Clustering spatial networks through latent mixture models

Leonardo Egidi, Francesco Pauli, Nicola Torelli and Susanna Zaccarin

Department of Economics, Business, Mathematics and Statistics 'Bruno de Finetti', University of Trieste, Trieste, Italy

Address for correspondence: Leonardo Egidi, Department of Economics, Business, Mathematics and Statistics, University of Trieste, Trieste, Italy. Email: legidi@units.it

Abstract

We consider a Bayesian model-based clustering technique that directly accounts for network relations between territorial units and their position in a geographical space. This proposal is motivated by a practical problem: to design administrative structures that are intermediate between the municipality and the province within an Italian region based on the existence of a relatively (to population) high commuting flow. In our social network model, the commuting flows are explained by the distances between the municipalities, i.e., the nodes, in a 3-dimensional space, where the 2 actual geographical coordinates and the third latent variable are modelled through a Gaussian mixture.

Keywords: Bayesian model-based clustering, commuting flows, geographical partitioning, Gaussian mixture

1 Introduction

The need to deal with territorial units better reflecting the inherent structure of the social and economic reality than administrative divisions of countries and regions is not new in regional geography and policy planning (for example, see Openshaw, 1977). In order to improve the socio-economic organization of the territory, it is crucial to pursue policies at the right scale, especially when it comes to issues such as service provision. In fact, social and economic activities unfold and affect areas that are not limited by administrative boundaries. On the contrary, they typically exert influence on neighbouring (administrative) areas. To elaborate, the latter may not be the most appropriate geographical scale to fully understand local economies and citizens' behaviour in a number of policy domains.

A 'functional' approach in defining territorial units, which is able to account for the socio-economic trends across the space, can improve the effectiveness of public policies. Functional units should be regarded as additional and complementary to the established administrative units, allowing a better understanding of the dynamics insisting on a spatial scale not necessarily properly captured by—although small—administrative geographies (Casado-Diaz Coombes, 2011).

The demand for meaningful geographies based on a functional subdivision of the space, generally not following the known administrative borders of the territory, has been addressed with different approaches and definitions: many of these stem from the fact that the mobility of people and goods, inherently connecting different areas, is a manifestation of the existing relations, and also that the existence and temporal evolution of these flows have important effects on several topics such as housing, transport, and land use. In regional science, a main stream of research focuses on the analysis of commuting flows observed between the smallest spatial units at which data are available (for example, municipalities, counties, and so on), leading to the various but similar concepts of metropolitan areas, labour market areas, daily urban systems, or more generally, functional areas (Eurostat, 2020; OECD, 2020). Commuting (usually travel-to-work) patterns are considered the

primary factors in defining and delineating functional areas, because the extent to which workers are willing to and are able to commute daily between two places is deemed to reflect the degree of economic integration between those places. The main methodology to obtain the functional territorial units, recently put forward by Eurostat (2020) as a comparative and harmonized framework for EU countries, follows the work of Coombes et al. (1986), Coombes Bond (2008), and Coombes et al. (2012). Further, it is based on an agglomerative clustering procedure, imposing several constraints on self-containment, size, and contiguity of the resulting subdivisions. A similar but more sophisticated approach is taken by Chakraborty et al. (2013) in that they model commuting flows using a hierarchical Bayesian model on individual data and then choose a partition based on self-containment as implied by model estimates. In gravity and spacial interactions models (Celik Guldmann, 2007; Haynes Fotheringham, 1985; Roy Thill, 2013), the focus is on the spatial structure of the flows, integrating measures of size and measures of distance (highway distance, great circle distance, and so on) among spatial units. As it is intuitive, the interaction between the origin and the destination locations decreases as the distance between them increases.

We address the problem of identifying functional areas inside a wider territory employing a model-based clustering technique that directly accounts for relations between units (municipalities in our case) and their position in the geographical space. This proposal is motivated by a practical issue, which arose when the 'Regione Autonoma Friuli Venezia Giulia' (one of the 20 Italian regions in the north-east of Italy) had to deal with the redaction of the territorial governance plan, mandated by the Regional Law n. 22 of 2009 and aimed at coordinating decisions at a supra-municipal level enhancing, at the same time, the role of local communities. One of the first steps of the plan was the redaction of the 'regional strategic territorial document' where the region had to identify the 'local territorial systems'.

In the policy initiatives at the regional level, the local territorial systems are viewed as functional areas in which the strategic planning at the regional level is linked with more operational choices at the municipality level, in order to provide conditions for a balanced and effective local development. Among the criteria put forward to define such territorial systems, there was the relationship structure of the network of the 218 municipalities belonging to the region.

In order to shed light on this structure, we studied the mobility flows between municipalities with the aim of designing territorial structures intermediate between the municipality and province level, comprising contiguous municipalities that are, to some extent, self-contained. Data on the flows of commuters daily travelling between the 218 municipalities of Friuli Venezia Giulia region (see Figure 1 for the region roadmap) are used to determine geographically connected groups of

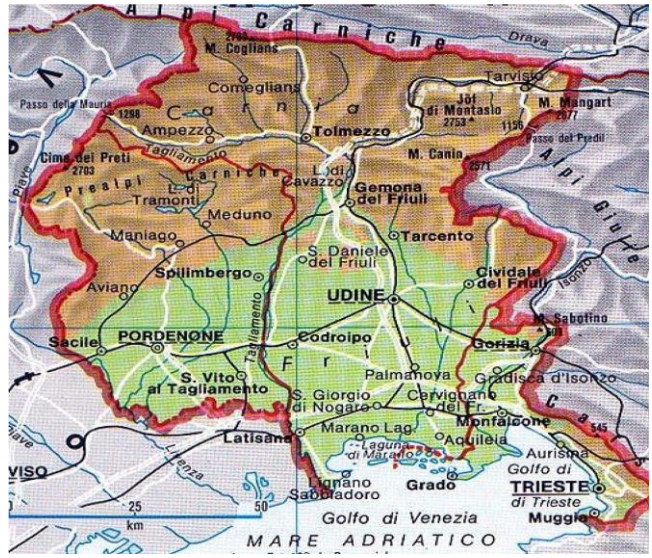

Figure 1. Roadmap of the Friuli Venezia Giulia region (north-east of Italy).

municipalities within which a high mobility exists. Two municipalities are strongly related if a high flow exists between them, where, high, means relative to the populations. The flows are clearly related to the populations of the municipalities and to the distances between them.

The task of designing new districts based on the internal flows can be naturally tackled using a social network framework, where municipalities are the nodes of the network and the relationship between them is measured by the commuting flows. In doing this, geographical partition should also be taken into account. As Daraganova et al. (2012) point out, the geographical distance between nodes in a network can have a powerful effect on the formation/intensity of a tie between them. Furthermore, spatial models have been used in analysing network data by Hoff et al. (2002), who proposed a latent space model for social network analysis, where the probability of a tie depends on the distance between actors in a latent Euclidean space. It is worth noting that network-based clustering techniques for community detection (Fortunato Hric, 2016) have been applied to commuting network, often adopting the concept of modularity (Newman Girvan, 2004) to drive the processes to delineate the resulting areas—e.g., see Farmer Fotheringham (2012) and Nelson Rae (2016). Handcock et al. (2007) used latent (unobserved) coordinates of nodes in the latent space to perform clustering by modelling them as a Gaussian mixture. More recent developments related to latent space models include their dynamic versions (Kim et al., 2018) or the possible use of nonEuclidean geometries for the space of latent variables (Smith et al., 2019).

The proposal by Handcock et al. (2007) is related to the stochastic blockmodels, which posit a latent membership vector for each node. Both approaches aim to achieve the same goal, i.e., detecting latent structures that explain the connectivity in an observed network. The former can be interpreted in terms of distances, where nodes are mapped to a Euclidean space. Conversely, the latter can be interpreted in terms of blocks of connectivity, or micro-communities, where the nodes are mapped to the space of cluster proportions (Goldenberg et al., 2010). As described in what follows, our approach coherently falls in the latent space field.

To this purpose, we modify the model by Handcock et al. (2007) in two main directions. First, the network we consider is undirected rather than directed, with valued (commuting flows) rather than binary ties. The second modification is a major (conceptual) extension, in that we allow for the actual spatial structure as well as a latent structure: nodes (municipalities) are positioned in a three-dimensional space structure, where two coordinates are the actual geographical coordinates of municipalities, and the third one is a latent variable.

The role of the latent variable is to allow for unaccounted features—such as the quality of the roads, the presence of railways, physical barriers (rivers, mountains), or socio-economic differences between the units—potentially affecting the relations among municipalities. The third dimension can be then interpreted as being a third coordinate; thus, its effect is to augment the bi-dimensional space in which the network is embedded in such a way that the connections are better explained by the distances between nodes in the augmented space than they were by the distances in the original bi-dimensional space. Moreover, some simulations along the paper suggest that the latent variable is susceptible to capture the effect of unobserved features and include that information in clusters determination, thus improving the clustering procedure.

The procedure in Handcock et al. (2007) relies on the idea that if a fictitious spatial structure of the nodes that explains the network relation can be obtained, then it is reasonable to use such a spatial structure to cluster the nodes. In practice, this means that the issue of obtaining clusters based on the relationships between the nodes is solved by performing clustering in a Euclidean space, which is a (more) standard task. In particular, this is done by specifying a Gaussian mixture distribution for the coordinates, which will then be used to identify the clusters.

Our setting is different, in that our nodes, being geographical entities, have a spatial structure, which is relevant for explaining the connections. The idea, then, is to combine the actual spatial structure with a latent structure. The addition of the third dimension improves the fit of the model: the distance in the three-dimensional space better describes the existing connections between the nodes. A linear component is also added to allow for the populations of the municipalities. Also, in our case, the model is completed by specifying a Gaussian mixture distribution for the coordinates: this model component allows us to obtain a clustering of the municipalities based on

the flows. The model is estimated on real data from the Friuli Venezia Giulia region using the Bayesian approach.

2 Data

We consider data on the number of commuters daily travelling either by train, bus, or private cars between municipalities in Friuli–Venezia Giulia. The flow between two municipalities is the result of combining direct observations on daily mobility patterns recorded in two different surveys carried out by the regional administration in 2010 (for railways and buses) and 2005 (for private cars). With respect to the more conventional data used in previous analyses, mainly referred to individual travel-to-work mobility from Population Census data (Franconi et al., 2017), the observed flows represent multipurpose travels. As a consequence, these data provide information on geographical patterns related to flows originated by other socio-economic activities, such as health and education service provision, consumption, and so on.

In the data set provided by the Regione Friuli Venezia Giulia, flows are reported regardless of direction, meaning that the total number of persons travelling between the two municipalities in a day is observed, and the origin and destination are not recorded. As there are 218 municipalities, we have an available sample of 23,653 ($218 \times 217/2$) flows, whose empirical distribution function is represented in Figure 2. A high number of zero flows (79%, 18,764 observations) and a relatively high

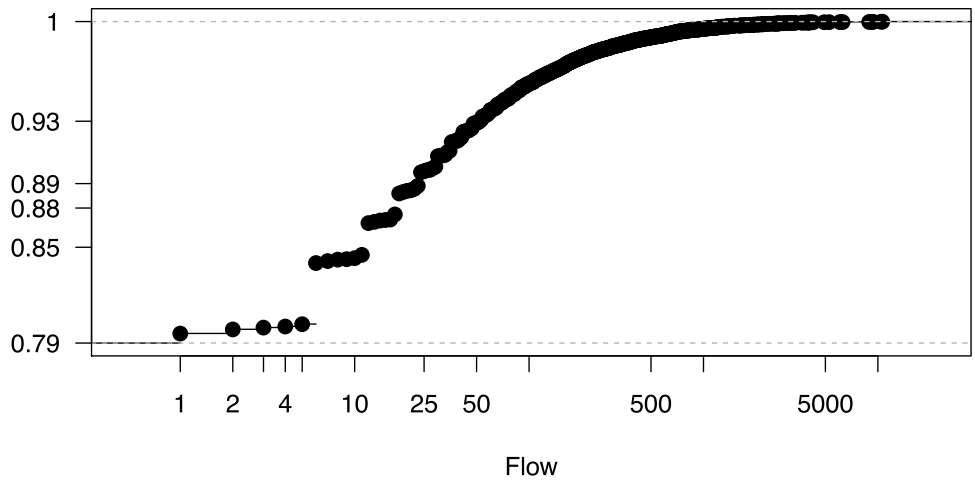

Figure 2. Empirical distribution function of flows (on log scale).

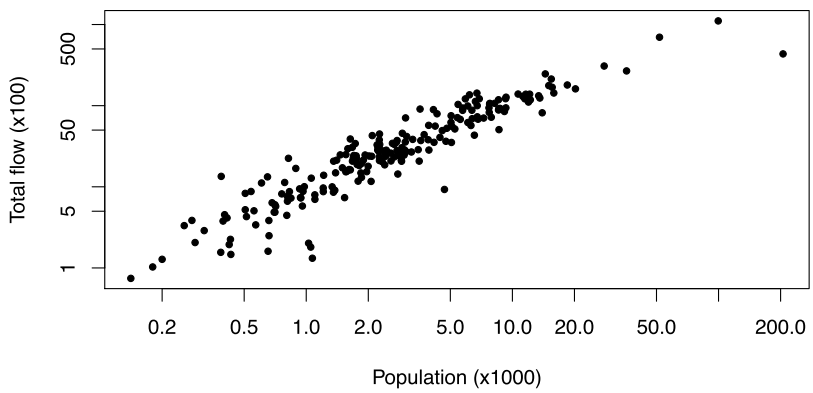

Figure 3. Total flows referred to a municipality versus the population of that municipality.

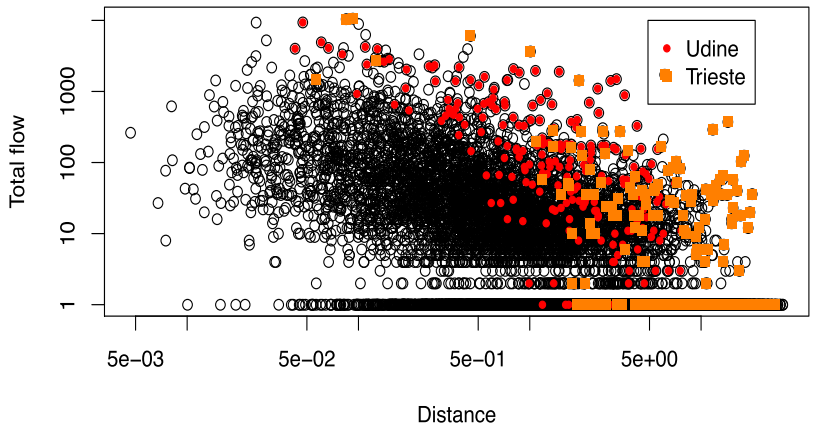

Fig. 4. Flows versus distance, with flows involving the two biggest municipalities: Trieste (orange) and Udine (red).

number of low values can be noted. The concentrations on some values, in particular the multiples of 6, remain unexplained, but they are not relevant for the conclusions according to a sensitivity check. It is worth noting that the concentration on these values occurs only in the data for private cars. The flows are positively related to the population, as may be easily seen in Figure 3, where the total flow involving a municipality is plotted against the population of that municipality, as observed in 2010 (coherently with the observation period of data on the commuting flows).

It is worthy to note that the distribution of the populations of the 218 municipalities is highly skewed (mean 5,669, median 2,398). The bigger municipalities are Trieste (205,535 inhabitants), Udine (99,627), Pordenone (51,723), Gorizia (35,798), Monfalcone (27,877), and Sacile (20,277), of the remaining 212, 47 have less than 1,000; 108 between 1,000 and 5,000. Thus, looking at the flows involving a municipality with a population of less than 1,000 (2,000) inhabitants, the percentage of zero flows is 92.2% (87.7%), while among the flows involving municipalities with both less than 1,000 (2,000) inhabitants, the share of zero flows is 94.2% (92.5%). In particular, there appears to be a linear relationship between the logarithms of the flow and of the population.

We consider a network whose nodes are geographical entities (municipalities), so a spatial structure does in fact exist, and it is clear even from very preliminary analysis (see Figure 4) that geographical distances are of great relevance in explaining the flows. As seen in the figure, the flows are also negatively related to the distance between the municipalities. Again, the trend appears to be reasonably well described by a straight line on the log-log scale.

3 Methods

3.1 Latent position models

Let us assume, we have a $n \times n$ *sociomatrix* Y, with entries y_{ij} denoting the value of the binary relation from node i to node j, along with other possible covariates X. Then, $y_{ij} = 1$ if this relationship exists; 0, otherwise. We can then follow a conditional independence approach as proposed by Hoff et al. (2002), assuming that the presence/absence of a tie between two nodes is independent of all the other ties, conditionally on the latent positions in the social space

$$P(Y \mid Z, X, \omega) = \prod_{i \neq j} P(y_{ij} \mid z_i, z_j, x_{ij}, \omega), \tag{1}$$

where x_{ij} are the observed pair-specific characteristics, whereas ω and (z_i, z_j) are a parameter-vector and the pair of unobserved latent positions in the latent social space, respectively: both these quantities need to be estimated. A typical choice is to assume a logistic regression model in which the probability of a tie depends on the Euclidean distance between nodes i and j and other relevant covariates as in Hoff et al. (2002) and Handcock et al. (2007)

$$\eta_{ij} = \text{logit}(P(Y_{ij} = 1 \mid z_i, z_j, x_{ij}, \delta, \theta)) = \delta^{\mathrm{T}} \log x_{ij} - \theta \log\left(\|z_i - z_j\|^2\right). \tag{2}$$

This specification is flexible and allows us to replace the Euclidean distances $\|z_i - z_j\|$ by any other arbitrary distance satisfying the triangular inequality. To represent clustering, a common choice (Banfield Raftery, 1993; Handcock et al., 2007) is to assume the latent positions z_is $\in \mathbb{R}^2$ as drawn from a mixture of multivariate Gaussian distributions

$$z_i \sim \sum_{g=1}^{G} \pi_g \mathcal{N}_d(\mu_g, \sigma_g^2 I_d), \tag{3}$$

where π_g is the mixing probability that a node belongs to group g, such that $\pi_g \geq 0$ and $\sum_{g=1}^{G} \pi_g = 1$, and I_d is the $d \times d$ identity matrix. Equation (3) implies spherical covariance matrices so that the likelihood is invariant to rotations of the latent social space.

3.2 Hybrid latent position model

In equations (2)–(3), the positions are unobserved in the social space and need to be estimated with suitable techniques. However, in many applications, the distances could be directly observed and plugged into a model specification similar to equation (2) as fixed quantities. Let us assume that equation (2) holds. In equation (3), the variable z_i is a d-dimensional latent process modelled as a Gaussian mixture. Thus, there is no true spatial information; rather, each node is assigned a position in a fictitious space. For our purposes, the observed geographical information matters and could be worth including in the final model. However, a complete drop of latent features may have detrimental implications in the final solutions: Equations (2) and (3) with $z_i \in \mathbb{R}^2$ equal to the longitude and latitude of the ith municipality would, in fact, constitute a reasonable model specification for the phenomenon we are studying; however, such a model would not help in determining clusters. Our proposal is to use then a *hybrid* latent position model, by using both actual and latent coordinates. In other words, we propose to combine the two approaches by specifying a space that is partially latent, partially observed. We assume $z_i \in \mathbb{R}^3$ and that the first two components are the (observed) latitude and longitude, while the third one is a latent variable, which augments the spatial structure and accounts for hidden features other than latitude and longitude implicitly contained in the bi-dimensional space. The vector of latent variables considered by Handcock et al. (2007) is, then, substituted by a vector comprising the (true) geographical coordinates of the municipalities and a latent variable. The latter can be interpreted as being a third coordinate; thus, the effect of the model is to augment the bi-dimensional space in which the network is embedded in such a way that the connections are better explained by the distances between nodes in the augmented space than they were by the distances in the original bi-dimensional space. Then, equation (3) is replaced by the following:

$$z_i \sim \sum_{g=1}^{G} \pi_g \mathcal{N}_d(\mu_g, \Lambda_d), \tag{4}$$

where Λ_d is a 3×3 diagonal matrix with diagonal elements equal to σ_z^2 for the first two components, and $\psi\sigma_z^2$ for the latent variable, with $\psi > 0$ acting as a scaling parameter. Equation (4) is similar to (3), although it has a different interpretation, as it is partially a model for observed variables (first two components) and partially for an unobserved variable (the latent third component). In a sense, we can interpret this as if there was a third coordinate that is a missing value for all nodes and so it is estimated. Handcock et al. (2007) consider spherical covariance matrices for each group due to the invariance of their likelihood to rotations of the latent social space. In our framework, the Euclidean space is not latent anymore, since we have the geographical dimensions clearly identified.

3.3 Introducing count data relationship

To extend the binary-data relationships model (2) to a count-data relationship model, such as the number of commuting flows between two given municipalities, we assume a zero-inflated distribution for the response $\mathscr{Y}_{ij}$ and the flow between two municipalities i and j. Then, let $\mathscr{Y}_{ij} = Y_{ij}S_{ij}$ where the r.v. S_{ij} follows a log-normal distribution:

$$\log(S_{ij} + 0.5) \mid z_i, z_j, x_{ij}, \beta, \gamma \sim \mathcal{N}(\beta^{\mathrm{T}} \log x_{ij} - \gamma \log(\|z_i - z_j\|^2), \sigma_y^2), \tag{5}$$

where $i < j$ and x_{ij} is the vector $(1, x_{ij}^{(1)}, x_{ij}^{(2)})$, where $x_{ij}^{(1)}$ ($x_{ij}^{(2)}$) is the smallest (biggest) observed population of the municipalities i and j; consequently, $\beta, \delta \in \mathbb{R}^3$. From a substantial point of view, the magnitude of the flow between the municipalities i and j depends on their populations and on the distance in the z-space (one expects $\beta_2, \beta_3 > 0$ and $\delta_2, \delta_3 > 0$, and $\gamma, \theta > 0$). In order to interpret the 'third coordinate' in the context of commuting flows, consider that even if we assume that, aside from populations, only the time to travel is relevant in explaining the flow intensity, the distance as calculated from longitudes and latitudes of the barycentres of the municipal territories would be an inadequate measure of distance. In fact, it would neglect unmeasured factors, such as the quality of the roads, the presence of railways, physical barriers (rivers, mountains), or socio-economic differences between the units; these would, then, be allowed by the latent variable.

To complete the model, we specify scarcely informative priors for the other parameters, in particular $\pi \sim \text{Dirichlet}(1_G)$; $\mu_g \sim \mathcal{N}_3(0, 10^3 I_3)$, $g = 1, \ldots, G$; $\psi \sim \mathcal{N}^+(0, 10)$; $\sigma_z^2 \sim \text{invGamma}(10^{-3}, 10^{-3})$; $\beta_j, \delta_j \sim \mathcal{N}(0, 10^6)$, $j = 1, 2, 3$; $\gamma, \theta \sim \mathcal{N}^+(0, 10^6)$, where $\mathcal{N}^+$ denotes the half-normal distribution. Note that this is the standard choice for these kinds of models with a hierarchical structure, similar to what is presented by Handcock et al. (2007).

We introduce the new variable K_i, equal to g if the ith municipality belongs to the gth group. Let the symbol $[\,\cdot\,]$ denote parameters that are not explicitly specified in the following formulae. Then, the full conditional posterior distributions are

$$z_{i3} \mid K_i = g, [\cdot] \propto \phi(z_{i3}; \mu_{g3}, \psi\sigma_z^2) P(\mathscr{Y}, z_{i1}, z_{i2} \mid X, \beta, \delta, \gamma, \theta, z_{i3}), \quad i = 1, \ldots, n \tag{6}$$

$$\beta_j \mid z, [\cdot] \propto \phi(\beta_j; 0, 10^6) P(\mathscr{Y}, z_{i1}, z_{i2} \mid X, \beta, \delta, \gamma, \theta, z_{i3}), \quad j = 1, \ldots, 3 \tag{7}$$

$$\pi \mid [\cdot] \sim \text{Dirichlet}(1_G + m) \tag{8}$$

$$\mu_g \mid [\cdot] \sim \mathcal{N}_3\left(\frac{m_g \bar{z}_g}{m_g + \psi\sigma_z^2/10^3}, \frac{\psi\sigma_z^2}{m_g + \psi\sigma_z^2/10^3} I\right), \quad g = 1, \ldots, G \tag{9}$$

$$\sigma_z^2 \mid [\cdot] \sim \text{invGamma}\left(10^{-3} + \frac{3n}{2}, 10^{-3} + \frac{3n}{2}\right) \tag{10}$$

$$P(K_i = g \mid [\cdot]) = \frac{\pi_g \phi(z_{i3}; \mu_{g3}, \psi\sigma_z^2)}{\sum_{r=1}^{G} \pi_r \phi_d(z_{i3}; \mu_r, \psi\sigma_z^2)}, \quad i = 1, \ldots, n, \quad g = 1, \ldots, G, \tag{11}$$

where

$$m_g = \sum_{i=1}^{n} I_{[K_i = g]},$$

$$\bar{z}_g = \frac{1}{m_g} \sum_{i=1}^{n} z_i I_{[K_i = g]},$$

where $I_{[A]}$ is the event indicator, which equals 1 if A is true, and zero otherwise. In the formulae above, $\phi(\cdot\,; \mu, \sigma^2)$ is the one-dimensional normal density and $P(\mathscr{Y}, z_{i1}, z_{i2}|X, \beta, \delta, \gamma, \theta)$ is the joint density for the flows $\mathscr{Y}$ and the two observed geographical coordinates z_{i1}, z_{i2}.

4. Results

4.1 Bayesian estimation via Gibbs sampling

Estimates from the model defined by equations (2), (4), and (5) were obtained by Markov Chain Monte Carlo (MCMC) methods (Gelman et al., 2004) using JAGS (Plummer, 2003) in R (R Core Team, 2018) via the rjags package (Plummer, 2018)—it is to be noted that we did not use the dnormmix procedure in JAGS to define the mixture of normal distribution; we rather used an additional parameter for group membership.

We estimated our model's parameters by running three chains, each consisting of $K = 4000$ iterations, and we monitored the chains' convergence through Gelman–Rubin statistics (Gelman Rubin, 1992). The internal JAGS algorithm proceeds as follows to sample from the full conditionals (6)–(11):

Step 1: An automatic simplification procedure to figure out if the two full conditional probabilities (6) and (7) for z_{i3} and β_j, respectively, can be reduced to a known statistical distribution. In the case, this reduction is not possible, and other techniques are used. A possibility is to code a Metropolis–Hastings step within the Gibbs sampling to sample $z_{i3}^{(t+1)}$ and $\beta_j^{(t+1)}$, respectively:

(a) At step $t+1$, propose $z_{i3}^{*(t+1)} \sim \phi(z_{i3}^{(t)}, \varepsilon_z^2)$ with probability equal to

$$\frac{\phi(z_{i3}^{*(t+1)}; \mu_g, \psi\sigma_z^2) P(\mathscr{Y}, z_{i1}, z_{i2} \mid X, \beta, \delta, \gamma, \theta, z_{i3}^{*(t+1)})}{\phi(z_{i3}^{(t)}; \mu_g, \psi\sigma_z^2) P(\mathscr{Y}, z_{i1}, z_{i2} \mid X, \beta, \delta, \gamma, \theta, z_{i3}^{(t)})},$$

set $z_{i3}^{(t+1)} = z_{i3}^{*(t+1)}$. Otherwise, $z_{i3}^{(t+1)} = z_{i3}^{(t)}$.

(b) At step $t+1$, propose $\beta^{*(t+1)} \sim \mathcal{N}_d(\beta^{(t)}, \varepsilon_\beta^2 I_3)$ with probability equal to

$$\frac{\phi_3(\beta^{*(t+1)}; 0, 10^6 I_3) P(\mathscr{Y}, z_{i1}, z_{i2} \mid X, \beta^{*(t+1)}, \delta, \gamma, \theta)}{\phi_3(\beta^{(t)}; 0, 10^6 I_3) P(\mathscr{Y}, z_{i1}, z_{i2} \mid X, \beta^{(t)}, \delta, \gamma, \theta)},$$

set $\beta^{(t+1)} = \beta^{*(t+1)}$. Otherwise, $\beta^{(t+1)} = \beta^{(t)}$ ($\phi_d(\cdot\,; \mu, \Sigma)$ denotes the d-dimensional multivariate Gaussian density distribution).

Step 2: Update π, μ_g, σ_z^2, and K_i from expressions (8)–(11).

In the algorithm above, the values for the proposal variance parameters, ε_z^2 and ε_β^2, should be fixed to achieve good performance.

4.2 Posterior estimates

We obtained estimates for different values of G, ranging from 2 to 11. In this section, we illustrate the method presenting a partitioning in nine groups, which is the maximum number of groups actually obtained by our model and has been judged as reasonable on substantive grounds by the regional planners involved in the redaction of the territorial plan of the 'Regione Autonoma Friuli Venezia Giulia'. The discussion on the choice of G from a purely statistical viewpoint is delayed to Section 5.1. It would suffice to note for now that we do not discuss the estimate of G within the model but only consider how model results can be used as a guide to choose a value for G.

It was already noted that the distribution of the flows has a particular form: specifically, the data for private car flows privilege the multiples of six. In order to assess whether this structure has an effect on the results, the model was re-estimated after adding noise to the preferred numbers (specifically, a discrete uniform distribution in $[-5, 0]$ was added). The results were not sensitive to such a manipulation.

Table 1. Summaries of the posterior distributions for the main parameters ($G = 9$). Four thousand Markov Chain Monte Carlo iterations, burn-in: 3,000 iterations, three parallel chains, Gibbs sampling with `rjags` software. The table also reports the effective sample size (neff), the integrated autocorrelation time (ACT), and the Gelman–Rubin statistic $\hat{R}$ for each parameter

					HPD 95%				
	Mean	Median	*SD*	Low	High	neff	ACT	$\hat{R}$	
Y	δ_1	2.11	2.11	0.07	1.97	2.24	866	13.85	1
	δ_2	1.30	1.30	0.04	1.22	1.37	7379	1.63	1
	δ_3	1.79	1.79	0.04	1.70	1.88	1524	7.87	1
	θ	2.06	2.05	0.03	1.99	2.12	2019	5.94	1
S	β_1	3.72	3.72	0.02	3.67	3.77	914	13.12	1
	β_2	0.52	0.52	0.02	0.48	0.55	7043	1.70	1
	β_3	0.87	0.87	0.02	0.83	0.91	3059	3.92	1
	γ	0.88	0.88	0.01	0.85	0.90	4237	2.83	1
	σ_y^2	0.99	0.99	0.02	0.95	1.03	10173	1.18	1
z	σ_z^2	0.16	0.16	0.01	0.14	0.18	5030	2.39	1
	ψ	0.25	0.24	0.04	0.18	0.35	2036	5.89	1

The likelihood of the model is invariant to relabelling of groups; thus, the label-switching problem arises. There are various solutions to the label-switching problem; in the Bayesian-MCMC setting, those solutions that postprocess the chains are particularly convenient (since the issue can be ignored by performing the MCMC and then dealt with later). The postprocessing techniques try to identify groups based on the value of some parameters (for example, group means) so that the first group is the one with the highest mean (it is to be noted that the results may change if another parameter is used). We employ a postprocessing method that, starting from a preliminary clustering of the samples, performs a relabelling with the purpose of obtaining an MCMC sample suitable to infer on the characteristics of the clustering in terms of both probabilities of each unit being in each group and the group parameters. The method is thoroughly described and discussed in Egidi et al. (2018).

In Table 1, posterior summary statistics for the main parameters are shown. As expected, higher flows are associated with higher populations: δ_2, δ_3 are positive, so the probability of a zero is lower, and β_2, β_3 are positive, so the intensity of the flow is higher on average. On the contrary, lower flows are associated with higher distances in the z space (θ, γ positive). The z_is variance σ_z^2 is about 0.16 for the latitude and longitude, whereas $\psi\sigma_z^2 = 0.04$ for the latent variable: the precision for this component is estimated to be four times greater than for the geographical coordinates. Diagnostic measures reported in the table confirm that the MCMC sampling properly converged for all the parameters.

The role, according to the model, of the three variables—populations of the two municipalities and the distance between them—in determining the commuting flow between them is shown in Figure 5. The figure depicts the expected flow (left) and the probability of the flow being nonzero (right) for two hypothetical municipalities as a function of the population of one municipality (x-axis) for selected population sizes of the other municipality and selected distances. There is a very strong expectation of a nonzero flow up to a distance of 50–60 km, at greater distances the probability is appreciably below one.

We depict in Figure 6a the clustering obtained by assigning each municipality to the most likely group, while Figure 6b shows the posterior median value of the third coordinate in a geographical setting and Figure 6c shows its numerical values sorted from the lowest to the highest; additionally, the colours correspond to the groups as depicted in Figure 6a. Further, Figure 6d depicts instead the clustering obtained by assigning each municipality to the most likely group according to the model with no latent coordinate, assuming only the two geographical coordinates, the latitude and the longitude. It is

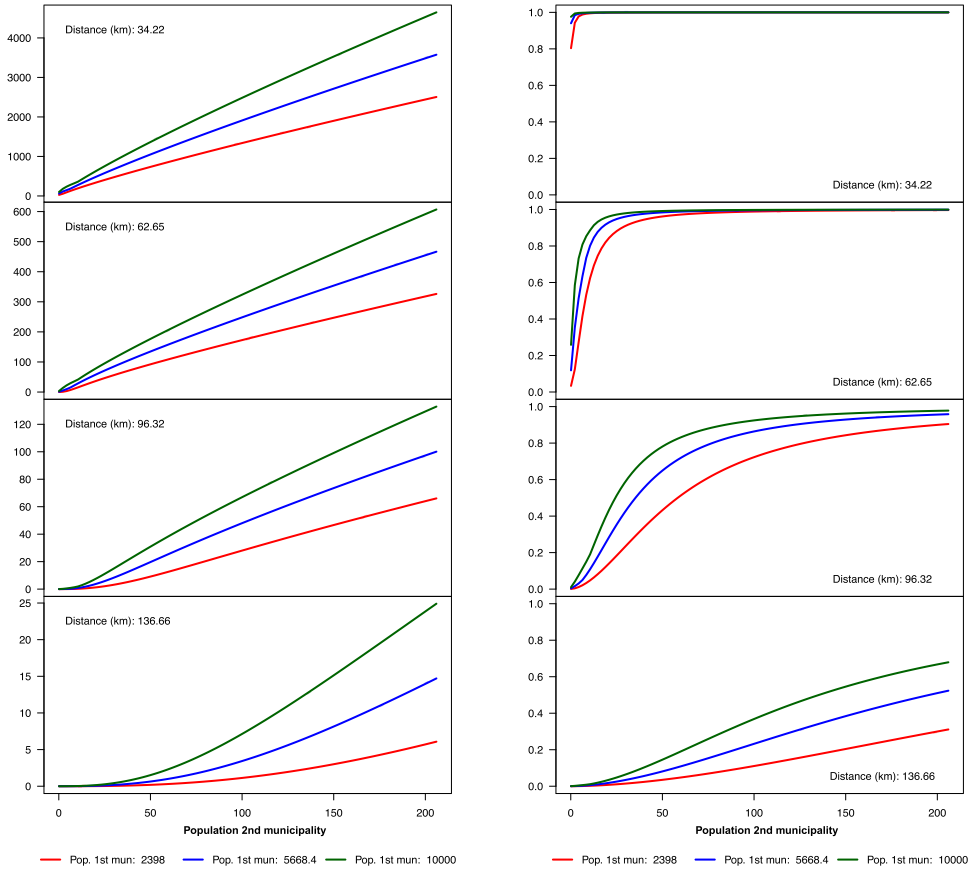

Fig. 5. Model estimates of the expected flow (left) and of the probability of the flow being nonzero (right) between two hypothetical municipalities as a function of the population of one of the two (*x*-axis) for a selection of distances between them (34, 63, 96, 137 km) and a selection of population sizes of the other municipality (2,398, 5,668, 10,000, the first two values are, respectively, the median and the mean of the populations for the 218 municipalities).

evident that this model does not capture the fixed number of groups and the whole clustering complexity. In Section 5.3, we confirm this intuition by computing some numerical comparisons through the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002).

The role of the third coordinate can be clarified looking at the three plots (a–c) in Figure 6. Recall that the third coordinate can amplify the distance between two municipalities: note the jumps in Figure 6c that occur between nearby municipalities reflecting a discontinuity despite their physical contiguity. This occurs in a particularly strong way between the group corresponding to the province of Trieste (light green) and the nearby group (light blue). Going from right to left, a second minor jump is seen, which does not lead to a separation (in the light blue group). Eventually, we notice a gap corresponding to the border of the province of Pordenone (dark blue and teal).

It is interesting to note that the geographical distribution of the third coordinate is broadly in agreement with administrative and physical borders, which are not included in the data (in particular, the municipalities belonging to the provinces of Pordenone and Trieste have, in the three-dimensional space, a greater distance from nearby municipalities belonging to the province of Udine compared with their distance in the actual bi-dimensional space).

4.3 Evidence based on a simulation experiment

The geographical clustering techniques illustrated here could not be directly validated through simulation procedures, since the final clustering solution, which represents the main task of these methods,

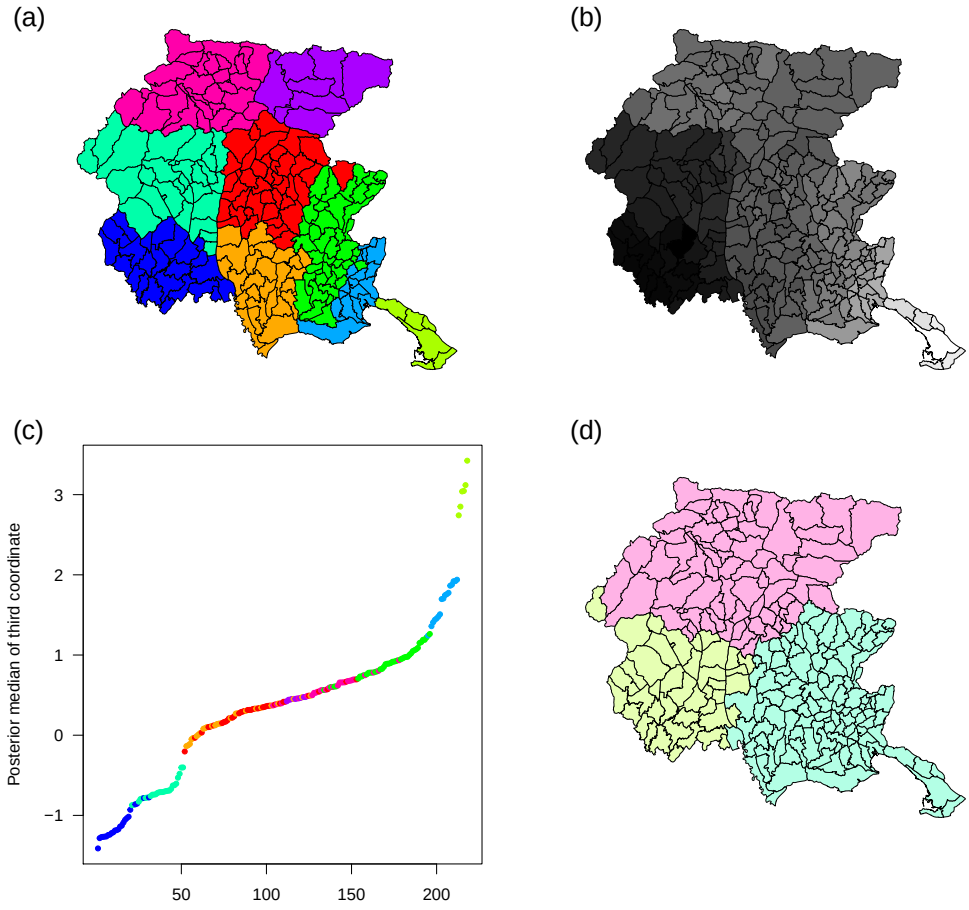

Fig. 6. (a) Groups, determined as the most likely from the model with the latent coordinate; (b) values of the (latent) third coordinate; (c) sorted posterior medians of the third coordinate; (d) groups determined as the most likely according to the model with no latent coordinate.

is a latent feature determined on the ground of a proper statistical model. However, we could assess the impact on the final shape of the clusters arising from distinct input values for some model parameters—in particular those related to the latent component and its relationship with the observed data. This could be another way to highlight how relevant the latent variable is in characterizing the clusters' shape and describing the nodes' connection in a three-dimensional space.

The goal of this simulation is to assess how latent features in a geographical region (for example, possibly, rivers, mountains, lakes, and so on) may have an impact on the commuting flows and on the final clusters' allocation. To achieve this task, we split a bi-dimensional imaginary geographical space in two sub-regions, each associated with a given value for the latent variable z_3.

Thus, our simulation strategy proceeds as follows:

(a) Simulate the longitude and the latitude $z_{\cdot 1}$, $z_{\cdot 2}$ for $n = 60$ geographical units (we can consider them as municipalities) in a $[0, 1] \times [0, 1]$ bi-dimensional space.

(b) Simulate the commuting flows S_{ij} for each pair of municipalities (i, j), $i, j = 1, \ldots, n, i \neq j$, from the following simplified model:

$$\log(S_{ij} + 0.5) | z_i, z_j \sim \mathcal{N}(\beta_0^* - \gamma^* (\|z_i - z_j\|^2), \sigma_y^2),$$

assuming that γ^* may adopt distinct input values, whereas $z_{i,3} = \omega$ if $z_{\cdot 1} + z_{\cdot 2} > 1$, and $z_{i,3} = -\omega$ otherwise. The remaining parameters are set to $\sigma_y^2 = 1, \beta_0^* = 1$.

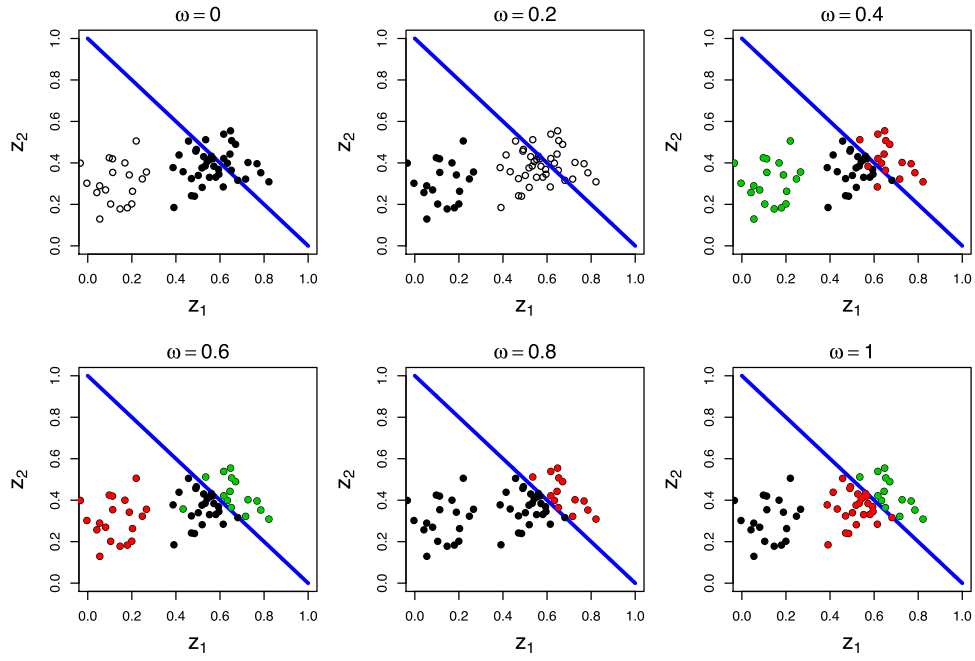

Fig. 7. Simulation study: $n = 60$ simulated nodes in a bi-dimensional space, $G = 3$, $\omega = 0.5$. Each colour represents the final cluster allocation. The solid line $z_{\cdot 2} = 1 - z_{\cdot 1}$ separates the region where $z_{\cdot 3} = \omega$ ($z_{\cdot 2} > 1 - z_{\cdot 1}$) from the region where $z_{\cdot 3} = -\omega$ ($z_{\cdot 2} \leq 1 - z_{\cdot 1}$).

(c) Run the JAGS models with $G = 3$ for different combinations of (ω, γ^*), and determine the final clusters.

We would expect that the higher is ω, and the higher should be the ability of the line $z_{\cdot 2} = 1 - z_{\cdot 1}$ to act as a group's discriminant, since points above the line are far away with respect to the third latent coordinate. Figure 7 depicts some clusters' configurations obtained for distinct values of ω, with γ^* set to 0.5. The blue line $z_{\cdot 2} = 1 - z_{\cdot 1}$ separates the region where $z_{\cdot 3} = \omega$ ($z_{\cdot 2} > 1 - z_{\cdot 1}$) from the region where $z_{\cdot 3} = -\omega$ ($z_{\cdot 2} \leq 1 - z_{\cdot 1}$). As may be noted, the higher is the value for ω, the better the ability of the latent variable to separate and define the final clusters: when ω is relatively high, the third variable is extremely helpful in determining $G = 3$ nonoverlapping groups. Figure 8 shows a similar experiment considering $\omega = 0.5$ and letting γ^* assume some input values. The actual groups are discovered more easily if the value of γ is not too low.

The simulation exercise presented, although limited, suggests that the latent variable is susceptible to capture the effect of unobserved features and include that information in clusters determination, thus improving the clustering procedure.

5 Model selection criteria and diagnostic checks

5.1 Number of groups

As suggested by Handcock et al. (2007), choosing the number of clusters may be framed as a model selection problem. One common measure to compare Bayesian models is the DIC, proposed by Spiegelhalter et al. (2002), and based on the trade-off between the fit of the data to the model and the corresponding complexity of the model. Denoted by $D(\theta) = -2 \log L(\theta; y)$, the deviance for a generic model with data y, parameter(s) θ, and likelihood L, the posterior mean deviance is $\bar{D} = E_{\theta|y}[D(\theta)]$, while the 'effective number of parameters' is $p_D = \bar{D} - D(E_{\theta|y}[\theta])$. Then, DIC is defined as a sum between a 'goodness of fit' measure and a 'complexity' measure

$$\text{DIC} = \bar{D} + p_D.$$

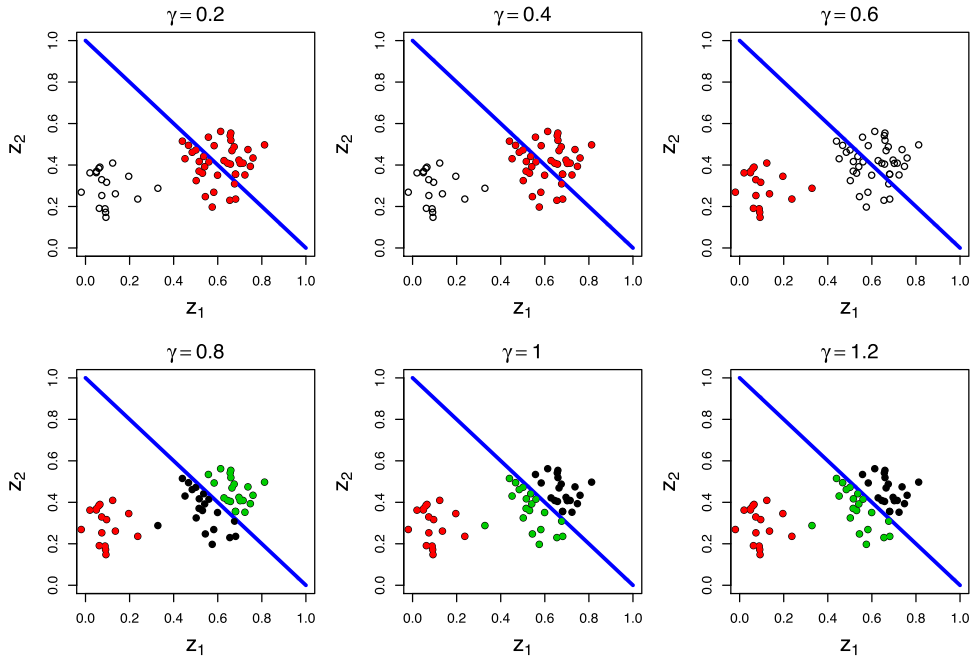

Fig. 8. Simulation study: $n = 60$ simulated nodes in a bi-dimensional space, $G = 3$, $\omega = 0.5$. Each colour represents the final cluster allocation. The solid line $z_{\cdot 2} = 1 - z_{\cdot 1}$ separates the region where $z_{\cdot 3} = \omega$ ($z_{\cdot 2} > 1 - z_{\cdot 1}$) from the region where $z_{\cdot 3} = -\omega$ ($z_{\cdot 2} \leq 1 - z_{\cdot 1}$).

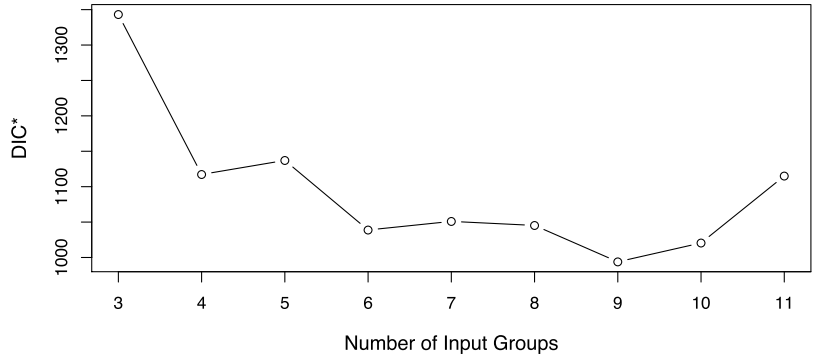

Fig. 9. Rescaled DIC values obtained by fitting the model (1)–(3) under different choices for the number of groups G.

The lower the DIC, the better is the model supported by the data. It is difficult to say what would constitute an important difference in DIC; as a rule of thumb, differences of more than 10 might definitely rule out the model with the higher DIC. As a matter of interpretability, we rescaled the DIC values using the following rule: $\text{DIC}^* = \text{DIC} - 4.73 \times 10^8$, since the DIC values turned out to be extremely large—due to the enormous amount of parameters—and difficult to interpret: the values for the above DIC^* are plotted in Figure 9 for models with G ranging from 3 to 11. According to this plot, a nonnegligible support in favour of $G = 9$ is expressed. We note that the model never leads to more than 9 groups (if $G > 9$, some groups are empty). Even when $G < 9$ it may happen that less than G groups are actually filled.

However, the choice of the number of groups is not a (fully) data-driven procedure and this is a common feature in geographical partitioning methods. In Feldman et al. (2005), the number of groups (zones) strictly depends on the purpose of the subdivision (of exogenous nature), so the

choice cannot be purely based on the model fit. Constraints such as a minimum or maximum dimension, a minimum or target group cohesion are commonly introduced (Coombes et al., 2012). While DIC can give a relevant guideline to select the models, some further indicators of group cohesion (such as those we introduce in Sections 5.4) may aid in choosing the number of groups.

5.2 Posterior predictive checks

To assess the quality of the model, the methodologies proposed by Gelman et al. (1996) extend classical goodness of fit procedures to the Bayesian settings. These are based on the posterior predictive distribution of hypothetical observations y^{rep}, whose general formulation is

$$\pi(y^{\text{rep}} \mid y) = \int \pi(y^{\text{rep}} \mid \phi)\pi(\phi \mid y)\, \mathrm{d}\phi, \tag{12}$$

where y are the data, $\pi(y^{\text{rep}} \mid \phi)$ is the distribution conditional on model parameter ϕ and $\pi(\phi \mid y)$ is the posterior distribution. Simulated values from $\pi(y^{\text{rep}} \mid y)$ are easily obtained using an MCMC sample from the posterior distribution.

For the model in Section 3, explicitly allowing for all involved variables, equation (12) becomes

$$\pi(y^{\text{rep}} \mid y, X, z_1, z_2) = \int \pi(y^{\text{rep}} \mid \phi, X, z)\pi(\phi, z_3 \mid y, z_1, z_2, X)\, \mathrm{d}\phi\, \mathrm{d}z_3, \tag{13}$$

where $\phi = (\delta, \theta, \beta, \gamma, \mu, \sigma)$ and $\pi(y^{\text{rep}} \mid \phi, X, z)$ represents the conditional distribution for the hypothetical flows as specified in equations (2), (4), and (5). The components of the variable z need to be dealt with separately due to their different nature, since the geographical coordinates z_1, z_2 are observed covariates, while z_3 behaves like a parameter; then, the results of inference is expressed by the posterior $\pi(\phi, z_3 \mid y, z_1, z_2, X)$.

Then, let $(\phi^{*(k)}, z_3^{*(k)})$, $k = 1, \ldots, K$ be the sample from the posterior distribution obtained through an MCMC procedure, a sample from (13) is obtained through simulations from equations (2), (4), and (5) for each value $(\phi^{*(k)}, z_3^{*(k)})$ (conditional on X, z_1, and z_2). The simulated values $y^{\text{rep}(k)}$, $k = 1, \ldots, K$ are then used to assess the quality of fit by comparing them with the observed values.

We now illustrate the goodness of fit of the model for $G = 9$, which turned out to be the model associated with the lowest DIC. The most obvious predictive check is to compare the predictions of the flows with their actual values. In Figure 10, we depict such a comparison for four municipalities, including the two cities with the largest populations. Posterior predictive 95% credibility

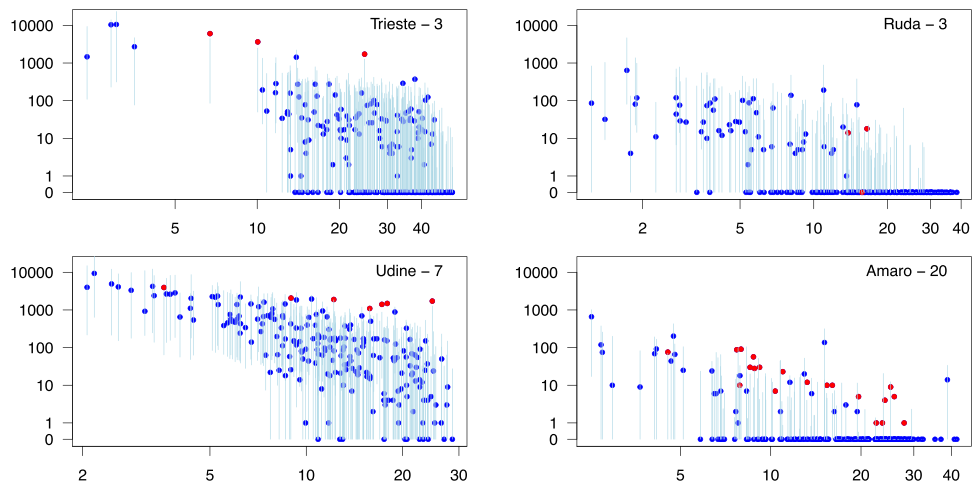

Fig. 10. $G = 9$: posterior predictive 95% credibility intervals (light vertical segments) for each flow associated to the municipalities Trieste, Ruda, Udine, and Amaro, along with the observed values (dark blue dots). The red dots denote those observations falling outside the credibility intervals.

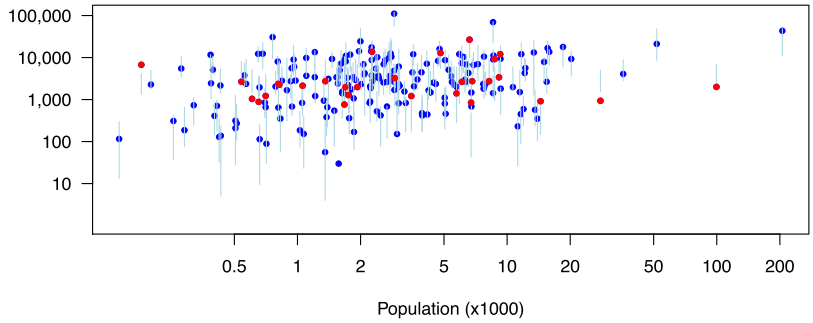

Fig. 11. $G = 9$: posterior predictive 95% credibility intervals (light vertical segments) for the total flows for each municipality along with the observed values (dark blue dots). The red dots denote those observations falling outside the credibility intervals.

intervals for each flow associated with the municipality are plotted against the distance between the two municipalities that the flow refers to. Overall, the number of observations falling outside the credibility intervals is lower than 5%; it is, however, to be said that independence is not guaranteed.

In Figure 11, we compare the observed total flows associated with each municipality and the corresponding flows implied by the model, which are calculated as follows:

$$f_c^{*(k)} = \sum_{j>c} y_{cj}^{\mathrm{rep}(k)} + \sum_{i<c} y_{ic}^{\mathrm{rep}(k)}, \quad c = 1, \ldots, 218.$$

In particular, in Figure 11 we depict, for each municipality c, the mean value of $f_c^{*(k)}$ and its 95% credibility interval (obtained considering the 0.025 and 0.975 quantiles). In 27 out of 218 cases, the credibility interval does not cover the actual observation, which is slightly more than the expected number (11). We also specifically consider the prediction of null flows: on average 86% are correctly predicted.

We also consider Bayesian predictive p-values for a number of statistics T, whose $T(y)$ is the observed value computed on the data y: the mean, maximum, and standard deviation of flows, and the mean, median, and number of nonzero flows. The observed value of the statistic $T(y)$ is compared with the predictive distribution

$$\pi_T(t \mid y) = \int_{y^{\mathrm{rep}} \mid T=t} \pi(y^{\mathrm{rep}} \mid \phi, z) p(\phi, z \mid y)\, \mathrm{d}\phi$$

by means of simulated values $t^{\mathrm{rep}(k)} = T(y^{\mathrm{rep}(k)})$. The comparison is made by graphical methods—for example, plotting the histogram of the simulated $y^{\mathrm{rep}(k)}$—or by synthesizing with the posterior predictive p-value

$$P(T(y^{\mathrm{rep}}) > T(y) \mid y),$$

which is evaluated as

$$\frac{1}{K} \#\{T(y^{\mathrm{rep}(k)}) > T(y)\}.$$

In practice, with respect to the classical testing procedure, the sampling distribution of T is substituted with its Bayesian predictive distribution.

We obtain satisfying results but for the maximum and the median of nonzero flows. In particular, the predicted maximum is high with respect to the observed one (values between 0.94 and 0.97) and the predicted median of nonzero flows is high with respect to observed ones (values between 0.95 and 0.99).

5.3 Comparison with the model having no latent (only geographical coordinates)

A relevant simplification of the model entails using the spatial coordinates only to cluster the municipalities, assuming that this would also reflect the connections measured by the flows.

We then consider the comparison of the proposed model with a simplified model not involving the latent coordinate—that is, a model where equation (4) involves only the geographical coordinates $z_i \in \mathbb{R}^2$. This comparison serves to unveil whether the addition of the latent third coordinate actually improves the model fit and is performed using models with $G = 9$ groups.

First of all, the model having no latent coordinate reports a DIC* of 2753.57, which is enormous if compared with the DIC* of the model with the third latent coordinate, 993.88. This result

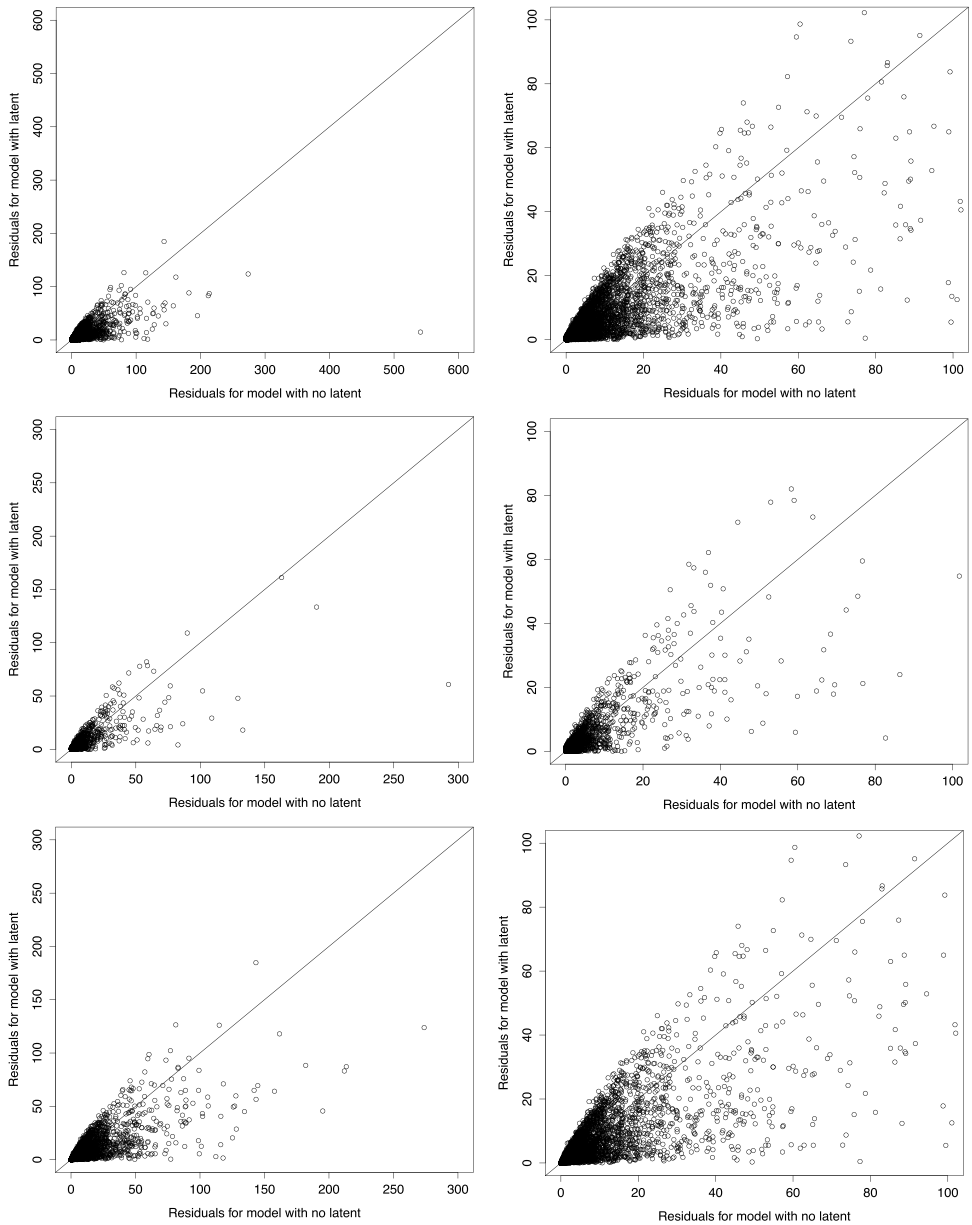

Fig. 12. Comparison of residuals from the model with no latent variable and the model with the latent variable (plots on the right are zoomed—in versions of plots on the left).

constitutes strong evidence in favour of the model with the latent coordinate, adding a further justification to the clustering provided by Figure 6a.

We compute $\hat{y} = \frac{1}{K}\sum_{k=1}^{K} y^{\text{rep}(k)}$ for the model with the latent and the model with no latent: $\hat{y}^{(L)}$, $\hat{y}^{(NL)}$; in addition, in Figure 12, we compare residuals. In particular, the first row depicts

$$\left(\frac{|\hat{y}_{ij}^{(NL)} - y_{ij}|}{y_{ij}+1}, \frac{|\hat{y}_{ij}^{(L)} - y_{ij}|}{y_{ij}+1}\right).$$

The second row considers

$$\left(\frac{|\hat{y}_{ij}^{(NL)} - y_{ij}|}{y_{ij}}, \frac{|\hat{y}_{ij}^{(L)} - y_{ij}|}{y_{ij}}\right)$$

for i, j such that $y_{ij} \neq 0$. The third row

$$\left(|\hat{y}_{ij}^{(NL)} - y_{ij}|, |\hat{y}_{ij}^{(L)} - y_{ij}|\right)$$

for i, j such that $y_{ij} = 0$. It is then apparent that the model with the latent variable has smaller residuals on average.

This constitutes an empirical confirmation of the fact that the geographical distance does not fully reflect the determinants of the commuting flows of the network, thus making the inclusion of the third variable a relevant improvement to the model for the reasons outlined at the end of Section 3.

5.4 Groups' cohesion

Let us then consider the cohesion of the groups, for which an overall measure is the ratio between the average internal flow to average flow depicted in Figure 13 for different MCMC runs indexed by the (effective) number of groups. This provides little help, perhaps advising against 2,3,4.

Finally, we consider the percentage of internal flows to total flows for each group for different MCMC runs indexed by the (effective) number of groups (Figure 14). We note that maximum cohesion is high when $G = 2$, while minimum cohesion is about the same for each G.

6 Concluding remarks

Several approaches to obtain geographical partitions, which take into account the relationships between territorial units measured by the commuting flows, have been considered in the literature, ranging from classical deterministic methods based on, more or less, efficient agglomerative clustering procedures to proper (stochastic) statistical models.

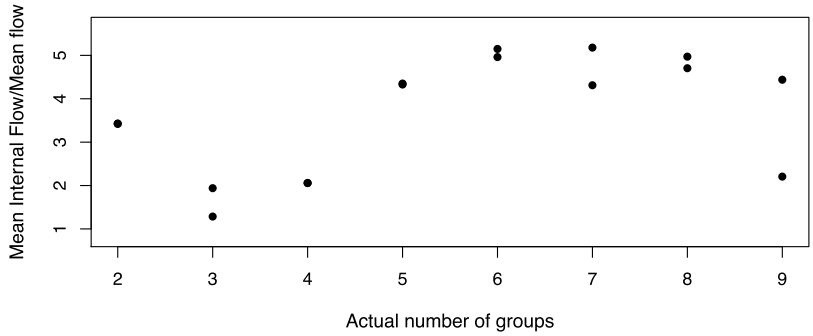

Fig. 13. Ratio between the average internal flow to average flow for different Markov Chain Monte Carlo runs indexed by the (effective) number of groups.

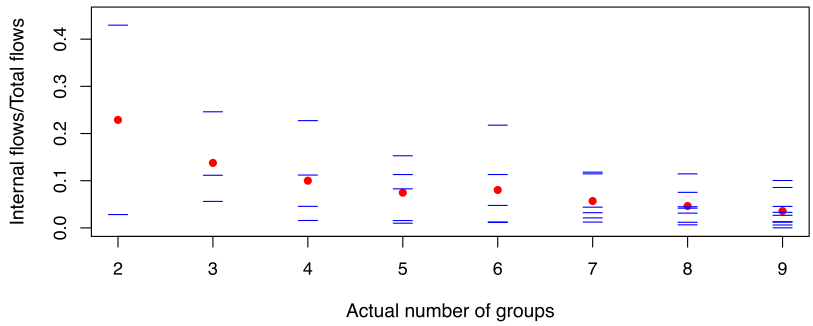

Fig. 14. Percentage of internal flows to total flows for each group for different MCMC runs indexed by the (effective) number of groups: each dash represents the cohesion in one of the groups; the dots represent the average cohesion.

The approach presented here relies on a model-based procedure and seems very promising. It proves useful, since the fit which is obtained using geographical distance alone is improved by adding the third (latent) coordinate, thus obtaining a distance in a three-dimensional space that better describes the existing connections between the nodes.

It is worth noting that the 'homegeneity' or self-containment of the final groups suggested by the model with the latent variable represents a nice and desirable result. Although self-containment is not explicitly allowed for in the model, the groups that are formed are, on this respect, better than the groups obtained by self-containment measures. Moreover, the introduction of the latent variable, accounting for unmeasured factors (for example, the quality of the roads, the presence of railways, physical barriers, and so on) which can explain the commuting flows among municipalities is useful to avoid partitions obtained by agglomerative procedure only based on flows that can be impractical from the perspective of service provision or accessibility. The quality of the results can be measured in terms of goodness of the fit of the model—through a pure Bayesian model selection procedure such as DIC— and the internal cohesion of the groups. Thus, the choice of the number of groups is less ambiguous and arbitrary with respect to a choice driven by a threshold or a tuning parameter (Casado-Dáaz et al., 2017). A relevant extension of the proposed model may be considered to allow for directed flows.

It should also be mentioned that a possible drawback of the model rests in the computational complexity of the procedure, which is higher compared with methods based on agglomerative clustering. As the number of territorial units gets larger, the Bayesian computations involved could become unfeasible. Nonetheless, since the partitions identified by the model are geographically connected, a possible solution is to preliminarily work on sub-matrices of reasonable size extracted from the main matrix of flows. In this first stage, partitions of subsets of units are obtained and a further analysis can be carried out on units that lie on the border of different areas obtained by analysing each sub-matrix.

With regard to the practical issue that inspired our work, the territorial governance plan of the region Friuli Venezia Giulia (delibera/ruling 1890, 31 October 2012) put forward a subdivision of the regional territory in 11 functional areas—'Sistemi territoriali locali (STL)'—combining (in an informal way) our model results and heuristic considerations on homogeneity among municipalities. Comparing the regional STL and the original groups identified by our procedure (Figure 15a), a substantial agreement in the areas identified in the north (mountain) and in the west (province of Pordenone) can be noted. The central area, which is highly interconnected, is where the two partitions are more differentiated and a finer partition to better address regional planning policies is proposed in the governance plan.

Moreover, in December 2014 (Regional law 26/2014), the Region Friuli Venezia Giulia also set up 18 'unioni territoriali intercomunali' (UTI), or unions of municipalities delineated at hand according several and different criteria (contiguity of the municipalities included in each UTI, homogeneity with respect to several socio-economic, cultural, environmental, among others aspects, the size of the resulting population, and so on). The UTI, which are meant to coordinate the administrative functions of the municipalities are shown in Figure 15b, and compared with the nine groups determined in our analysis. It is noted that three UTI coincide with our groups; six of them are contained in one group; seven are almost contained in a group. The remaining two

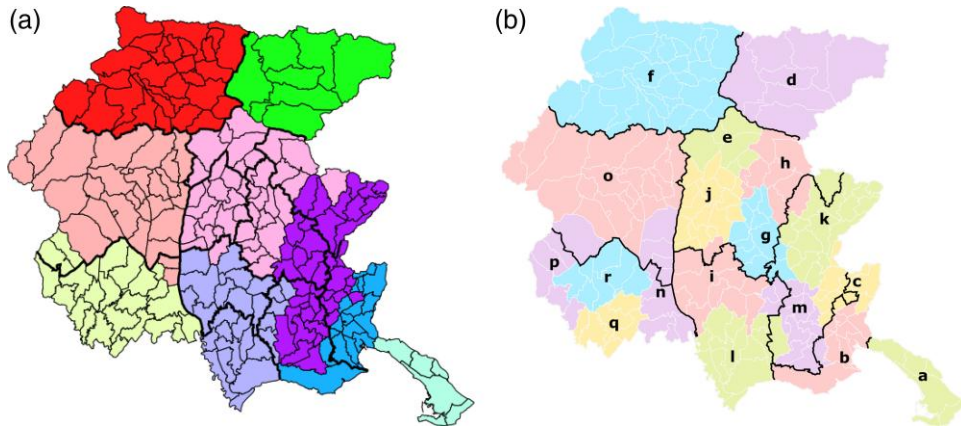

Fig. 15. (a) Comparison of Sistemi territoriali locali, and groups identified by our method, coded by colour; (b) comparison of unioni territoriali intercomunali, coded by colours and letters, and the groups identified by our method.

UTI, n and c, imply a markedly different subdivision as they are split in half according to our grouping. UTI were created in 2014 and were substituted from 1 December 2021 by four 'Enti di decentramento regionale', which coincide with the former provinces (abolished in 2017): Trieste (group 1), Pordenone (groups 7, 8), Udine (groups 4, 5, 6, 9, part of group 3), and Gorizia (group 2 and part of group 3).

These various subdivisions adopted by the regional government, although based on different criteria, underline the importance of functional geographies to complement the established system of the administrative boundaries in a territory as well as the proposal of novel methods to address the statistical problem.

Acknowledgements

The authors are particularly grateful to the Regione Friuli–Venezia Giulia for providing us with the data, and in particular, to Andrea Battiston and his collaborators at 'Direzione centrale infrastrutture, mobilità, pianificazione territoriale e lavori pubblici'.

Data availability

Given their size, the data, code, and final results used in this paper are available upon request from the authors.

Conflict of interest: The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding statement

The authors received no financial support for the research, authorship and/or publication of this article.

References

Banfield J. D., & Raftery A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*(3), 803–821. https://doi.org/10.2307/2532201

Casado-Dáaz J. M., Martínez-Bernabéu L., & Flórez-Revuelta F. (2017). Automatic parameter tuning for functional regionalization methods. *Papers in Regional Science*, *96*(4), 859–879. https://doi.org/10.1111/pirs.12199

Casado-Diaz J., & Coombes M. (2011). The delineation of 21st century local labour market areas: A critical review and a research agenda. *Boletín de la Asociación de Geógrafos españoles*, *57*, 7–32.

Celik H. M., & Guldmann J.-M. (2007). Spatial interaction modeling of interregional commodity flows. *Socio-Economic Planning Sciences*, *41*(2), 147–162. https://doi.org/10.1016/j.seps.2005.10.003

Chakraborty A., Beamonte M., Gelfand A. E., Alonso M. P., Gargallo P., & Salvador M. (2013). Spatial interaction models with individual-level data for explaining labor flows and developing local labor markets. *Computational Statistics & Data Analysis*, *58*, 292–307. https://doi.org/10.1016/j.csda.2012.08.016

Coombes M., & Bond S. (2008). *Travel-to-work areas: The 2007 review*. Office for National Statistics.

Coombes M., Casado-Diaz J., Martìnez-Bernabeu L., & Carausu F. (2012). *Study on comparable labour market areas: Final research report*. Eurostat.

Coombes M. G., Green A. E., & Openshaw S. (1986). An efficient algorithm to generate official statistical reporting areas: The case of the 1984 travel-to-work areas revision in Britain. *The Journal of the Operational Research Society*, *37*(10), 943–953. https://doi.org/10.1057/jors.1986.163

Daraganova G., Pattison P., Koskinen J., Mitchell B., Bill A., Watts M., & Baum S. (2012). Networks and geography: Modelling community network structures as the outcome of both spatial and network processes. *Social Networks*, *34*(1), 6–17. https://doi.org/10.1016/j.socnet.2010.12.001

Egidi L., Pappadà R., Pauli F., & Torelli N. (2018). Relabelling in Bayesian mixture models by pivotal units. *Statistics and Computing*, *28*(4), 957–969. https://doi.org/10.1007/s11222-017-9774-2

Eurostat (2020). *European harmonised labour market areas - methodology on functional geographies with potential*. Eurostat. https://ec.europa.eu/eurostat/web/products-statistical-working-papers/-/ks-tc-20-002

Farmer C., & Fotheringham A. (2012). Network-based functional regions. *Environment and Planning A: Economy and Space*, *43*(11), 2723–2741. https://doi.org/10.1068/a44136

Feldman O., Simmonds D., Troll N., & Tsang F. (2005). Creation of a system of functional areas for England and wales and for Scotland. Proceedings of the European Transport Conference 2005, PTRC, London (on CD).

Fortunato S., & Hric D. (2016). Community detection in networks: A user guide. *Physics Reports*, *659*, 1–44. https://doi.org/10.1016/j.physrep.2016.09.002

Franconi L., Ichim D., D'Alò M., & Cruciani S. (2017). *Guidelines for labour market area delineation process: From definition to dissemination*. Istat.

Gelman A., Carlin J., Stern H., & Rubin D. (2004). *Bayesian data analysis*. Chapman & Hall/CRC.

Gelman A., Meng X.-L., & Stern H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–760.

Gelman A., & Rubin D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Goldenberg A., Zheng A. X., Fienberg S. E., & Airoldi E. M. (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning*, *2*(2), 129–233. https://doi.org/10.1561/2200000005

Handcock M. S., Raftery A. E., & Tantrum J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(2), 301–354. https://doi.org/10.1111/rssa.2007.170.issue-2

Haynes K., & Fotheringham A. (1985). *Gravity and spatial interaction models*. Reprint. Grant Ian Thrall. WVU Research Repository, 2020.

Hoff P. D., Raftery A. E., & Handcock M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, *97*(460), 1090–1098. https://doi.org/10.1198/016214502388618906

Kim B., Lee K. H., Xue L., & Niu X. (2018). A review of dynamic network models with latent variables. *Statistics Surveys*, *12*, 105–135. https://doi.org/10.1214/18-SS121

Nelson G., & Rae A. (2016). An economic geography of the United States: From commutes to megaregions. *PLoS ONE*, *11*(11), e0166083. https://doi.org/10.1371/journal.pone.0166083

Newman M., & Girvan M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113

OECD (2020). *Delineating functional areas in all territories*. OECD Territorial Reviews. OECD Publishing. https://doi.org/10.1787/07970966-en

Openshaw S. (1977). Optimal zoning systems for spatial interaction models. *Environment and Planning A*, *9*(2), 169–184. https://doi.org/10.1068/a090169

Plummer M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vienna, 20–22 March 2003. p. 1–10

Plummer M. (2018). *rjags: Bayesian graphical models using MCMC*. R package version 4–8.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Roy J., & Thill J. (2013). Spatial interaction modelling. *Papers in Regional Science*, *83*(1), 339–361. https://doi.org/10.1007/s10110-003-0189-4

Smith A. L., Asta D. M., & Calder C. A. (2019). The geometry of continuous latent space models for network data. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *34*(3), 428–453. https://doi.org/10.1214/19-STS702

Spiegelhalter D. J., Best N. G., Carlin B. P., & Der Linde A. Van (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639. https://doi.org/10.1111/rssb.2002.64.issue-4