Statistics and Computing manuscript No.
(will be inserted by the editor)

Dynamic model-based clustering for spatio-temporal data

Lucia Paci · Francesco Finazzi

Received: date / Accepted: date

Abstract In many research fields, scientific questions are investigated by analyzing data collected over space and time, usually at fixed spatial locations and time steps and resulting in geo-referenced time series. In this context, it is of interest to identify potential partitions of the space and study their evolution over time. A finite space-time mixture model is proposed to identify level-based clusters in spatio-temporal data and study their temporal evolution along the time frame. We anticipate space-time dependence by introducing spatio-temporally varying mixing weights to allocate observations at nearby locations and consecutive time points with similar cluster's membership probabilities. As a result, a clustering varying over time and space is accomplished. Conditionally on the cluster's membership, a state-space model is deployed to described the temporal evolution of the sites belonging to each group. Fully posterior inference is provided under a Bayesian framework through Monte Carlo Markov Chain algorithms. Also, a strategy to select the suitable number of clusters based upon the posterior temporal patterns of the clusters is offered. We evaluate our approach through simulation experiments and we illustrate using air quality data collected across Europe from 2001 to 2012.

The research was partially funded by a FIRB2012 grant (project no. RBFR12URQJ) provided by the Italian Ministry of Education, Universities and Research.

L. Paci
Department of Statistical Sciences, University of Bologna
Tel.: +39 0512094657
E-mail: lucia.paci2@unibo.it

F. Finazzi
Department of Management, Information and Production Engineering, University of Bergamo
Tel.: +39 0352052363
E-mail: francesco.finazzi@unibg.it

Keywords Bayesian analysis · Finite mixture models · Markov chain Monte Carlo · State-space modeling

1 Introduction

In many research fields, scientific questions are investigated by analyzing data collected over space and time, i.e., spatio-temporal data. Customarily, data gathered at fixed spatial locations and regular time steps is referred to as point-referenced time series. In order to understand complex systems, it is important to extract useful information from such spatio-temporal datasets. In this work, extracting useful information is referred to as identifying spatial and temporal patterns in the observed phenomenon, with the main assumption that the temporal patterns are relatively small in number. This might be helpful to understand the problem at hand and, eventually, to make decisions on the basis of concise information.

When the interest is on the temporal evolution of an observed phenomena, geo-referenced time series can be pooled over space to look at the overall temporal pattern, ignoring the spatial dependence across time series at different locations. However, this approach yields to bias results if the data-generating process differs between the time series. Rather, at each time, information can be pooled over space within a small number of groups according to the underlying process driving the data. Moreover, such spatial partition can vary dynamically along time depending upon the temporal evolution of the underlying process.

When dealing with spatio-temporal data, statistical models are widely adopted to understand and predict responses of interest across space and over time. Customary, spatio-temporal modeling (Cressie and Wikle,

2011; Banerjee et al, 2014) relies on an implicit idea of grouping which depends upon model choices (e.g. neighborhood structure, covariance function). Spatial clustering within the Bayesian framework is often performed via nonparametric approaches; see, for instance, methods based on the spatial Dirichlet process (Gelfand et al, 2005; Duan et al, 2007), the spatial stick-breaking process (Reich and Fuentes, 2007), the Dirichlet labeling process (Nguyen and Gelfand, 2011) and the recent spatial product partition model (Page and Quintana, 2016). Although these approaches are quite appealing, their benefit can be limited when covariates are available to help cluster's identification. With regard to time series clustering, Nieto-Barajas and Contreras-Cristán (2014) proposed a Bayesian semiparametric mixture model centered in a state-space to induce clustering of time series. Finazzi et al (2015) developed a modified version of the state-space model to study the temporal coherence of ecological time series assuming that the observed time series share common temporal patterns along the entire temporal frame of observation.

Model-based clustering of spatio-temporal data can be described within the class of finite mixture models under a Bayesian perspective. In this framework, Fernández and Green (2002) developed a spatial mixture model for areal data with a variable number of mixing components and spatially dependent mixing weights. Frühwirth-Schnatter and Kaufmann (2008) proposed a clustering approach based on finite mixtures of dynamic regression models that allows for pooling within clusters. In Viroli (2011), a finite mixture model is employed to study three-way data which includes, among others, spatio-temporal data. Neelon et al (2014) used a finite mixture model to analyze multivariate areal-referenced data, introducing spatial random effects for each mixture component, as well as for the mixing weighs. Hossain et al (2014) used a space-time mixture of Poisson regression models to investigate relabeling algorithms and model selection issues. However, there is little work in the spatio-temporal setting for clustering point-referenced time series with spatial partitions that are allowed to vary dynamically over time.

Our contribution is to propose a space-time model-based approach to identify dynamic clusters in spatio-temporal data. Our approach builds upon the finite mixture modeling, where each mixture component describes a cluster with a level-based meaning. Within finite mixture modeling, space-time dependence is anticipated by introducing spatio-temporally varying mixing weights. Indeed, we envision a latent spatial process that evolves dynamically over time and drives the mixing probabilities. Also, spatial and temporal covariates can be easily included in the mixing weights to facilitate groups' identification. As a result, data observed at nearby locations and consecutive time points is assigned with similar cluster membership's probabilities. According to such probabilities, data collected at spatial locations are partitioned into K mutually exclusive groups at each time. In other words, a clustering varying over time and space is accomplished. Conditionally on the cluster's membership, a state-space model is deployed to described the temporal evolution of the sites belonging to each cluster. Hence, we borrow strength of information of all sites belonging to a given cluster at a given time to estimate the average level of the cluster at that time. We interpret the cluster level, varying over time, as the temporal pattern of the cluster.

Fully posterior inference is provided under a Bayesian framework through Monte Carlo Markov Chain (MCMC) algorithms. Moreover, we offer a strategy to select the suitable number of clusters using the posterior inference on the temporal pattern of the clusters.

Our application is the assessment of air quality trends from 2001 to 2012 over Europe. We illustrate our modeling approach by analyzing the annual mean of daily particulate matter to provide dynamic grouping of monitoring sites according to pollution levels.

The remainder of the manuscript is organized as follows. In Section 2 we describe our model developments, including mixing weights specification and state-space modeling. Section 3 outlines model fitting and discusses prior specification, posterior computation and posterior classification, with details deferred to the Appendix. Section 4 presents a simulation study while in Section 5 we illustrate our approach with an application on air quality data. Finally Section 6 provides a brief review and indications for future work. Supplementary materials are available online.

2 A Bayesian space-time mixture model

Let $y_t(\mathbf{s})$ be a response variable observed at time t ($t = 1, ..., T$) and location $\mathbf{s} \in \mathbb{R}^2$. We assume that observation $y_t(\mathbf{s})$ comes from a finite mixture model, that is

$$f\left(y_t(\mathbf{s}) \mid \boldsymbol{\pi}, \boldsymbol{\Theta}\right) = \sum_{k=1}^{K} \pi_{t,k}(\mathbf{s}) f\left(y_t(\mathbf{s}) \mid \boldsymbol{\Theta}_k\right) \quad (1)$$

where K is the number of components. The distribution under the k-th component ($k = 1, \ldots, K$) is denoted by $f(\cdot \mid \boldsymbol{\Theta}_k)$ where f is a density function of specified form and $\boldsymbol{\Theta}_k$ denotes the set of parameters of each component distribution. The mixing probability $\pi_{t,k}(\mathbf{s})$ is the probability that the location $\mathbf{s}$ belongs to

component k at time t and it satisfies $\pi_{t,k}(\mathbf{s}) > 0$ with $\sum_{k=1}^{K} \pi_{t,k}(\mathbf{s}) = 1$ for each $\mathbf{s}$ and t.

As usual in Bayesian analysis, a hierarchical formulation of the mixture model is exploited to facilitate the computation. For each observation, we introduce a latent allocation variable, $w_t(\mathbf{s})$, that identifies the component membership of $y_t(\mathbf{s})$, that is $Pr(w_t(\mathbf{s}) = k) = \pi_{t,k}(\mathbf{s})$. In other words, we assume that the allocation variables $w_t(\mathbf{s})$ are conditionally independently distributed given $\pi_{t,k}(\mathbf{s})$ and they come from a multinomial distribution. Given the latent $w_t(\mathbf{s})$, the observations $y_t(\mathbf{s})$ are independent with

$$f\big(y_t(\mathbf{s}) \mid w_t(\mathbf{s}) = k, \boldsymbol{\Theta}\big) = f\big(y_t(\mathbf{s}) \mid \boldsymbol{\Theta}_k\big). \tag{2}$$

The allocation variables $w_t(\mathbf{s})$ define a random partition of the data in the sense of Lau and Green (2007), as $y_t(\mathbf{s})$ and $y_{t'}(\mathbf{s}')$ belong to the same component if and only if $w_t(\mathbf{s}) = w_{t'}(\mathbf{s}')$. As customary in model-based clustering, we interpret each mixture component as a cluster, such that observations are partitioned into mutually exclusive K groups. Alternatively, clusters can be determined by merging mixture components according to some criterion (Hennig, 2010; Melnykov, 2016).

2.1 Spatio-temporally varying mixture weights

The mixing probabilities, $\pi_{t,k}(\mathbf{s})$, are allowed to vary from observation to observation, i.e., across space and over time. In particular, we introduce space-time dependence in the observations through the prior distribution of the weights such that observations corresponding to nearby locations and consecutive time points are more likely to have similar allocation probabilities than observations that are far apart in space and time.

For each location $\mathbf{s}$ and time t, the weights take the form

$$\pi_{t,k}(\mathbf{s}) = \frac{\exp\left(\mathbf{x}_t'(\mathbf{s})\boldsymbol{\beta}_k + \phi_{t,k}(\mathbf{s})\right)}{\sum_{l=1}^{K} \exp\left(\mathbf{x}_t'(\mathbf{s})\beta_k + \phi_{t,l}(\mathbf{s})\right)} \tag{3}$$

where $\mathbf{x}_{t,k}(\mathbf{s})$ is a $p \times 1$ vector of covariates, $\phi_{t,k}(\mathbf{s})$ are spatio-temporal random effects and $\boldsymbol{\beta}_1 = 0$ and $\phi_{t,1}(\mathbf{s}) = 0$ $(t = 1, \ldots, T)$ to ensure identifiability. The logistic-type transformation in (3) guarantees that the two conditions mentioned in Section 2 are satisfied (Fernández and Green, 2002). When available, covariates may help in predicting group membership's probabilities, yielding useful insights into the factors that determine group membership. Moreover, random effects provide adjustment in space and time to the explanation provided by covariates. Therefore, the response distribution is allowed to vary in flexible ways across time, space and covariate profiles.

To allow for dynamics over time and dependence over space we assume, for $k = 2, \ldots, K$,

$$\phi_{t,k}(\mathbf{s}) = \rho_k \phi_{t-1,k}(\mathbf{s}) + \zeta_{t,k}(\mathbf{s}) \tag{4}$$

where $\zeta_{t,k}(\mathbf{s})$ are independent-in-time spatially correlated errors coming from a zero-mean Gaussian process equipped with a spatial covariance function, i.e., $\zeta_{t,k}(\mathbf{s}) \overset{ind}{\sim} \mathrm{GP}\big(\mathbf{0}, \lambda_k^2 C(\cdot\,; \theta)\big)$. Several function can be employed to describe the spatial correlation between sites. For instance, a popular example of isotropic correlation function is provided by the exponential function that is, $C(\mathbf{s}_i, \mathbf{s}_j\,; \theta) = \exp\{-\theta \left\|\mathbf{s}_i - \mathbf{s}_j\right\|\}$ where θ describes the decay rate of correlation as a function of the distance between locations.

The space-time structure of random effects $\phi_{t,k}(\mathbf{s})$ induces space-time dependence among the mixing probabilities, allowing to borrow strength information from nearby sites and consecutive time steps. As a result, similar outcomes at near space and time points are assigned with similar cluster membership's probabilities. Finally, model simplifications can be easily achieved assuming only spatially or only temporally varying mixing weights π's.

2.2 State-space modeling

Model (1) requires the specification of the sampling density $f\left(y_t(\mathbf{s}) \mid \boldsymbol{\Theta}_k\right)$. The approach pursued in this work is based on dynamic linear modeling, often referred to as state-space models (West and Harrison, 1997). In particular, we assume a dynamic linear model to describe the temporal dynamic evolution of all the sites within component k.

Let $\mathbf{y}_t = \left(y_t(\mathbf{s}_1), \ldots, y_t(\mathbf{s}_n)\right)'$ be the $n \times 1$ observation vector at time t, where n is the number of locations. Conditionally on the allocation variables, the space-state model is provided by

$$\begin{aligned} \mathbf{y}_t &= \mathbf{H}_t \mathbf{z}_t + \boldsymbol{\varepsilon}_t \\ \mathbf{z}_t &= \mathbf{G}\mathbf{z}_{t-1} + \boldsymbol{\eta}_t \end{aligned} \tag{5}$$

where $\mathbf{z}_t = (z_{t,1}, \ldots, z_{t,K})'$ is the $K \times 1$ state vector, $\mathbf{H}_t$ is a $n \times K$ matrix defined below, and $\mathbf{G}$ is a $K \times K$ stable transition matrix. Finally, $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \sigma^2 I_n)$ is the $n \times 1$ measurement error vector and $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \Sigma_\eta)$ is the $K \times 1$ innovation vector. This formulation is very general and flexible and it allows to handle different time series analysis problems in a single framework.

We now turn to matrix $\mathbf{H}_t$. Suppose that site $\mathbf{s}$ belongs to component k at time t. Then, the i-th row of matrix $\mathbf{H}_t$ contains a single element equal to one at position k, while all the other elements are filled with zeros (Inoue et al, 2007; Finazzi et al, 2015). Note that, the

one-zero structure of matrix $\mathbf{H}_t$ is allowed to vary over time according to mixing probabilities $\pi_{t,k}(\mathbf{s})$. Moreover, we benefit from the borrowing strength of information of all sites belonging to component k at time t, since they all contribute in estimating the common latent state $z_{t,k}$. Given the specification in (5), the desired temporal pattern of cluster k is represented by latent state $z_{t,k}$.

3 Model fitting

3.1 Prior distributions

We complete the hierarchy of the model by specifying the prior distribution for all the hyperparameters. In particular, we place flat normal priors on the regression coefficients of the mixing weights, i.e., $\beta_k \sim N(0, 10^4)$, $k = 2, \ldots, K$. We assume a diagonal matrix $\mathbf{G} = \text{diag}(g_1, \ldots, g_K)$ and we specify flat normal priors on its diagonal entries restricted in the interval $(-1, 1)$. Similarly, for ρ_k $(k = 2, \ldots, K)$, we place a flat normal prior distribution truncated in $(-1, 1)$. We assume a diagonal matrix $\Sigma_\eta = \text{diag}(\tau_1^2, \ldots, \tau_K^2)$ and independent inverse gamma distributions, $\text{IG}(a, b)$, on its diagonal entries. Moreover, variance components λ_k^2 and σ^2 are assumed to follow an inverse gamma distribution, independently. In our implementation we take $a = 2$ and $b = 1$ to have a proper vague prior specification for each of these variance components.

The dynamic structure of the mixing weights requires an initial condition for the initial states of random effects $\boldsymbol{\phi}_{1,k} = \big(\phi_{1,k}(\mathbf{s}_1), \ldots, \phi_{1,k}(\mathbf{s}_n)\big)'$; and we assume $\boldsymbol{\phi}_{1,k} \sim N\big(\mathbf{0}, \lambda_k C(\theta)\big)$. Similarly, the autoregressive state equation requires a prior distribution for the initial states $z_{1,k}$ that are assumed to be independent normal distributions centered in zero with large variance 10^4.

Finally, a prior distribution for the spatial decay parameter of the exponential correlation function is needed to provide its full posterior inference. Customary choices are vague gamma priors or uniform prior distributions. However, under weak prior distributions, the MCMC algorithm for λ_k^2 and θ is often poorly behaved due to the weak identifiability and the slow-mixing of the associated Markov chains. Hence, we adopt an empirical Bayes approach by setting the value of parameter θ as suggested by standard exploratory spatial analysis (e.g. variogram) and then we infer about the variance conditional on this value. Moreover, with no updating of θ in the MCMC, the covariance matrix of $\zeta_{t,k}(\mathbf{s})$ and its inversion needs to be calculated only once, expediting substantially the computation.

3.2 Posterior computation

Recalling allocation variables $w_t(\mathbf{s})$ and using the conditional independence assumption, the joint posterior distribution is expressed as

$$
\begin{aligned}
& p\left(\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{\lambda}^2, \mathbf{z}, \mathbf{G}, \Sigma_\eta, \sigma^2 \mid \mathbf{y}\right) \propto \\
& \times \prod_{k=1}^{K} \prod_{t=1}^{T} \prod_{i=1}^{n} \Big[\pi_{t,k}(\mathbf{s}_i)\ \mathrm{N}\big(y_t(\mathbf{s}_i);\ z_{t,k},\ \sigma^2\big)\Big]^{I(w_t(\mathbf{s})=k)} \\
& \times \prod_{k=2}^{K} \prod_{t=2}^{T} \mathrm{N}\big(\boldsymbol{\phi}_{t,k};\ \rho_k\, \boldsymbol{\phi}_{t,k},\ \lambda_k^2 C(\theta)\big) \\
& \times \prod_{k=1}^{K} \prod_{t=1}^{T} \mathrm{N}\big(z_{t,k};\ g_k\, z_{t-1,k},\ \tau_k^2\big) I(z_{t,1} < \cdots < z_{t,K}) \\
& \times \prod_{k=2}^{K} p(\beta_k)\, p(\rho_k)\, p(\lambda_k^2)\, p(\boldsymbol{\phi}_{1,k}) \\
& \times \prod_{k=1}^{K} p(g_k)\, p(\tau_k^2)\, p(z_{1,k})
\end{aligned}
\tag{6}
$$

where bold symbols represent all the elements associated with the corresponding parameter, $I(\cdot)$ denotes the indicator function and $p(\cdot)$ represents the prior distributions for their respective parameters, as described in the previous subsection. The order constraint in the distribution of the latent state z's is imposed to ensure identifiability of the estimates as discussed in Subsection 3.3.

We employ MCMC algorithms to evaluate the joint posterior distribution, using Metropolis steps for updating the mixing parameters and Gibbs steps for updating all the other parameters. In particular, after assigning initial values to the model parameters, the sampling scheme is given by the following steps:

1. for $k = 2, \ldots, K$, sample coefficients β_k using a random-walk Metropolis step;
2. for $k = 2, \ldots, K$, update $\boldsymbol{\phi}_{t,k} = \big(\phi_{t,k}(\mathbf{s}_1), \ldots, \phi_{t,k}(\mathbf{s}_n)\big)'$ using a Metropolis step with a conditional prior proposal (Knorr-Held, 1999);
3. for $k = 2, \ldots, K$, update ρ_k and λ_k^2 from their closed form full conditional distribution;
4. for $t = 1, \ldots, T$ and $i = 1, \ldots, n$, sample allocation variables $w_t(\mathbf{s}_i)$ from a multinomial distribution taking values $\{1, \ldots, K\}$ with posterior probabilities $\pi_{t,k}(\mathbf{s})^\star$ described in the Appendix;
5. for $t = 1, \ldots, T$, sample latent state $\mathbf{z}_t$ from its closed form full conditional distribution and apply the order restriction $I(z_{t,1} < \cdots < z_{t,K})$;
6. for $k = 1, \ldots, K$, update g_k and τ_k^2 from their closed form full conditional distribution;
7. update σ^2 from its closed form full conditional distribution.

The closed form full conditional distributions of the model parameters are deferred to the Appendix. Con-

vergence of the chains is monitored using standard MCMC diagnostics.

3.3 Identification and posterior classification

Within the Bayesian analysis, if exchangeable priors are placed upon the parameters of a mixture model, then the resulting posterior distribution is invariant to permutations in the labeling of the parameters (Jasra et al, 2005). As a result, the estimates of mixture components in every iteration of MCMC algorithm are not sensitive to the estimates of allocation variables.

To perform posterior classification and to estimate the group-specific parameters, the finite mixture model must be identified to avoid label switching. Since this is a common issue in Bayesian mixture modeling, many ideas has been proposed to deal with label switching (see e.g. Stephens 2000; Frühwirth-Schnatter 2006; Sperrin et al 2010). Here, we elicit the idea of level-based partitions by assuming that, a priori, the mean level of the groups, $z_{t,k}$, are in increasing order at each time. This order restriction enables to avoid the label switching problem and to identify the model (Fernández and Green, 2002).

Once the model has been identified it is possible to classify the spatio-temporal observations into the different groups. Under the assumption that each mixture component is interpreted as a cluster, we allocate the observations using the posterior classification probability. Therefore, given the MCMC draws, we register the cluster membership and we estimate the posterior classification probability for each observation, $Pr(w_t(\mathbf{s}) = k|\mathbf{y})$, as the relative frequency (relative to the number of posterior samples) corresponding to the event $w_t(\mathbf{s}) = k$. To provide the clustering, we assign each observations to their most likely group according to the posterior probabilities of $w_t(\mathbf{s})$, that is using the maximum a posteriori probability (MAP) rule.

3.4 Number of clusters

The number of mixture components K is usually unknown in practice and needs to be estimated. In this case, reversible jump (Richardson and Green, 1997; Dellaportas and Papageorgiou, 2006) or birthdeath (Viroli, 2011) MCMC methods can be employed. Although these approaches enable full posterior inferences on the number of components, they are computationally intensive, particularly with big spatio-temporal datasets.

Therefore, when the number of mixing components is relatively small, a simpler way to estimate the number of components is by comparing the values of model selection criteria calculated for various mixture models with fixed number of components (Hossain et al, 2014). Recently, Malsiner-Walli et al (2016) proposed the use of sparse finite mixture models together with standard MCMC methods to estimate the number of mixture components and identify cluster-relevant variables, simultaneously, for multivariate Gaussian mixtures.

As an alternative, for level-based meaning clusters, we propose to look at the posterior temporal patterns to identify the number of clusters. In fact, when the interest is on the temporal evolution of the groups, it seems sensible to estimate the number of clusters such that the corresponding posterior temporal patterns show significant differences along the observational time frame. For instance, for a given K, when the analysis provides at least two temporal patterns that show no significant differences at most of the time steps, then a smaller number of clusters would be preferred. Therefore, starting with only two clusters, we proceed by fitting the model with an increasing fixed number of components and stop when the posterior inference on the temporal patterns results in a sensible picture.

Finally, in many real applications it is reasonable to choose the number of clusters that emerges as meaningful with respect to the phenomena, say relying on a sort of 'scientific significance'.

4 Simulation study

We carry out simulation experiments in order to investigate the performance of our approach in identifying dynamic clusters. In particular, we consider the number of the clusters as well as their spatial structure and evolution over time. In this section, we show the results of a simulation study designed as follows. At $n = 100$ sparse locations and $T = 20$ times, we generate a realization of a dynamic space-time model, that is

$$O_t^\star(\mathbf{s}) = \rho^\star O_{t-1}^\star(\mathbf{s}) + \eta_t^\star(\mathbf{s}) \tag{7}$$

where $\eta_t^\star(\mathbf{s}) \sim \mathrm{GP}(0, C(\theta^\star))$ equipped with an exponential correlation function. Then, at each time, we slice the process realization with respect to $K^\star$ equidistant levels giving rise to a spatial partition. Locations within the same partition are assigned to the same cluster. Each cluster is associated with a different temporal trend $z_{t,k}^\star$, shown through black lines in Figure 1. Finally, we simulate data from $\mathrm{N}(z_{t,k}^\star, \sigma^{2\star})$. In our simulation setting we consider the following factors: small ($K^\star = 2$) and relatively large ($K^\star = 5$) number of clusters; low ($\rho^\star = 0.2$) and high ($\rho^\star = 0.9$) temporal correlation; low ($\theta^\star = 0.3$) and high ($\theta^\star = 1.2$) spatial correlation corresponding to roughly 10% and 90% of

the maximum distance between locations, respectively. For each factor combination, we simulate 50 datasets to investigate the performance of our approach in recovery the clusters.

First, we show our strategy to choose the number of clusters for a simulated dataset. Such data and the associated spatial partitions are shown in the Supplementary material. Again, starting with only two groups, we fit the model with an increasing number of clusters until the temporal patterns are significant different each other at most of the time steps. For instance, Figure 1 shows the posterior 95% credible interval of the latent z's obtained fitting the model with $K = 2, \ldots, 5$, given a 'true' number of clusters $K^\star = 4$. Fitting the model using $K = 2$, $K = 3$ and $K = 4$, provides significant differences among the posterior temporal patterns at most of the time. Rather, with $K = 5$, yellow and green clusters in Figure 1 exhibit overlapped trends along the time frame, suggesting a smaller number of groups. So, according to our strategy, the 'true' number of clusters, $K^\star = 4$, is recovered.

For each simulated dataset, we fit our model setting the number of clusters equal to 'true' number of clusters, i.e., $K = K^\star$. We offer a comparison of our approach with a simpler Bayesian mixture model with spatio-temporally invariant mixing probabilities. We fit, at each time, a univariate Gaussian mixture model with standard Dirichlet prior on the mixing weight. The estimation via MCMC methods is provided by the `R` package `BayesMix` (`https://cran.r-project.org/package=bayesmix`). Then, an ordering constraint on the component's means is imposed to avoid label switching.

The performance of the two approaches is evaluated through the misclassification error rate (Ranciati et al, 2016), i.e., the average number of units not correctly allocated when compared to the known simulated membership over time. Figures 2 and 3 show the misclassification error rate for each time for two simulation settings with $K = 2$ and $K = 4$, respectively. Identifying the group to which each observation belongs at the beginning and ending of the time frame is a relatively easy task because the temporal patterns of clusters are well-separated. So, both approaches do recover well the spatial partitions. Conversely, as all the latent trends approach zero, the allocation problem becomes more challenging. Figures 2 and 3 show that our approach outperforms the simpler mixture model. Clearly, the benefit of considering spatio-temporally varying weights is appreciated when the underlying process generating the data is strongly correlated over time and space. Indeed, for the simulation setting with $\rho = 0.9$ and $\theta = 1.2$ we obtain a reduced misclassification error of roughly 60% and 30% corresponding to $K = 2$ and $K = 4$, respectively.

The interested reader is referred to the Supplementary material that provides the results for the other two simulation settings and a second simulation experiment over a spatial grid.

5 Analysis of air quality trends

Understanding the status and trends of air quality levels is crucial to support policy development and implementation, as well as to inform about the challenges in air quality assessment and management that need to be met (Guerreiro et al, 2014). The assessment of air quality is based on ambient air measurements collected from monitoring stations at fixed locations over times. Monitoring data is often aggregated over space to estimate a joint temporal pattern, ignoring the spatial dependence between locations. Alternatively, information is pooled over space within a small number of groups according to specific features of the monitoring stations (Bruno et al, 2013).

According to the European Ambient Air Quality Directive (AQD, EU 2008), air quality stations for compliance monitoring are classified as traffic (stations located in proximity to a single major road), industrial (stations located in proximity to a single industrial source or industrial area) and background (any location which is neither to be classified as traffic or industrial). For particulate matter, for instance, it is expected that concentrations collected at traffic stations are higher than those gathered at background sites. However, such classification does not always corresponds to similar observed pollution levels within the groups. Moreover, the classification is fixed over time, despite the environment surrounding the sites may change considerably along the time. The Implementing Provisions on Reporting of AQD (EU, 2011) has clarified that each station should be classified according to the predominant emission sources relevant for the measurement configuration for each pollutant. In other words, each station could have a number of different classifications for different pollutants with a classification that may change over time (Vincent and Stedman, 2013).

We illustrate our approach by modeling the annual mean of daily fine particulate matter (PM_{10}) gathered from 523 monitoring sites across western Europe, from 2001 to 2012. Data is free available from the European air quality database AirBase maintained by the European Environmental Agency (http://www.eea.europa.eu/). Figure 4 shows the PM_{10} monitoring stations used in the analysis. For each site, we also consider its elevation that we employ as a spatially varying covariate in the

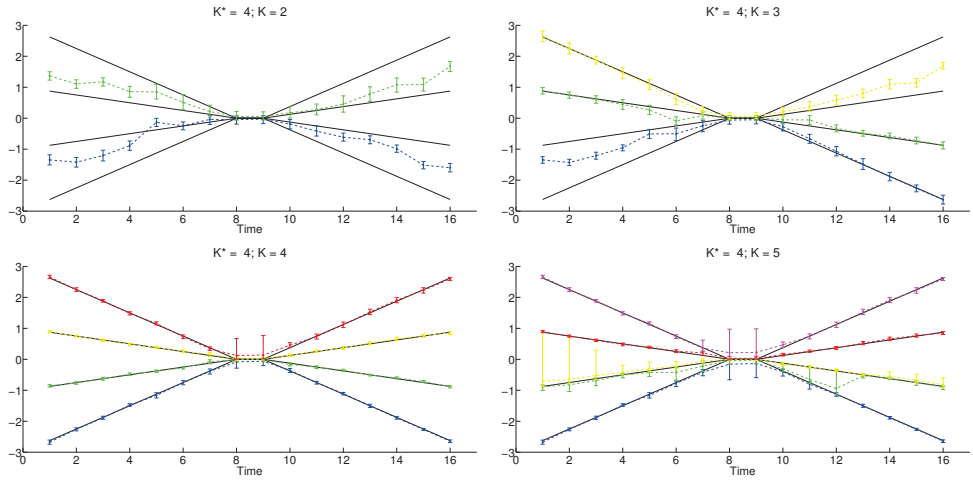

Fig. 1 Posterior 95% credible interval of the temporal patterns z's obtained using an increasing number of clusters for a datasets generated according to the simulation design. Black lines represent temporal patterns $z^{\star}_{t,k}$.

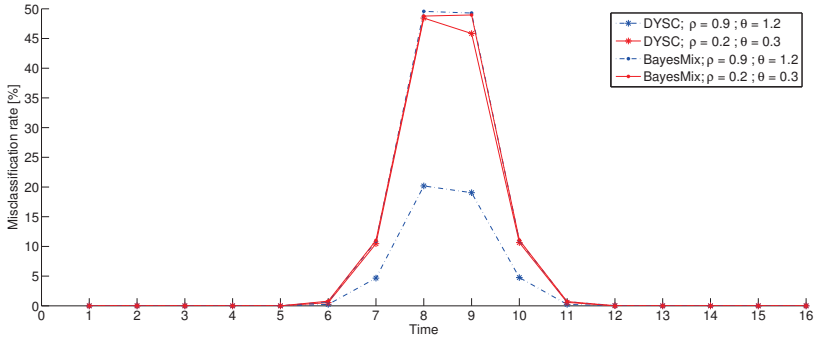

Fig. 2 Misclassification error rate over time resulting from our dynamic space-time clustering (DYSC) and a simple Bayesian mixture model (BayesMix) for two simulation settings with $K = 2$.

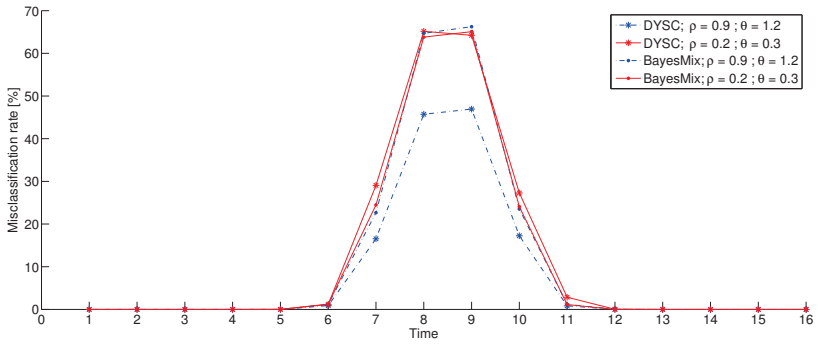

Fig. 3 Misclassification error rate over time resulting from our dynamic space-time clustering (DYSC) and a simple Bayesian mixture model (BayesMix) for two simulation settings with $K = 4$.

mixing weights model (3) to help in identifying the clusters. Usually, transformations are applied to PM_{10} data in order to make the normality assumption acceptable to fit customary space-time models (see e.g. Cocchi et al 2007). Here, we benefit from the Gaussian mixture that is well suited to model not normally distributed data.

Following our strategy described in Subsection 3.4, the number of clusters corresponding to significant different posterior temporal patterns would be five. However, we show here the results obtained by setting $K = 3$ that represents, according to our experience, a more sensible choice in understanding air quality trends. Posterior inference is carried out by implementing the al-

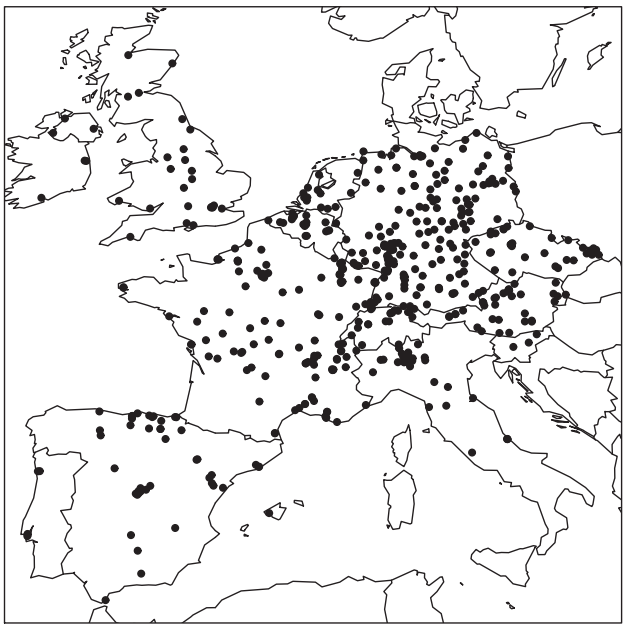

Fig. 4 PM_{10} monitoring stations across Europe used in our study.

gorithm described in Subsection 3.2 with 100,000 iterations, 60,000 as burn-in period, and keep one of every 15th iteration to reduce the autocorrelation of the chains. Running time on a Intel(R) Core(TM) i7-3537U CPU (2.50GHz, 8 GB RAM) is roughly 0.05 seconds per iteration.

Table 1 shows the posterior summaries of model parameters. The coefficients associated with the elevation are significant and reveal that the relative probability of belonging to clusters 2 and 3 rather than being in cluster 1 is roughly 50% lower for a meter increase in the elevation. The variances λ's of the latent space-time processes show significant differences while all the autoregressive coefficients ρ's and g's are very close to one, because of the strong temporal correlation of PM_{10} concentrations. The posterior average maps of the latent processes $\phi_{t,k}(\mathbf{s})$ are presented in the Supplementary materials and show the spatial patterns of the processes and their evolution over years. Figure 5 shows the posterior 95% credible interval of the temporal trends z's; following the increasing order of the temporal patterns, we denote the low-level, middle-level and high-level clusters as the first, second and third clusters, respectively. For each year, the spatial partition is obtained using the maximum a posteriori probability rule and displayed in Figure 6, that allows to appreciate the borrowing of strength across space and over time in the resulting clustering. Overall, we note decreasing levels of fine particulate matter over the years, with a clearer drop in the temporal pattern of the third cluster. This is likely due to the effect of stronger policies for air pollution reduction that has been applied in particular regions of Europe over the last years. Indeed, the monitoring sites belonging to the third cluster are located in regions well-known for their bad air quality conditions, such as the Po Valley (Italy), the eastern Czech Republic, the south of Spain and the Benelux, see Figure 6. From Figure 6 we can also extract some interesting stories. For instance, the monitoring sites located in the north-western of Germany move from the second to the first cluster, suggesting an improvement of the air quality status in that part of the country over years. Similarly, we note that monitoring sites of the northern Spain move from the third cluster to the second one, showing a reduction of air pollution concentrations over such region. Conversely, sites located in France (outside the center of the country) move from the first to the second cluster, showing a worsening situation in air quality over years. Moreover, we look at the composition of the clusters with respect to the station type. As expected, the 84% of sites belonging to the first cluster are background stations; however, the third cluster does not contain only traffic sites, rather it consists of a similar number of background and traffic sites, with changes over time. Finally, we can assess model adequacy by computing the empirical coverage of the 95%

Table 1 Posterior means and 95% credible intervals for model parameters.

	Posterior mean	95% credible interval
$\beta_{0,2}$	0.215	[0.006, 0.422]
$\beta_{0,3}$	-1.292	[-1.682, -0.891]
$\beta_{1,2}$	-0.761	[-0.830, -0.698]
$\beta_{1,3}$	-0.737	[-0.871, -0.606]
ρ_2	0.996	[0.987, 0.999]
ρ_3	0.998	[0.994, 0.999]
λ_2^2	1.081	[1.013, 1.215]
λ_3^2	3.137	[2.772, 3.446]
g_1	0.992	[0.970, 0.999]
g_2	0.991	[0.966, 0.999]
g_3	0.993	[0.974, 0.999]
τ_1^2	0.874	[0.293, 2.101]
τ_2^2	2.620	[1.155, 5.939]
τ_3^2	4.408	[1.630, 9.894]
σ^2	21.653	[20.780, 22.515]

predictive interval, i.e., generating the replicate observations under the model and look at the proportion of predictive intervals containing the observations. Averaging over time and space, we find that 97% contains the respective observed concentrations.

6 Summary and future works

We have proposed a finite mixture model to provide a dynamic clustering of spatio-temporal data. We have introduced spatio-temporally varying mixing weights to accommodate space-time dependence and assign data observed at nearby locations and consecutive time points with similar cluster membership's probabilities. Conditionally on the cluster's membership, a state-space model has been employed to describe the temporal evolution of the sites belonging to each cluster. Also, a procedure to select the number of clusters has been offered. The approach is very flexible and allows clusters identification also with geo-referenced time series affected by missingness.

Currently, we are developing a MATLAB Toolbox, called DYnamic Spacetime Clustering (DYSC), to provide easy implementation of our approach. With regard to the computation, the MCMC algorithm can suffer of poor mixing with a large K. In this case, alternative augmentation approaches and sampling schemes can be employed. For instance, a data augmentation step based on auxiliary Pólya-Gamma variables (Polson et al, 2013) can be used. Finally, a natural extension of our approach will move from the univariate to the multivariate setting, in order to identify clusters in spatio-temporal multivariate responses, such multiple pollutants.

Acknowledgements The authors thank the air quality service at ARPAE Emilia-Romagna for helpful discussions.

A Appendix

The full conditional distribution of the variances λ_k^2, for $k = 2, \ldots, K$, is

$$\lambda_k^2 \mid \text{rest} \sim \text{IG}\left(a + \frac{Tn}{2}, b + \frac{1}{2}\sum_{t=1}^{T}\left(\boldsymbol{\phi}_{t,k} - \rho_k \boldsymbol{\phi}_{t-1,k}\right)' C(\theta)^{-1} \left(\boldsymbol{\phi}_{t,k} - \rho_k \boldsymbol{\phi}_{t-1,k}\right)\right).$$

The full conditional distribution of the variances τ_k^2, for $k = 1, \ldots, K$, is

$$\tau_k^2 \mid \text{rest} \sim \text{IG}\left(a + \frac{T}{2}, b + \frac{1}{2}\sum_{t=1}^{T}\left(z_{t,k} - g_k z_{t-1,k}\right)^2\right).$$

The full conditional distribution of the error variance σ^2 is given by

$$\sigma^2 \mid \text{rest} \sim \text{IG}\left(a + \frac{Tn}{2}, b + \frac{1}{2}\sum_{t=1}^{T}\left(\mathbf{y}_t - \mathbf{H}_t \mathbf{z}_t\right)'\left(\mathbf{y}_t - \mathbf{H}_t \mathbf{z}_t\right)\right).$$

The full conditional distribution of ρ_k, $k = 2, \ldots, K$, is a univariate normal distribution $N(vd, v)$ restricted in the interval $I(-1 < \rho_k < 1)$, where

$$v^{-1} = \frac{1}{\lambda_k^2}\boldsymbol{\phi}'_{t-1,k} C(\theta)^{-1} \boldsymbol{\phi}_{t-1,k} + 10^{-4}$$
$$d = \frac{1}{\lambda^2}\boldsymbol{\phi}'_{t-1,k} C(\theta)^{-1} \boldsymbol{\phi}_{t,k}.$$

The full conditional distribution of g_k, $k = 1, \ldots, K$, is a univariate normal distribution[1] $N(vd, v)$ truncated in the interval $I(-1 < g_k < 1)$, where

$$v^{-1} = \frac{1}{\tau_k^2}\sum_{t=1}^{T} z_{t-1,k}^2 + 10^{-4}$$
$$d = \frac{1}{\tau_k^2}\sum_{t=1}^{T} z_{t-1,k} z_{t,k}.$$

The full conditional distribution of the allocation variables $w_t(\mathbf{s})$ is given by

$$w_t(\mathbf{s}) \mid \text{rest} \sim \text{Multinomial}\left(\pi_{t,1}(\mathbf{s})^\star, \ldots, \pi_{t,K}(\mathbf{s})^\star\right)$$

where the posterior probabilities are

$$\pi_{t,k}^\star(\mathbf{s}) = \frac{\pi_{t,k}(\mathbf{s}) N(z_{t,k}, \sigma^2)}{\sum_{l=1}^{K} \pi_{t,l}(\mathbf{s}) N(z_{t,l}, \sigma^2)}.$$

The full conditional distribution of the latent states $\mathbf{z}_t$ is a multivariate normal distribution $N(VD, V)$, where

[1] For notation simplicity, same symbols are re-used.

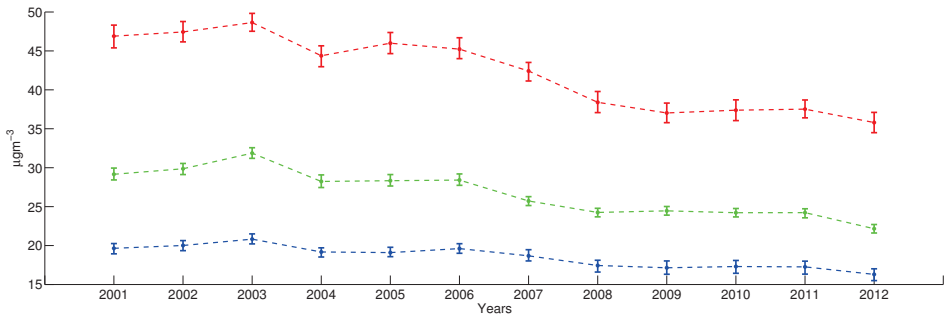

Fig. 5 Posterior 95% credible interval of the temporal patterns $z_{t,k}$ of PM_{10} in μgm^{-3}; first, second and third clusters are displayed in blue, green and red, respectively.

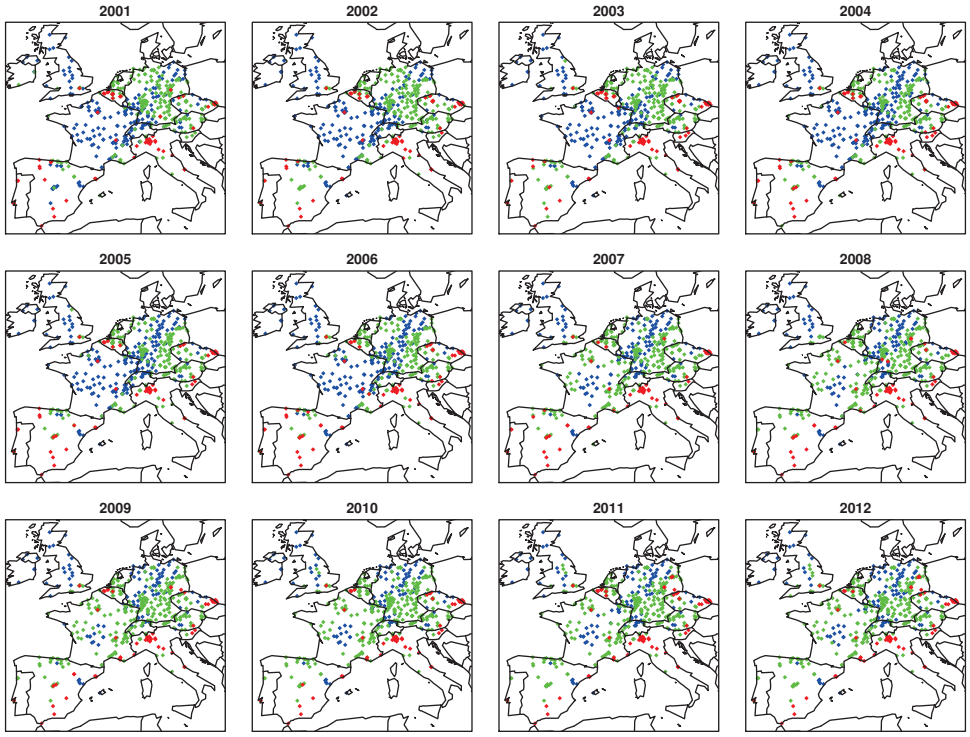

Fig. 6 Spatial partitions for each year; first, second and third clusters are displayed in blue, green and red, respectively.

- $t = 1$

$$V^{-1} = \frac{1}{\sigma^2}\mathbf{H}_t'\mathbf{H}_t + \mathbf{G}'\Sigma_\eta^{-1}\mathbf{G} + 10^{-4}I_K$$

$$D = \frac{1}{\sigma^2}\mathbf{H}_t'\mathbf{y}_t + \mathbf{G}'\Sigma_\eta^{-1}\mathbf{z}_{t+1}$$

- $t = 2, \ldots, T-1$

$$V^{-1} = \frac{1}{\sigma^2}\mathbf{H}_t'\mathbf{H}_t + \mathbf{G}'\Sigma_\eta^{-1}\mathbf{G} + \Sigma_\eta^{-1}$$

$$D = \frac{1}{\sigma^2}\mathbf{H}_t'\mathbf{y}_t + \mathbf{G}'\Sigma_\eta^{-1}\mathbf{z}_{t+1} + \Sigma_\eta^{-1}\mathbf{G}\mathbf{z}_{t-1}$$

- $t = T$

$$V^{-1} = \frac{1}{\sigma^2}\mathbf{H}_t'\mathbf{H}_t + \Sigma_\eta^{-1}$$

$$D = \frac{1}{\sigma^2}\mathbf{H}_t'\mathbf{y}_t + \Sigma_\eta^{-1}\mathbf{G}\mathbf{z}_{t-1}.$$

References

Banerjee S, Carlin BP, Gelfand AE (2014) Hierarchical Modeling and Analysis for Spatial Data, 2nd edn. Chapman and Hall/CRC, Boca Raton, FL

Bruno F, Cocchi D, Paci L (2013) A practical approach for assessing the effect of grouping in hierarchical spatio-temporal models. AStA Advances in Statistical Analysis 97(2):93–108

Cocchi D, Greco F, Trivisano C (2007) Hierarchical space-time modelling of PM10 pollution. Atmospheric Environment 41(3):532 – 542

Cressie N, Wikle CK (2011) Statistics for Spatio-Temporal Data. Wiley, Hoboken, NJ

Dellaportas P, Papageorgiou I (2006) Multivariate mixtures of normals with unknown number of components. Statistics and Computing 16(1):57–68

Duan JA, Guindani M, Gelfand AE (2007) Generalized spatial Dirichlet process models. Biometrika 94:809–825

EU (2008) Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on Ambient Air Quality and Cleaner Air for Europe. Official Journal of the European Union L 152:1–44, URL http://eur-lex.europa.eu/eli/dir/2008/50/oj

EU (2011) Commission Implementing Decision 2011/850/EU of 12 December 2011 laying down rules for Directives 2004/107/EC and 2008/50/EC of the European Parliament and of the Council as regards the reciprocal exchange of information and reporting on ambient air quality. Official Journal of the European Union L 335:86–106, URL http://data.europa.eu/eli/dec_impl/2011/850/oj

Fernández C, Green PJ (2002) Modelling spatially correlated data via mixtures: A Bayesian approach. Journal of the Royal Statistical Society, Series B 64:805–826

Finazzi F, Haggarty R, Miller C, Scott M, Fassò A (2015) A comparison of clustering approaches for the study of the temporal coherence of multiple time series. Stochastic Environmental Research and Risk Assessment 29:463–475

Frühwirth-Schnatter S (2006) Finite Mixture and Markov Switching Models. Springer, New York

Frühwirth-Schnatter S, Kaufmann S (2008) Model-based clustering of multiple time series. Journal of Business & Economic Statistics 26:78–89

Gelfand AE, Kottas A, MacEachern SN (2005) Bayesian nonparametric spatial modeling with Dirichlet process mixing. Journal of the American Statistical Association 100(471):1021–1035

Guerreiro CB, Foltescu V, de Leeuw F (2014) Air quality status and trends in europe. Atmospheric Environment 98:376 – 384

Hennig C (2010) Methods for merging gaussian mixture components. Advances in Data Analysis and Classification 4(1):3–34

Hossain MM, Lawson AB, Cai B, Choi J, Liu J, Kirby RS (2014) Space-time areal mixture model: relabeling algorithm and model selection issues. Environmetrics 25:84–96

Inoue LYT, Neira M, Nelson C, Gleave M, Etzioni R (2007) Cluster-based network model for time-course gene expression data. Biostatistics 8:507–525

Jasra A, Holmes CC, Stephens DA (2005) Markov chain monte carlo methods and the label switching problem in Bayesian mixture modeling. Statistical Science 20(1):50–67

Knorr-Held L (1999) Conditional prior proposals in dynamic models. Scandinavian Journal of Statistics 26(1):129–144

Lau JW, Green PJ (2007) Bayesian model-based clustering procedures. Journal of Computational and Graphical Statistics 16(3):526–558

Malsiner-Walli G, Frühwirth-Schnatter S, Grün B (2016) Model-based clustering based on sparse finite gaussian mixtures. Statistics and Computing 26(1):303–324

Melnykov V (2016) Merging mixture components for clustering through pairwise overlap. Journal of Computational and Graphical Statistics 25(1):66–90

Neelon B, Gelfand AE, Miranda ML (2014) A multivariate spatial mixture model for areal data: examining regional differences in standardized test scores. Journal of the Royal Statistical Society, Series C 63:737–761

Nguyen X, Gelfand AE (2011) The Dirichlet labeling process for clustering function data. Statistica Sinica 21:1249–1289

Nieto-Barajas LE, Contreras-Cristán A (2014) A Bayesian nonparametric approach for time series clustering. Bayesian Analysis 9(1):147–170

Page GL, Quintana FA (2016) Spatial product partition models. Bayesian Analysis 11:265–298

Polson NG, Scott JG, Windle J (2013) Bayesian inference for logistic models using PlyaGamma latent variables. Journal of the American Statistical Association 108(504):1339–1349

Ranciati S, Viroli C, Wit E (2016) Mixture model with multiple allocations for clustering spatially correlated observations in the analysis of ChIP-Seq data. ArXiv e-prints 1601.04879

Reich BJ, Fuentes M (2007) A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. The Annals of Applied Statistics 1(1):249–264

Richardson S, Green PJ (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). Journal of the Royal Statistical Society: Series B 59(4):731–792

Sperrin M, Jaki T, Wit E (2010) Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. Statistics and Computing 20(3):357–366

Stephens M (2000) Dealing with label switching in mixture models. Journal of the Royal Statistical Society: Series B 62(4):795–809

Vincent K, Stedman J (2013) A review of air quality station type classifications for uk compliance monitoring. Tech. rep., The Department for Environment, Food and Rural Affairs, Welsh Government, Scottish Government and the Department of the Environment for Northern Ireland, rICARDO-AEA/R/3387, https://uk-air.defra.gov.uk/library/reports?report_id=765

Viroli C (2011) Model based clustering for three-way data structures. Bayesian Analysis 6(4):573–602

West M, Harrison J (1997) Bayesian forecasting and dynamic models, second edition. Springer, New York