. The resulting model defines each mixture component as a mixture of autoregressive polynomial regressions in which the weights consider the spatial and temporal information with logistic links. Under the maximum likelihood framework, parameter estimation is carried out via an expectation–maximization algorithm while classical information criteria can be used for model selection. The proposed model is applied to air quality monitoring data from the periphery of Paris considering one of the critical pollutants, nitrogen dioxide, at different times during the day. The STM model is implemented in the R package `SpaTimeClust`.

KEYWORDS

air quality, clustering, EM algorithm, functional data, mixture model, polynomial regression, spatiotemporal clustering

1 | INTRODUCTION

Clean air is a primary requirement for human health and well-being. According to the World Health Organization, air pollution remains a significant threat to health worldwide. Several guidelines and standards, published by World Health Organization, the US Environmental Protection Agency, and the European Union, exist for air pollutants in ambient air. Various local governments use air quality monitoring for the management and evaluation of air quality status. An air monitoring network is composed of sites that measure atmospheric pollutants (e.g., ozone, nitrogen dioxide, and particulate matter) and weather variables (e.g., quantity of rain, speed of wind, etc.) at different times. As a consequence, statistical methods are needed to evaluate spatiotemporal data obtained from numerous sites (Armanino, Roda, Ius, Casolino, & Bacigalupo, 1996; Fassò, Cameletti, & Nicolis, 2007; Lawson & Lance, 1996; Lee & Sarran, 2015). Because of the complexity of the data, it is often useful to partition these data into homogeneous subgroups by considering temporal and spatial information. The challenge is to cluster 101 days according to air quality monitoring in Paris. One observation indicates the quantity of nitrogen dioxide collected every hour during a day from nine different stations located in the periphery of Paris, that is, Ivry-sur-Seine, Neuilly-sur-Seine, Pantin, Périphérique Est, Place Basch, Porte d'Auteuil, Rue Eastmen, Rue Flacon, and Stade Lenglen. The clustering problem resembles a group multivariate time series with spatial dependencies. Note that the time series are observed only at discrete times. Many clustering methods have been developed for continuous time series and require transformation of the observed data. A drawback is that these transformations are not unique, and the choice of method is arbitrary. For example, Ignaccolo, Ghigo, and Giovenali (2008) suggest the use of a two-stage partitioning around medoids algorithm to transform discrete observations into functional data via B-splines, with a fixed number of knots.

An interesting perspective to tackle the aforementioned problem is to adopt model-based approaches that consider finite mixture models (McLachlan & Peel, 2000; McNicholas, 2016a). Thus, selecting the number of clusters is realizable by probability theory and interpreting the results is easier. Moreover, parameter estimation is often manageable through the expectation–maximization (EM)

algorithm (Dempster, Laird, & Rubin, 1977). Nowadays, finite mixtures are well established as a tool for clustering (cf. McNicholas, 2016a; 2016b). They can be used to cluster various types of data, such as continuous (Banfield & Raftery, 1993; McNicholas & Murphy, 2008; Andrews & McNicholas, 2012), categorical (Goodman, 1974); Gollini & Murphy, 2014; Marbac, Biernacki, & Vandewalle, 2015), ordinal (Jollois & Nadif, 2009; McParland & Gormley, 2013; Biernacki & Jacques, 2016), mixed (Browne & McNicholas, 2012; Hennig & Liao, 2013; Kosmidis & Karlis, 2016; McParland & Gormley, 2016), and other data types. Because of their flexibility, mixture models are employed in functional data clustering. For instance, Fernández and Green (2002) proposed modelling spatial structure by a weighted sum of Poisson distributions, while Green and Richardson (2002) and Forbes et al. (2013) used different approaches based on a hidden Markov model. On the other hand, to cluster time series, Wong and Li (2000) proposed an autoregressive version of the Gaussian mixture transition distribution model, and Nguyen, McLachlan, and Ullmann (2016) extended this idea to Laplace mixtures. Other approaches require a predefined functional basis, such as Fourier, wavelet, splines, and more. The choice of functional basis is arbitrary but impacts the results, because two different bases can lead to different partitions even if the same model is used to cluster. Jacques and Preda (2014a) give a good overview of functional data by dividing into four types of methods: raw data, filtering, adaptive, and distance based. Unfortunately, very few models consider multivariate functional data. Often, they assume independence within classes (Jacques & Preda, 2014b), which restricts them to the consideration of spatial dependencies.

Samé, Chamroukhi, Govert, and Aknin (2011) proposed a mixture model where each component follows a polynomial regression mixture in which the logistic weights depend on the time. Each observation of a time series arises independently from one of the polynomial regression models specific to the cluster to which it belongs. Therefore, unlike other approaches, the observed data do not require any transformation. The aforementioned model is only applied to a univariate temporal framework. Extending this concept to an autoregressive model, we propose spatiotemporal data clustering. The resulting method, named STM (spatiotemporal model), is a mixture model where each component is an autoregressive polynomial regression mixture in which the logistic weights depend on the spatial and temporal dimensions. The EM algorithm is used to obtain the maximum likelihood estimates of the parameters of interest. A key contribution of our work is the introduction of an autoregressive component to the model proposed by Samé et al. (2011) and the ability to model spatial dependencies for multivariate functional data.

The remainder of the paper is organized as follows. In Section 2, we provide a brief background on finite mixture models and introduce the STM model. Parameter estimation, which is achieved via the EM algorithm, is detailed in Section 3 followed by model selection in Section 4. In Section 5, numerical experiments on simulated data are performed to illustrate the consistency of the inference and the impact of the spatiotemporal dependencies. Air pollution data are analyzed in Section 6, and some concluding remarks are given in Section 7.

2 | FINITE MIXTURE MODEL FOR SPATIOTEMPORAL DATA

One may take the view that the purpose of clustering is to identify the inner structure of clustered data when no information other than the observed values are available. To achieve this in the model-based context, this section introduces the proposed autoregressive spatiotemporal model, called STM.

2.1 | Finite mixture model

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be a sample of n independent and identically distributed observations. We suppose that the density of $\boldsymbol{X}_i$ is a mixture of G components so that

$$f(\boldsymbol{x}_i|\boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g f_g(\boldsymbol{x}_i|\boldsymbol{\vartheta}_g), \tag{1}$$

where $f_g(\boldsymbol{x}_i|\boldsymbol{\vartheta}_g)$ is the gth component density of the mixture with parameters $\boldsymbol{\vartheta}_g$, $\pi_g \in (0,1]$ is the mixing proportion for the gth component with $\sum_{g=1}^{G} \pi_g = 1$, and the set of all model parameters is $\boldsymbol{\theta} = \{\pi_g, \boldsymbol{\vartheta}_g; g = 1, \ldots, G\}$. The categorical variable $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iG})$ is introduced, where

$$z_{ig} = \begin{cases} 1, \text{if observation } \boldsymbol{x}_i \text{ arises from the } g\text{th component of the mixture,} \\ 0, \text{otherwise.} \end{cases}$$

Therefore, $\boldsymbol{z}_i$ represents the membership of the mixture component and is a realization of a random variable from a multinomial distribution $\mathcal{M}_G(\pi_1, \ldots, \pi_G)$. Note that the distribution defined in Equation 1 is usually interpreted as the marginal distribution of $\boldsymbol{X}_i$ based on the distribution of the pair $(\boldsymbol{X}_i, \boldsymbol{Z}_i)$.

2.2 | Polynomial regression mixtures to model the components

Each observation $\boldsymbol{x}_i = \{x_{ijt}\}_{j=1,\ldots,J}^{t=1,\ldots,T}$ consists of $J \times T$ observations over a prespecified time grid $\mathbf{m} = (m_1, \ldots, m_T)$ and spatial grid $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_J)$, where $x_{ijt} \in \mathbb{R}$ is the realization of X_{ijt} corresponding to observation i at site j and time t. The spatial coordinates of site j are given by the bivariate vector $\mathbf{s}_j = (u_j, v_j)$. The STM supposes that each site j is independent

conditional on z_i and the dependency in time is Markov, that is, autoregressive. Hence,

$$P(\mathbf{X}_i|\boldsymbol{\vartheta}_g, Z_{ig}=1)=\prod_{j=1}^{J}\prod_{t=1}^{T}P(X_{ijt}|\boldsymbol{\vartheta}_g, Z_{ig}=1, X_{ij(t-1)}=x_{ij(t-1)}), \tag{2}$$

where $\mathbf{X}_i=\{X_{ijt}\}_{j=1,\dots,J}^{t=1,\dots,T}$. The distribution of $X_{ijt}|\boldsymbol{\vartheta}_g$, $Z_{ig}=1$, $X_{ij(t-1)}=x_{ij(t-1)}$ is a K-component mixture of Q-order polynomial regressions. Thus, the gth component density of the STM is defined as follows:

$$f_g(\boldsymbol{x}_i|\boldsymbol{\vartheta}_g)=\prod_{j=1}^{J}\prod_{t=1}^{T}\sum_{k=1}^{K}\omega_{gjtk}(\boldsymbol{\lambda}_g) \times\phi\left(x_{ijt}|[\mathbf{M}_t-\mathbf{M}_{t-1}]'\boldsymbol{\beta}_{gk}+x_{ij(t-1)},\sigma_{gk}^2\right), \tag{3}$$

where $\boldsymbol{\vartheta}_g=\{\boldsymbol{\lambda}_g, \boldsymbol{\beta}_{gk}, \sigma_{gk}; k=1,\dots,K\}$ is the set of parameters of the gth component, $\mathbf{M}_t=(1, m_t, \dots, m_t^Q)$ represents the vector of the Q-degree polynomial of $\mathbf{M}_t$, $\boldsymbol{\beta}_{gk}=(\beta_{gk0}, \dots, \beta_{gkQ})'$ is the coefficient vector of the kth regression model for the gth component, and $\phi(\cdot|\mu,\sigma^2)$ is the density of the univariate Gaussian distribution with mean μ and variance σ^2. Note that $\mathbf{M}_0$ is a null vector of size $Q+1$, and we define $x_{ij0}:=0$. The weight of this mixture, $\omega_{gjtk}(\boldsymbol{\lambda}_g)$, depends on both the spatial and time dimensions using a logistic function. Because the gth component is itself a mixture, it requires a second latent variable $\boldsymbol{y}_{ijt}=(y_{ijt1}, \dots, y_{ijtK})$ conditional on $z_{ig}=1$, with

$$y_{ijtk}=\begin{cases}1, \text{if } x_{ijt}\in k\text{th polynomial regression of the } g\text{th component},\\ 0, \text{otherwise.}\end{cases}$$

It can be viewed as a label assignment of the polynomial regression mixture that adjusts the observation x_{ijt} conditionally on $z_{ig}=1$. This categorical variable Y_{ijt} follows a multinomial distribution $\mathcal{M}_K(\omega_{jt1}(\boldsymbol{\lambda}_g), \dots, \omega_{jtK}(\boldsymbol{\lambda}_g))$ with

$$\omega_{gjtk}(\boldsymbol{\lambda}_g)=\frac{\exp(\lambda_{gk1}u_j+\lambda_{gk2}v_j+\lambda_{gk3}m_t+\lambda_{gk4})}{\sum_{\ell=1}^{K}\exp(\lambda_{g\ell1}u_j+\lambda_{g\ell2}v_j+\lambda_{g\ell3}m_t+\lambda_{g\ell4})} \quad \forall\,(g,j,t,k), \tag{4}$$

where $\boldsymbol{\lambda}_g=(\lambda_{g11}, \dots, \lambda_{gK4})$ is the parameter of the logistic function. To ensure model identifiability, we impose that $\lambda_{g1h}=0$ for $h=1,\dots,4$. The proof of model identifiability is presented in the Appendix. Note that the elements of the observation $\boldsymbol{x}_i$ do not necessarily arise from the same polynomial regression.

The distribution defined by Equations 1 and 3 implies the following three steps of the generative model, see Figure 1:

1. Sampling of the component membership:

$$\boldsymbol{Z}_i\sim\mathcal{M}_G(\pi_1, \dots, \pi_G).$$

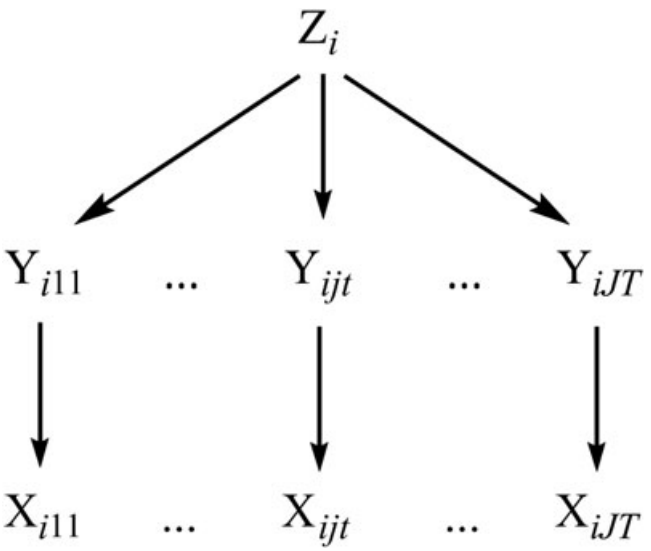

FIGURE 1 Diagram describing the three steps of the generative model

2. Conditional sampling of the regression membership for all (j, t):

$$\boldsymbol{Y}_{ijt}|Z_{ig}=1\sim\mathcal{M}_K(\omega_{jt1}(\boldsymbol{\lambda}_g), \dots, \omega_{jtK}(\boldsymbol{\lambda}_g)).$$

3. Conditional sampling of the observation for all (j, t):

$$X_{ijt}|Z_{ig}=1, Y_{ijtk}=1\sim\mathcal{N}\left([\mathbf{M}_t-\mathbf{M}_{t-1}]'\boldsymbol{\beta}_{gk}+x_{ij(t-1)},\sigma_{gk}^2\right).$$

Remark 1. To easily summarize and visualize each component, we can plot an average curve of each component thanks to

$$\mathbb{E}\left[X_{ijt}|Z_{ig}=1\right]=\sum_{k=1}^{K}\omega_{gjtk}(\boldsymbol{\lambda}_g)\mathbf{M}_t'\boldsymbol{\beta}_{gk}, \tag{5}$$

because

$$\mathbb{E}\left[X_{ijt}|Z_{ig}=1, Y_{ijtk}=1\right]=\mathbf{M}_t'\boldsymbol{\beta}_{gk}$$

and

$$\mathbb{E}\left[X_{ijt}|Z_{ig}=1, Y_{ijtk}=1, X_{ij(t-1)}=x_{ij(t-1)}\right]=[\mathbf{M}_t-\mathbf{M}_{t-1}]'\boldsymbol{\beta}_{gk}+x_{ij(t-1)}.$$

Remark 2. Note that Y_{ijt} is used to model spatial and temporal dependencies, though it can also be employed to interpret components. Conditionally on the component, the posterior distribution of Y_{ijt} defines the probability that the observation x_{ijt} arises from the kth regression. Thus, a segmentation can be considered because, conditional on the component, we can obtain the regression with the greatest probability at time m_t for site s_j. Figure 2 presents the average curve and the segmentation obtained for the site Neuilly-sur-Seine by considering the three-component mixture model. Thanks to the logistic

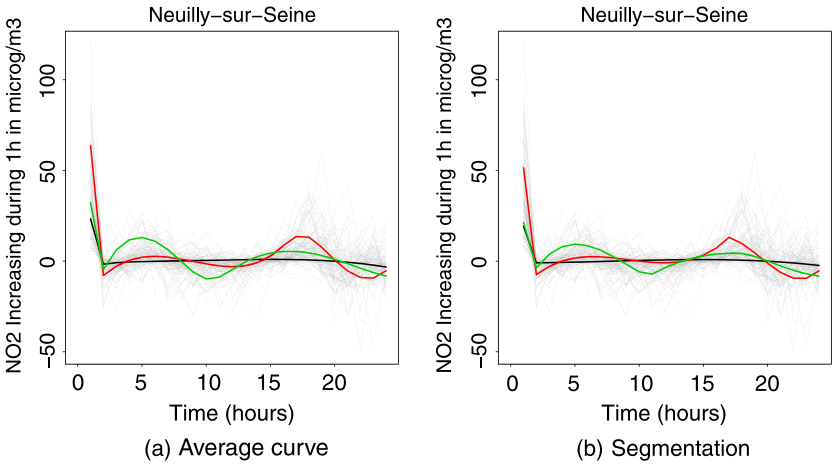

FIGURE 2 Summary of the three-component mixture model for the site Neuilly-sur-Seine. Colours indicate component 1 (black), component 2 (red), component 3 (green), and the data (gray)

weights, for a fixed time and location, each component is mainly described by a single regression. As a consequence, the average curve and the segmentation curve are similar.

3 | MAXIMUM LIKELIHOOD INFERENCE OF THE PARAMETERS

Given a sample of n independent observations from a mixture defined in Equation 1, the observed log-likelihood function is

$$\log\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{n} \log\left(\sum_{g=1}^{G} \pi_g f_g(\boldsymbol{x}_i|\boldsymbol{\vartheta}_g)\right). \tag{6}$$

The EM algorithm is an appealing tool to obtain maximum likelihood estimates (MLE) for mixture models. Recall that the STM model, see Equations 1 and 3, implies two latent variables: the component membership z_i and the label assignment for the polynomial regression $\boldsymbol{y}_{ijt}$. The complete-data log-likelihood is given by

$$\begin{aligned}&\log\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}_1, \dots \mathbf{x}_n, \mathbf{y}_1, \dots, \mathbf{y}_n, \mathbf{z}_1, \dots, \mathbf{z}_n)\\ &= \log\mathcal{L}_1(\boldsymbol{\pi}) + \log\mathcal{L}_2(\lambda) + \log\mathcal{L}_3(\boldsymbol{\beta}, \boldsymbol{\sigma}),\end{aligned} \tag{7}$$

where

$$\log\mathcal{L}_1(\boldsymbol{\pi}) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\log\pi_g,$$

$$\log\mathcal{L}_2(\lambda) = \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{ig}y_{ijtk}\log\omega_{gjtk}(\lambda_g),$$

$$\log\mathcal{L}_3(\boldsymbol{\beta}, \boldsymbol{\sigma}) = \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{ig}y_{ijtk} \times \log\left\{\phi\left(x_{ijt}|[\mathbf{M}_t - \mathbf{M}_{t-1}]'\boldsymbol{\beta}_{gk} + x_{ij(t-1)}, \sigma^2_{gk}\right)\right\}.$$

Let $\boldsymbol{\theta}^{(r)}$ represents the value of the model parameters $\boldsymbol{\theta}$ at iteration r of the EM algorithm. Starting with an initial value of the parameters $\boldsymbol{\theta}^{(0)}$, the algorithm iterates between the expectation (E-step) and the maximization (M-step) until some convergence criterion is met.

Expectation step (E-step):

Consider the expected value of the complete-data log-likelihood, conditional on the current parameter estimates, that is, $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(r)}) := \mathbb{E}\left[\log\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}_1, \dots \mathbf{x}_n, \mathbf{y}_1, \dots, \mathbf{y}_n, \mathbf{z}_1, \dots, \mathbf{z}_n)\right]$. The following expectations are needed in the E-step: $m_{ig}(\boldsymbol{\vartheta}^{(r)}) := \mathbb{E}[Z_{ig} = 1|\boldsymbol{x}_i, \boldsymbol{\vartheta}^{(r)}]$ and $s_{igjtk}(\boldsymbol{\vartheta}^{(r)}) := \mathbb{E}[Y_{ijtk} = 1|\boldsymbol{x}_i, \boldsymbol{\vartheta}^{(r)}, Z_{ig} = 1]$, where

$$m_{ig}(\boldsymbol{\vartheta}^{(r)}) = \frac{\pi_g^{(r)} f_g\left(\boldsymbol{x}_i|\boldsymbol{\vartheta}_g^{(r)}\right)}{\sum_{\ell=1}^{G}\pi_\ell^{(r)} f_\ell\left(\boldsymbol{x}_i|\boldsymbol{\vartheta}_\ell^{(r)}\right)}, \tag{8}$$

$$s_{igjtk}(\boldsymbol{\vartheta}^{(r)}) = \frac{\omega_{jtk}^{(r)}\phi\left(x_{ijt}|[\mathbf{M}_t - \mathbf{M}_{t-1}]'\boldsymbol{\beta}_{gk}^{(r)} + x_{ij(t-1)}, \sigma_{gk}^{(r)2}\right)}{\sum_{\ell=1}^{K}\omega_{jt\ell}^{(r)}\phi\left(x_{ijt}|[\mathbf{M}_t - \mathbf{M}_{t-1}]'\boldsymbol{\beta}_{g\ell}^{(r)} + x_{ij(t-1)}, \sigma_{g\ell}^{(r)2}\right)}. \tag{9}$$

Maximization step (M-step):

We maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$. Note that the M-step is achieved by optimizing three groups of parameters independently; see Equation 7.

1. The **mixture weights** are updated as follows:

$$\pi_g^{(r+1)} = \frac{n_g^{(r)}}{n}, \tag{10}$$

where $n_g^{(r)} = \sum_{i=1}^{n} m_{ig}(\boldsymbol{\vartheta}^{(r)})$.

2. The **parameter of regression weights** for the gth component implies finding the MLE of a weighted logistic regression. Thus, the update is defined by

$$\boldsymbol{\lambda}_g^{(r+1)} = \arg\max_{\boldsymbol{\lambda}_g} \sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{k=1}^{K} r_{igjtk}(\boldsymbol{\vartheta}^{(r)}) \log \omega_{gjtk}(\boldsymbol{\lambda}_g), \tag{11}$$

where $r_{igjtk}(\boldsymbol{\vartheta}^{(r)}) = m_{ig}(\boldsymbol{\vartheta}^{(r)}) s_{igjtk}(\boldsymbol{\vartheta}^{(r)})$. From Equation 4, recall that $\lambda_{g1h} = 0$ for $h = 1, \ldots, 4$ and $g = 1, \ldots, G$. The assessment of $\boldsymbol{\lambda}_g^{(r+1)}$ is performed by the Newton-Raphson algorithm (cf. Nocedal & Wright, 2006).

3. The **regression parameters** entail the computation of the regression coefficient $\boldsymbol{\beta}_{gk}$ and the variance σ^2_{gk}. The updated regression coefficient is defined as

$$\boldsymbol{\beta}_{gk}^{(r+1)} = \left[\mathbf{A}'\mathbf{D}_{gk}^{(r)}\mathbf{A}\right]^{-1}\left[\mathbf{A}'\mathbf{C}_{gk}^{(r)}\right], \tag{12}$$

where $\mathbf{A}_t = [\mathbf{M}_t - \mathbf{M}_{t-1}]$ is the tth row of a $T \times (Q+1)$ matrix $\mathbf{A}$, $\mathbf{D}_{gk}^{(r)}$ is a $T \times T$ diagonal matrix where the tth diagonal element is $\sum_{i=1}^{n}\sum_{j=1}^{J} r_{igjtk}(\boldsymbol{\vartheta}^{(r)})$, and $\mathbf{C}_{gk}^{(r)} = (C_{gk1}^{(r)}, \ldots, C_{gkT}^{(r)})'$ with $C_{gkt}^{(r)} = \sum_{i=1}^{n}\sum_{j=1}^{J} r_{igjtk}(\boldsymbol{\vartheta}^{(r)})(x_{ijt} - x_{ij(t-1)})$. The updated variance associated with the kth regression of component g is written as

$$(\sigma^2_{gk})^{(r+1)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{t=1}^{T} r_{igjtk}(\boldsymbol{\vartheta}^{(r)})\left(x_{ijt} - x_{ij(t-1)} - \mathbf{A}_t'\boldsymbol{\beta}_{gk}^{(r+1)}\right)^2}{\text{trace}(\mathbf{D}_{gk}^{(r)})}. \tag{13}$$

The EM algorithm converges toward a local optimum that is only determined by the starting point $\boldsymbol{\vartheta}^{(0)}$. Many initializations are required to achieve the maximum likelihood inference. Here, we adopt the xem-EM strategy (Biernacki, Celeux, & Govaert, 2003). The strategy is to execute several short runs of EM from random starting points, for a few iterations, followed by a long run of EM from the solution maximizing the log-likelihood.

4 | MODEL SELECTION

Competing models

Let $\mathcal{M}$ denote the set of competing candidate models. According to the distribution from Equations 1 and 3, each model $\mathscr{M} \in \mathcal{M}$ is defined by three elements: the number of components, G; the number of regressions per component, K; and the degree of polynomial regressions, Q. Hence, a model $\mathscr{M}$ is given by

$$\mathscr{M} = (G, K, Q) \qquad \text{with } G, K, Q \in \mathbb{N}^*. \tag{14}$$

Thus, $\mathcal{M}$ is obtained by fixing the maximum number of each element listed above. Therefore, assuming $G_{\min} = K_{\min} = Q_{\min} = 1$, the number of candidate models is

$$\text{card}(\mathcal{M}) = G_{\max} \times K_{\max} \times Q_{\max}. \tag{15}$$

Model selection is carried out using an information criterion whose value is computed for each model in $\mathcal{M}$. Then, the "best model" is selected according to the information criterion. For this study, we employed the Bayesian information criterion (BIC; Schwarz, 1978), but one may choose to use the Akaike information criterion (Akaike, 1974) or the integrated completed likelihood (Biernacki, Celeux, & Govaert, 2000) as an alternative.

The BIC criterion

In a Bayesian framework, the "best model" is characterized by the model $\widehat{\mathscr{M}}$ with the highest posterior probability among $\mathcal{M}$. By assuming equal prior distributions for each $\mathscr{M}$, $\widehat{\mathscr{M}}$ maximizes the integrated likelihood. However, this integral does not have a closed-form for mixture models. To circumvent this issue, many methods propose to approximate its value. The most appealing one consists of using the BIC, which approximates the logarithm of the integrated likelihood using a Laplace approximation. The BIC is defined as

$$\text{BIC}(\mathscr{M}) = \log\, \mathcal{L}(\mathscr{M}, \hat{\boldsymbol{\vartheta}}_{\mathscr{M}}) - \frac{\nu_{\mathscr{M}}}{2}\log\, n, \tag{16}$$

where $\log\mathcal{L}(\mathscr{M}, \hat{\boldsymbol{\vartheta}}_{\mathscr{M}})$ is the maximized log-likelihood, $\hat{\boldsymbol{\vartheta}}_{\mathscr{M}}$ is the MLE of $\boldsymbol{\vartheta}_{\mathscr{M}}$, and $\nu_{\mathscr{M}}$ is the number of free parameters in the model $\mathscr{M}$.

5 | NUMERICAL EXPERIMENTS ON SIMULATED DATA

In this section, we conduct simulation studies to illustrate the consistency of inference of the model and the importance of modelling the spatiotemporal dependencies.

The simulated data

Observations are described by the values measured on $J = 25$ sites at $T = 10$ moments. For each scenario, 20 replicates are generated by the STM model with two components defined by two linear regressions each (i.e., $G = 2$,$K = 2$, $Q = 1$). We consider unbalanced mixing proportions $\pi_1 = 1/3$ and $\pi_2 = 2/3$, and fix $G = 2$; thus, the worst misclassification rate is 0.33. The other parameters are

$$\boldsymbol{\beta}_{g1} = (0, a), \boldsymbol{\beta}_{g2} = (a, (-1)^g a),$$
$$\boldsymbol{\lambda}_{g2} = \begin{bmatrix} \textit{abscissa} & \textit{ordinate} & \textit{time} & \textit{constant} \\ 2 & -2 & -1 & 4 \end{bmatrix} \text{ and } \sigma^2_{gk} = 1$$

where a permits the consideration of three different levels of overlap (5, 10, and 15%). Results yielded by the STM model are compared with those provided by the `mclust` software (Fraley & Raftery, 2002; Fraley, Raftery, & Scrucca, 2015) for R.

Consistency of estimates

As expected, the estimates are consistent because they rapidly converge toward zero as n increases; see Table 1. For example, the mean distance of $|\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}|$, at a 5%

TABLE 1 Mean of the distance between the true parameters and the estimates: proportions ($|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}|$), logistic weight parameters ($|\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}|$), regression parameters ($|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|$), and standard deviation parameters ($|\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}|$)

Size	$\lvert\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\rvert$			$\lvert\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\rvert$			$\lvert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rvert$			$\lvert\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\rvert$		
(n)	5%	10%	15%	5%	10%	15%	5%	10%	15%	5%	10%	15%
50	0.01	0.01	0.05	1.63	3.57	4.65	0.09	0.10	0.24	0.01	0.02	0.04
100	0.01	0.02	0.06	0.86	2.36	4.61	0.12	0.09	0.27	0.01	0.01	0.03
200	0.00	0.01	0.02	0.60	2.43	4.30	0.05	0.10	0.24	0	0.01	0.02
400	0.00	0.00	0.01	0.42	2.05	3.62	0.02	0.03	0.08	0.00	0.00	0.00

TABLE 2 Frequencies of selecting (G, K, and Q) with the STM model where the true value is (2, 2, and 1)

Size	Overlap	G			K			Q		
(n)	(%)	1	2	3	1	2	3	0	1	2
50	5	0	20	0	0	20	0	0	19	1
	10	9	11	0	0	20	0	0	17	3
	15	18	2	0	6	14	0	0	14	6
100	5	0	20	0	0	20	0	0	19	1
	10	6	14	0	0	20	0	0	19	1
	15	17	3	0	0	20	0	0	17	3
200	5	0	20	0	0	20	0	0	20	0
	10	0	20	0	0	20	0	0	20	0
	15	11	9	0	0	20	0	0	19	1
400	5	0	20	0	0	20	0	0	20	0
	10	0	20	0	0	20	0	0	20	0
	15	0	20	0	0	20	0	0	20	0

TABLE 3 Mean misclassification rates for each method applied to the simulated data, with class overlaps of 5, 10, and 15%

Size	5%		10%		15%	
(n)	MCLUST	STM	MCLUST	STM	MCLUST	STM
50	0.17	0.08	0.29	0.16	0.34	0.28
100	0.10	0.06	0.27	0.14	0.33	0.23
200	0.09	0.07	0.22	0.13	0.34	0.21
400	0.09	0.05	0.21	0.11	0.33	0.17

misclassification rate, drops rapidly from 1.63 at $n = 50$ to 0.42 at $n = 400$.

Consistency of model

The following step is used to verify the consistency of model selection because, in practice, the model is most likely unknown. Therefore, for each replicate, all possible models, where $G_{\max} = K_{\max} = 3$ and $Q_{\max} = 2$, are fitted to compute the BIC criterion. The winner is the model that maximizes this criterion. Table 2 presents the statistics of the winning models. We observe, as n increases, that the STM model is able to detect the true grouping, even at 15% class overlap. As suspected, G is underestimated when n is small at an overlap of 15% but, as n grows, the probability of selecting the true number of components converges to one. Note that the true number of regressions and the degree of polynomial regression are well estimated by STM in all scenarios.

Importance of dependencies

Table 3 presents the mean of the misclassification rates obtained by two competing models on the 20 replicates. For MCLUST, we used the classic dependency, which is not necessarily specific to spatiotemporal data. This negligence results in greater misclassification rates and is unlike the STM, which converges faster. This shows the importance of incorporating the spatiotemporal dependencies in the model.

6 | APPLICATION TO AIR POLLUTION DATA

The air quality monitoring network in Paris and the Île-de-France region (the capital city region) is operated by Airparif-Design Clair et Net (2010), a nonprofit organization accredited by the Ministry of Environment. Hourly quantities of nitrogen dioxide (NO_2), measured at nine Airparif stations distributed around the periphery of Paris for 101 days in 2014, are used to illustrate the proposed model. We extracted data for $J = 9$ stations that describe the quantity of NO_2 (in $\mu g/m^3$) for $T = 24$ hr per day for $n = 101$ days. The nine stations include Ivry-sur-Seine, Neuilly-sur-Seine, Pantin (RN2), Périphérique Est (abbreviated as Periph-Est), place Basch (Paris XVIeme), Porte d'Auteuil, Rue Eastmen (Paris XIIIeme), rue Flacon (Paris XVIIIeme), and Stade Lenglen. Additional variables (e.g., percentage of weekend days, days between April 1 and October 1, holidays in July

TABLE 4 Selected model and information criterion for the air pollution dataset

	G	K	Q	No. of free parameters	BIC
MCLUST (VEI)	6	–	–	1522	– 93241
TM	4	4	4	123	– 79175
STM	3	4	4	110	– 79006

TABLE 5 Confusion table of ARI values computed between the three resulting partitions for the air pollution data

	MCLUST	TM	STM
MCLUST	1.00	0.13	0.25
TM	–	1.00	0.78
STM	–	–	1.00

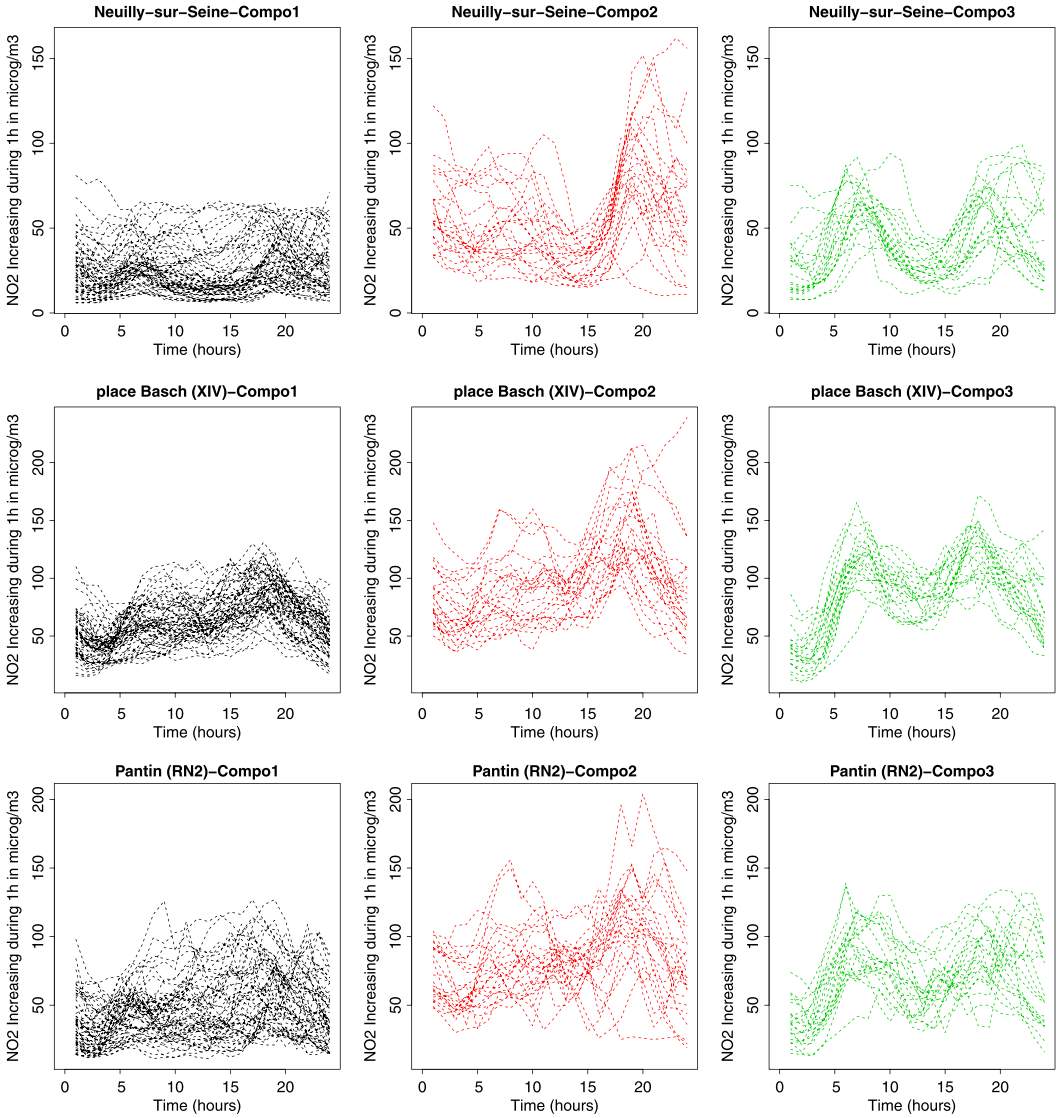

FIGURE 3 Hourly quantity of NO_2 ($\mu g/m^3$) measured at three different stations in the periphery of Paris where each class member represents days with similar curvature or characteristic for 101 days observed

TABLE 6 Justification of classes using other explicative variables for air pollution

Variables	Component (g)		
	1	2	3
Weekend (%)	81.0	69.5	10.0
Total rain (cm)	0.154	0.147	0.111
Wind_{avg} (km/h)	4.638	2.652	3.400
Days between Apr 1 - Oct 1(%)	56.9	26.1	40.0
Spring holidays in July & Aug (%)	20.7	0.0	0.15
Wind_{max} (km/h)	11.155	11.652	13.050

and August, total rain, and average and maximum wind speed) are also examined to support and interpret our clustering results.

Software conditions

The air pollution data analysis is performed using the STM model via the `SpaTimeClust` package (Marbac, Cheam, & McNicholas, 2016) for R (R Core Team, 2016) with the default options. Note that this package also includes the data. For this application, we defined $G_{\max} = 6$, $K_{\max} = 6$ and $Q_{\max} = 5$. The spatial coordinates of each site j are determined by the geographical coordinates of the region's capital. Note that we also fit the model TM defined by the proposed model, STM, without the spatial dependencies, that is, $\lambda_{gk1} = \lambda_{gk2} = 0\ \forall(g, k)$ in Equation 4. This approach can be viewed as an extension of ClustSeg, proposed by Samé et al. (2011), by incorporating an autoregressive regression and allowing multivariate functional data. Thus, the TM model can also be regarded as an approach to clustering multivariate functional data by assuming independence within classes between different functions.

Model comparison

Interestingly, the selected STM outperforms the TM and MCLUST (VEI) models in terms of the BIC, that is, $\text{BIC}_{\text{STM}} > \text{BIC}_{\text{TM}} > \text{BIC}_{\text{MCLUST}}$ (see Table 4). Moreover, the STM model produces a straightforward interpretation

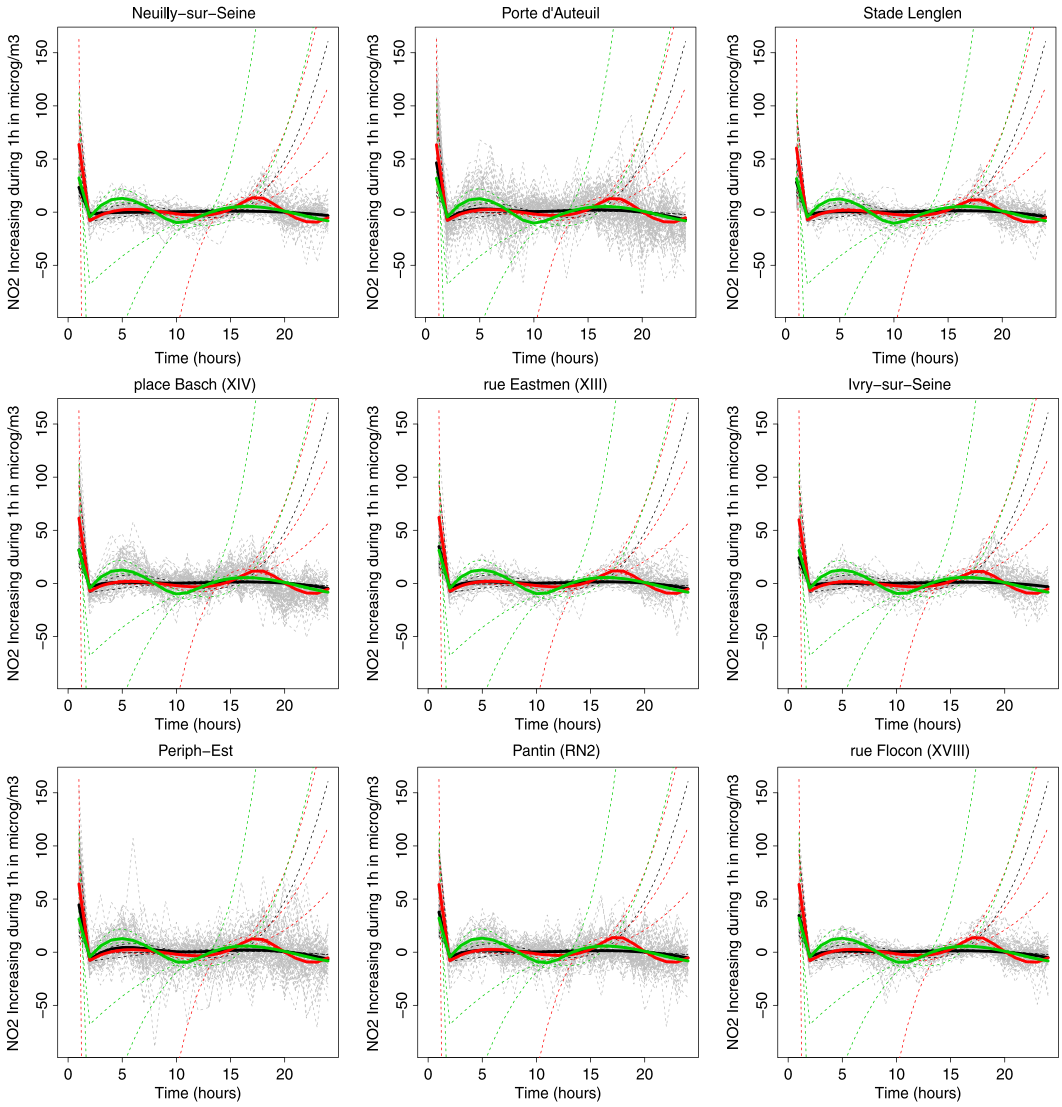

FIGURE 4 Hourly averaged curves (in bold) of the increasing quantity of NO_2 ($\mu g/m^3$), i.e., $x_{ijt} - x_{ij(t-1)}$, at nine different stations in the periphery of Paris under different components (in black for Component 1, red for Component 2, and green for Component 3). The dashed line represents each label assignment of the polynomial regression depending on the class membership. The grey curves represents the observations, that is, 101 days. The station Périphérique Est is abbreviated as Periph-Est

because it considers fewer classes and requires fewer parameters than the other models. As suspected, there is a major disparity in the partitioning of classes between MCLUST and STM because the adjusted Rand index (ARI; Hubert & Arabie, 1985) equals 0.25; see Table 5. This statement is not surprising because MCLUST does not consider the spatial and temporal dimensions unlike STM. Moreover, the distinct partitioning of classes between TM and STM (ARI = 0.78) highlights the importance of modelling the spatial dependencies.

Best model interpretation

First, recall the aim is to group days presenting similarity in the curvature or characteristic. The selected STM model produces three components. Figure 3 illustrates the behaviour of each component for three sites (Neuilly-sur-Seine, place Basch, and Pantin). Hence, the 101 days are divided within these three components containing 58, 23, and 20 observations, respectively. The first component contains principally weekends and vacation periods, but also days with the most rain and the highest average wind speed of 4.638 km/h (Table 6). Combining all of these factors, we expect a low and constant quantity of NO_2, as shown in Figure 3. Conversely, the third component is composed of weekdays with strong wind, that is, 13.05 km/h, but little rain.

Figure 4 shows the averaged curves of the model for each component obtained by using Equation 5 for $K = 4$. For the third component (green bold line), we notice the curve have two peaks possibly caused by rush hours. This finding remains consistent with Figure 3. This implies that the STM model is able to capture the curvatures of the data. From Table 7, NO_2 reaches its maximum at 8 am (see $\arg\max_{t=1,\ldots,24} \mathbb{E}_{gjt}$), for three sites, unlike other components. The second component includes days with the least vacations as well as many weekend days. Adding the slow average

TABLE 7 Class descriptor with respect to the time, where $\mathbb{E}_{gjt} = \mathbb{E}\left[x_{ij1}|Z_{ig} = 1\right]$

		Component		
		$g = 1$	$g = 2$	$g = 3$
Neuilly-sur-Seine	$\mathbb{E}_{gj1}$	23.32	63.65	32.14
	$\mathbb{E}_{gj24}$	19.83	67.87	51.91
	$\max_{t=1,\dots,24}\mathbb{E}_{gjt}$	27.27	96.93	76.05
	$\arg\max_{t=1,\dots,24}\mathbb{E}_{gjt}$	19	20	8
Ivry-sur-Seine	$\mathbb{E}_{gj1}$	24.16	59.93	32.99
	$\mathbb{E}_{gj24}$	24.33	51.35	51.33
	$\max_{t=1,\dots,24}\mathbb{E}_{gjt}$	30.76	80.26	74.71
	$\arg\max_{t=1,\dots,24}\mathbb{E}_{gjt}$	20	20	8
Pantin	$\mathbb{E}_{gj1}$	37.72	63.62	32.10
	$\mathbb{E}_{gj24}$	38.17	69.11	54.56
	$\max_{t=1,\dots,24}\mathbb{E}_{gjt}$	49.86	98.06	79.34
	$\arg\max_{t=1,\dots,24}\mathbb{E}_{gjt}$	20	20	8

wind speed (2.652 km/h) implies the quantity of NO_2 accumulates throughout the day. This explains the continuously high level of NO_2 over a period of 24 hr, with a peak at 8 pm (Table 7).

Figure 5 illustrates the averaged curves of $x_{ijt} - x_{ij(t-1)}$ per hour for nine sites. We observe, especially for the first component, that the nine curves are different. For instance, we notice one particular station has a higher quantity of NO_2. Thus, the spatial information is modelled with respect to each component allowing to obtain nine curves in each component instead of only one.

7 | CONCLUSION

A new model for spatiotemporal data has been introduced. The proposed model, STM, is a mixture model where each component is an autoregressive polynomial regression mixture with logistic weights that depend on the spatial and temporal dimensions. In the simulation studies, STM showed consistency of estimates, and the importance of spatial dependencies were illustrated. The STM model was applied to air pollution data and gave excellent results when compared to MCLUST. Moreover, interpretation of the spatiotemporal clustering is straightforward.

The superior performance of the STM model when compared to MCLUST on the simulated data can be explained, in part, by the modelling of spatial dependencies. Another appealing aspect of the STM model is the ability to work directly with the data without any transformations or selection of functional basis, such as Fourier. Thus, the computation of the information criteria is achievable. An interesting avenue for future work would be to generalize the proposed model by replacing the polynomial regressions with different functions and then using the information criteria to determine which function is best with respect to the data. Hence, by determining the model, we could simultaneously choose the most relevant functional basis per component. Because observations are often influenced by auxiliary variables, the introduction of covariates such as quantity of rain, wind speed, and others, in the model would be very useful and instructive. These

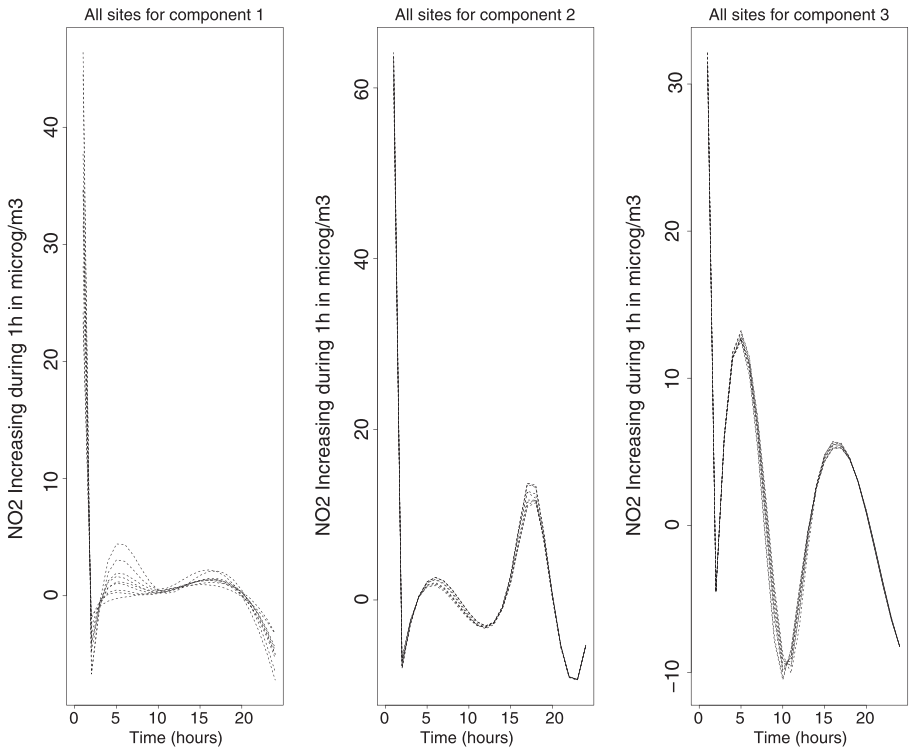

FIGURE 5 Hourly averaged curves of the increasing quantity of NO_2 ($\mu g/m^3$), i.e., $x_{ijt} - x_{ij(t-1)}$, for all nine stations in the periphery of Paris under their respective class

covariates may, in turn, impact the weights of the polynomial regression mixtures.

REFERENCES

Airparif-Design Clair et Net (2010). Airparif, association de surveillance de la qualité de l'air. Retrieved from: http://www.airparif.asso.fr/. Accessed on March 2016.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Andrews, J. L., & McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis via mixture of multivariate t-distributions. *Statistics and Computing*, *22*(5), 1021–1029.

Armanino, C., Roda, A., Ius, A., Casolino, M. C., & Bacigalupo, M. A. (1996). Pattern recognition of particulate-bound pollutants sampled during a long term urban air monitoring scheme. *Environmetrics*, *7*(6), 537–550.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*(3), 803–821.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–725.

Biernacki, C., Celeux, G., & Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, *41*(3), 561–575.

Biernacki, C., & Jacques, J. (2016). Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. *Statistics and Computing*, *26*(5), 929–943.

Browne, R. P., & McNicholas, P. D. (2012). Model-based clustering and classification of data with mixed type. *Journal of Statistical Planning and Inference*, *142*(11), 2976–2984.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38.

Fassò, A., Cameletti, M., & Nicolis, O. (2007). Air quality monitoring using heterogeneous networks. *Environmetrics*, *18*(3), 245–264.

Fernández, C., & Green, P. J. (2002). Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*, 805–826.

Forbes, F., Charras-Garrido, M., Azizi, L., Doyle, S., Abrial, D., et al. (2013). Spatial risk mapping for rare disease with hidden markov fields and variational em. *The Annals of Applied Statistics*, *7*(2), 1192–1216.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*(458), 611–631.

Fraley, C., Raftery, A. E., & Scrucca (2015). mclust: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. R package version 5.1.

Gollini, I., & Murphy, T. B. (2014). Mixture of latent trait analyzers for model-based clustering of categorical data. *Statistics and Computing*, *24*(4), 569–588.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231.

Green, P. J., & Richardson, S. (2002). Hidden markov models and disease mapping. *Journal of the American statistical association*, *97*(460), 1055–1070.

Hennig, C., & Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *62*(3), 309–369.

Horn, R. A., & Johnson, C. R. (1991). *Topics in Matrix Analysis*. New York: Cambridge University Press.

Hubert, L., & Arabie P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Ignaccolo, R., Ghigo, S., & Giovenali, E. (2008). Analysis of air quality monitoring networks by functional clustering. *Environmetrics*, *19*(7), 672–686.

Jacques, J., & Preda, C. (2014a). Functional data clustering: a survey. *Advances in Data Analysis Classification*, *8*(3), 231–255.

Jacques, J., & Preda, C. (2014b). Model-based clustering for multivariate functional data. *Computational Statistics and Data Analysis*, *71*, 92–106.

Jollois, F. X., & Nadif, M. (2009). Classification de données ordinales: modèles et algorithmes. In *41ème Journées de Statistique*, Bordeaux.

Kosmidis, I., & Karlis, D. (2016). Model-based clustering using copulas with applications. *Statistics and Computing*, *26*(5), 1079–1099.

Lawson, A. B., & Lance, W. A. (1996). A review of point pattern methods for spatial modelling of events around sources of pollution. *Environmetrics*, *7*(5), 471–487.

Lee, D., & Sarran, C. (2015). Controlling for unmeasured confounding and spatial misalignment in long-term air pollution and health studies. *Environmetrics*, *26*(7), 477–487.

Marbac, M., Biernacki, C., & Vandewalle, V. (2015). Model-based clustering for conditionally correlated categorical data. *Journal of Classification*, *32*(2), 145–175.

Marbac, M., Cheam, A. S. M., & McNicholas, P. D. (2016). SpaTimeClust: Model-Based Clustering of Spatio-Temporal Data. R package version 1.0.

McLachlan, G., & Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

McNicholas, P. D. (2016a). *Mixture Model-Based Classification*. Boca Raton: Chapman & Hall/CRC Press.

McNicholas, P. D. (2016b). Model-based clustering. *Journal of Classification*, *33*(3), 331–373.

McNicholas, P. D., & Murphy, T. B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, *18*(3), 285–296.

McParland, D., & Gormley, I. C. (2013). Clustering Ordinal Data via Latent Variable Models. In *Algorithms from and for Nature and Life*. (pp. 127–135). Springer International Publishing.

McParland, D., & Gormley, I. C. (2016). Model based clustering for mixed data: clustmd. *Advances in Data Analysis and Classification*, *10*(2), 155–169.

Nguyen, H. D., McLachlan, G. J., & Ullmann, J. F. P. (2016). Laplace mixture autoregressive models. *Statistics and Probability Letters*, *110*, 18–24.

Nocedal, J., & Wright, S. (2006). *Numerical Optimization* (2nd ed.)., Springer series in operations research and financial engineering. New York, NY: Springer.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Samé, A., Chamroukhi, F., Govert, G., & Aknin, P. (2011). Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis Classification*, *5*, 301–321.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Silvapulle, M. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *43*(3), 310–313.

Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, *34*(4), 1265–1269.

Wong, C. S., & Li, W. K. (2000). On a mixture autoregressive model. *Journal of the Royal Statistical Society: Series B (Methodological)*, *62*(1), 95–115.

How to cite this article: Cheam ASM, Marbac M, McNicholas PD. Model-based clustering for spatio temporal data on air quality monitoring, *Environmetrics*. 2017;e2437. https://doi.org/10.1002/env.2437

APPENDIX

Identifiability of a model means that a given mixture is uniquely characterized in the sense that two distinct sets of the

parameters (e.g., $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$) cannot yield the same distribution. Although it is well known that mixture models are not identifiable due to the possibility of obtaining the same distribution by rearranging class labels, this issue is trivially soluble here, for example, by setting $\pi_1 > \pi_2 > \cdots > \pi_G$. Setting aside the issue of class labels, we establish that parameters of the STM model are in fact generically identifiable. This implies that the set of points for which identifiability does not hold has measure zero. In other words, any observed data set has probability one of being drawn from a distribution with identifiable parameters. The STM model is generically identifiable, setting apart the issue of class labels, if

1. $T > Q$;
2. we can construct a matrix of size $(Q + 1) \times 2$ composed of spatial coordinates of the observed sites where the row rank equals to $Q + 1$.

Suppose that $\boldsymbol{\Theta}$ is the parameter space for the model. There exists a subspace $A \subset \Theta$ of measure zero such that $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta \setminus A$. Thus, we have

$$\forall \quad \boldsymbol{x}, \quad f(\boldsymbol{x}|\boldsymbol{\theta}) = f(\boldsymbol{x}|\tilde{\boldsymbol{\theta}}) \;\Rightarrow\; \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}. \tag{A1}$$

An equivalent condition to prove Appendix (A1) is to verify that

$$\forall \quad \boldsymbol{x}, \quad \forall \quad (j,t) \quad f_{jt}(\boldsymbol{x}_{jt}|\boldsymbol{\theta}) = f_{jt}(\boldsymbol{x}_{jt}|\tilde{\boldsymbol{\theta}}) \;\Rightarrow\; \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}, \tag{A2}$$

where $f_{jt}(\boldsymbol{x}_{jt}|\boldsymbol{\theta})$ is the marginal probability of $\boldsymbol{x}_{jt}$ obtained by marginalizing over $f(\boldsymbol{x}|\boldsymbol{\theta})$. Note that if $U_1 \sim \mathcal{N}(\mu_1, \sigma^2)$ and $U_2|U_1 \sim \mathcal{N}(\mu_2 - \mu_1 + u_1, \sigma^2)$, then $U_2 \sim \mathcal{N}(\mu_2, 2\sigma^2)$. From this result, the marginal probability density function of x_{jt} can be written as

$$f_{jt}(\boldsymbol{x}_{jt}|\boldsymbol{\theta}) = \sum_{g=1}^{G}\sum_{k=1}^{K} \pi_g \omega_{gjtk}(\lambda_g)\varphi(x_{jt}|\mathbf{M}_t'\boldsymbol{\beta}_{gk}; 2^{t-1}\sigma_{gk}^2). \tag{A3}$$

We observe that Equation (A3) is a univariate Gaussian mixture with $G \times K$ components. Thus, from Teicher (1963), the identifiability of Gaussian mixtures implies that $\forall(g, k, j, t)$

$$\begin{cases} \sigma_{gk}^2 = \tilde{\sigma}_{gk}^2, \\ \mathbf{M}_t'\boldsymbol{\beta}_{gk} = \mathbf{M}_t'\tilde{\boldsymbol{\beta}}_{gk}, \\ \pi_g\omega_{gjtk}(\boldsymbol{\lambda}_g) = \tilde{\pi}_g\omega_{gjtk}(\tilde{\boldsymbol{\lambda}}_g). \end{cases} \tag{A4}$$

From Equation (A4), the identifiability of the parameters σ_{gk}^2 is proven because $\sigma_{gk}^2 = \tilde{\sigma}_{gk}^2$. The next step is to show the identifiability of the parameters $(\boldsymbol{\beta}_{gk}, \boldsymbol{\lambda}_g, \boldsymbol{\pi}_g)$.

Identifiability of $\boldsymbol{\beta}_{\mathrm{gk}}$

From Equation (A4), we have that $\forall(g,k)$, $\mathbf{M}'\boldsymbol{\beta}_{gk} = \mathbf{M}'\tilde{\boldsymbol{\beta}}_{gk}$, where $\mathbf{M} = (\mathbf{M}_1, \ldots, \mathbf{M}_T)$ is a $T \times (Q + 1)$ Vandermonde matrix (cf. Horn and Johnson, 1991, Sec. 6.1). Suppose $T \geqslant Q + 1$ and recall that all numbers m_t are distinct, then we can construct a square submatrix of $\mathbf{M}$ of size $(Q + 1) \times (Q + 1)$ that is also a Vandermonde matrix. Hence, that matrix is said to be positive definite, and we obtain $\boldsymbol{\beta}_{gk} = \tilde{\boldsymbol{\beta}}_{gk}$.

Identifiability of π_g

Recall that $\sum_{k=1}^{K} \omega_{gjtk}(\lambda_g) = 1$. From Equation (A4), we have $\forall(g,j,t,k)$,

$$\sum_{k=1}^{K} \pi_g\omega_{gjtk}(\lambda_g) = \sum_{k=1}^{K} \tilde{\pi}_g\omega_{gjtk}(\tilde{\lambda}_g) \qquad \Rightarrow \qquad \pi_g = \tilde{\pi}_g,$$

which proves the identifiability of π_g.

Identifiability of $\boldsymbol{\lambda}_{\mathrm{g}}$

We have $\forall(g, j, t, k)$, $\omega_{gjtk}(\lambda_g) = \omega_{gjtk}(\tilde{\lambda}_g)$, where ω_{gjtk} is a logistic function. From the identifiability conditions proposed by Silvapulle (1981), we obtain $\lambda_g = \tilde{\lambda}_g$.