XXXV CONFERENZA ITALIANA DI SCIENZE REGIONALI

GEOGRAPHICAL PARTITIONING OF SPATIAL NETWORKS THROUGH LATENT MIXTURE MODELS

Francesco Pauli, Nicola Torelli, Susanna Zaccarin[1]

Abstract

We consider a model based clustering technique that directly accounts for network relations between subjects and their position in geographical space. This proposal is motivated by a practical problem: to design administrative structures intermediate between the municipality and the province where the criterion for municipalities aggregation is the existence of a relatively (to population) high commuting flow.
A natural framework to deal with this issue is social network analysis, where municipalities are the nodes and the relationship between the nodes is measured by the commuting flows. In our model the flows are explained by the distances between the nodes in a three-dimensional space where two coordinates are the actual geographical coordinates and the third one is a latent variable. 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 (having discounted for the populations).
The model is estimated on real data from Friuli-Venezia Giulia region using the Bayesian approach.

[1]Dipartimento di Scienze Economiche, Aziendali,
Matematiche e Statistiche 'B. de Finetti'
Università degli Studi di Trieste

1 Introduction

We consider a model based clustering technique that directly accounts for network relations between subjects and their position in geographical space.
This proposal is motivated by a practical problem: to design administrative structures intermediate between the municipality and the province comprising areas which are to some extent self-contained. Data on the flows of commuters travelling between the municipalities of an Italian region, as collected by Istat in 2001 Population Census, are used to determine geographically connected groups of municipalities within which a high mobility exists. Two municipalities are strongly related if a high flow exists between them, where high is meant relative to the populations. (The flows are clearly related to the populations of the municipalities and to the distances between them.)
A natural framework to deal with this issue is social network analysis, where municipalities are the nodes and the relationship between the nodes is measured by the commuting flows. As Daraganova et al. [2012] point out the geographical distance between actors in a network can have a powerful effect on the formation/intensity of a tie between them. Despite this, however, Daraganova et al. [2012] note that spatial models for network data have been seldom used.
One of the exceptions is Hoff et al. [2002], who propose using 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. Handcock et al. [2007] propose using such latent (unobserved) coordinates of nodes in the latent space to perform clustering by modelling them as a gaussian mixture.
We propose an extension of Handcock et al. [2007] approach considering a valued and not binary network. This implies that we must change the model specification from binomial to lognormal with zero inflation. The populations are easily included in the model aside of the spatial structure. The second and major (conceptual) extension is to allow for the actual spatial structure as well as a latent one: nodes are positioned in a three-dimensional space where two coordinates are the actual geographical coordinates and the third one is a latent variable. The addition of the third coordinate improves the fit of the model: the distance in three-dimensional space better describes the existing connections between the nodes.
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 (having discounted for the populations).
The model is estimated using the Bayesian approach. We also propose a comparison of the resulting clusters with "Sistemi Locali del Lavoro" (SLL) by ISTAT (2005).

2 Data

We consider data on the number of commuters travelling between municipalities in Friuli Venezia Giulia (Italy's region in the North-East of the country) as measured by ISTAT using data from the 2001 Population Census.

Flows are measured regardless of direction, meaning that the number of commuters travelling daily between the two cities is observed, the origin and destination are ignored. As there are 218 municipalities, we have then a total of 23653 ($218 \times 217/2$) flows, whose empirical distribution function is represented in figure 1. We note a high number of zero flows (68.8% , 16274 observations) and a relatively high number of low values.
That the flows are positively related to the population is easily seen in figure 2 where the sum of the flows involving a municipality is plotted against the population of that municipality.
As seen in figure 3 the flows are also negatively related to the distance between the municipalities. We consider a network whose nodes are geographical entities (municipalities), so a spatial structure does in fact exist, and geographical distances are of great relevance in explaining the flows.

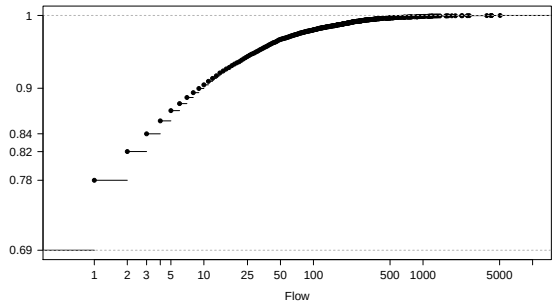

Figure 1: Empirical distribution function of flows (depicted on log x-axis and a truncated y-axis)

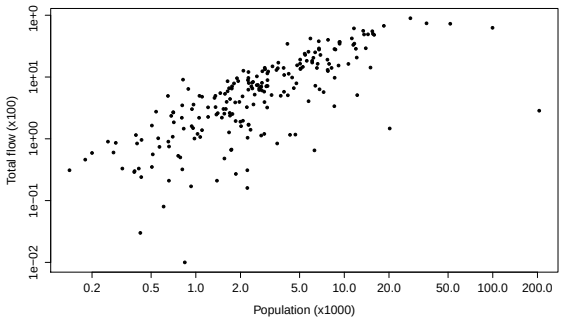

Figure 2: Total flows referred to a municipality versus population of that municipality (log scales)

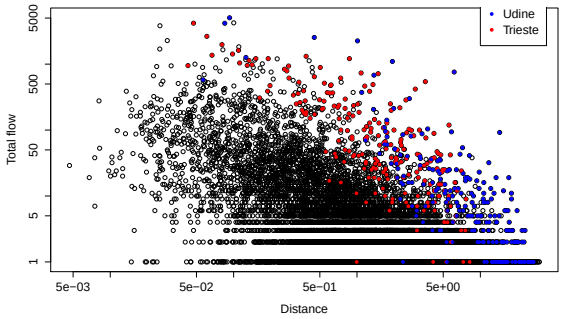

Figure 3: Flows (log) versus distance, with flows involving the two biggest municipalities evidenciated

3 Method

We assume a zero-inflated distribution for the response Y_{ij}, let then $Y_{ij} = V_{ij}S_{ij}$ where V_{ij} is a bernoulli r.v. with

$$\text{logit}(P(V_{ij} = 0|z_i, z_j, x_{ij}, \delta, \theta)) = \delta x_{ij} + \theta \log\left(||z_i - z_j||^2\right) \tag{1}$$

while the r.v. S_{ij} follows a log-normal distribution, that is

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

where $i < j$ and x_{ij} is the vector $(1, x_{ij}^{(1)}, x_{ij}^{(2)})$ where $x_{ij}^{(1)}$ and $x_{ij}^{(2)}$ are, respectively, the populations of the smallest and the biggest among municipalities i and j, consequently, $\beta, \delta \in \mathbb{R}^3$. [2] From a substantial point of view, the magnitude of the flow between the municipalities i and j depends on their populations (one expects $\beta_2, \beta_3 > 0$, $\delta_2, \delta_3 < 0$) and on the distance in the z-space.
Equations (1) and (2) with $z_i \in \mathbb{R}^2$ equal to the longitude and latitude of the i-the municipality would be a reasonable model specification for the phenomenon we are studying, however, such a model would not help in determining clusters.
In the approach by Handcock et al. [2007], on the other hand, the variable z_i is a d-dimensional latent process modelled as a gaussian mixture. So there is no true spatial information, rather nodes are assigned to positions in a fictitious space.
Our proposal is to combine the two approaches by specifying a space which 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 is a latent variable, which augments the spatial structure. In

[2] In Handcock et al. [2007] the existence of a tie from node i to node j is modeled by a logistic regression specification

$$\text{log-odds}(y_{ij} = 1|z_i, z_j, x_{ij}, \beta, \gamma) = \beta^T x_{ij} - \gamma||z_i - z_j||^2 \tag{3}$$

where z represent a d-dimensional (bi-dimensional in the examples in the paper) latent variable and $||\cdot||$ is a suitable norm (Euclidean norm in the examples).

other words, the vector of latent variables considered by Handcock et al. [2007] is 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.
The z_is are assumed to be drawn from a mixture of gaussian distribution

$$z_i \sim \sum_{g=1}^{G} \pi_g MVN_d(\mu_g, \sigma^2 I_d) \tag{4}$$

(in the original specification group specific variances are considered). Equation (4) is identical to Handcock et al. [2007] although it has a different interpretation, as it is partially a model for observed variables (first two components) and partially for un unobserved variable (the latent third component). In a sense, we can interpret this as if there was a third coordinate which is a missing value for all nodes and so is estimated.
We also assume $\pi \sim \text{Dirichlet}(1_G)$; $\mu \sim MVN(0, 10^3 I_3)$; $\sigma^2_\bullet \sim \text{invGamma}(10^{-3}, 10^{-3})$; $\beta_i, \delta_i \sim \mathcal{N}(0, 10^6)$, $i = 1, 2, 3$; $\gamma, \theta \sim \text{half}\mathcal{N}(0, 10^6)$.

4 Results

Estimates were obtained by MCMC [Gelman et al., 2004] using JAGS (Plummer [2003]) in R (R Development Core Team [2011]). (It is to be noted that we did not use the *dnormmix* procedure in JAGS to define the mixture of normal distribution, but rather used an additional parameter for group belonging.)
Proposed estimates are obtained conditional on $G = 12$. In principle G could be estimated as a model parameter in the bayesian framework or selected based on goodness of fit of the model; this issue is not discussed in the present work. The choice of $G = 12$ is driven by the aim of comparing the partition we obtain with ISTAT SLLs, which are 13.
Another issue arising in estimation of the model is that of label switching: the likelihood of the model is invariant to relabelling of groups. We analysed the Markov chain in post processing to check for occurrence of label switching, this check revealed that no switching occurred.
The partition of the municipalities in groups is represented in figure 4. It is to be noted that 10 groups are formed despite the fact that $G = 12$ in model specification, this can occur because a group may be empty. It is tempting to interpret this as an indication that 12 groups are too many, however we must be cautious with such an interpretation since it is possible that if we estimate the model with a greater value of G, we will obtain more non empty groups.
The partition in figure 4 is obtained by associating each municipality to the group it is most likely to belong to, such probabilities are represented in figure 5. There we see that two of the 12 groups nominally present in the model specification are empty. It can also

be argued that for most municipalities there is a high probability associated to only one group, thus suggesting a much stronger relationship with the other municipalities of the same group than with the municipalities belonging to the other groups. If we look at this from a group point of view, we note that most of them include municipalities only with a high probability, which can be loosely interpreted as evidence of a good fit. The main exception to this is given by groups 3 and 5, which might be joined in the same group with a relatively high probability, while the other groups appear to be more clear cut.
The estimated values of the latent coordinate (medians) are shown in fig. 6. It should capture various unobserved factors which affects the strength of the connection between municipalities. For example, if we assume that, aside of populations, only the time to travel is relevant in explaining the flow intensity, there are physical attributes which affect the time to travel between municipalities: the quality of the roads, the presence of railways and so on. It is interesting to note that the estimate increases the distance – in the three dimensional Z-space – between municipalities belonging to the province of Pordenone from the nearby municipalities belonging to the province of Udine.

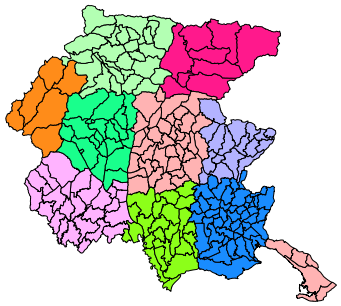

Figure 4: Groups, determined as the most likely

5 Comparison with SLL

It is interesting to compare the partition obtained with the mixture model (MM in what follows) with the one obtained by ISTAT in defining the SLL (Local labour Systems). The latter is a core-based non-hierarchical agglomerative clustering procedure (also called rule-based) with an elaborated linkage criteria where a minimum level of self-containment (minSC) is enforced for all the resulting LMAs ([ISTAT, 1997]).
It is to be born in mind that our model is estimated taking only the FVG municipalities under consideration, while ISTAT SLL are obtained considering the whole country. As a consequence, while the groups we obtain are forced to be within the region boundaries, ISTAT SLL may cross them (That is, a group may join municipalities of FVG and municipalities from Veneto).
The two partitions are represented side by side in figure 7, note that in the SLL parti-

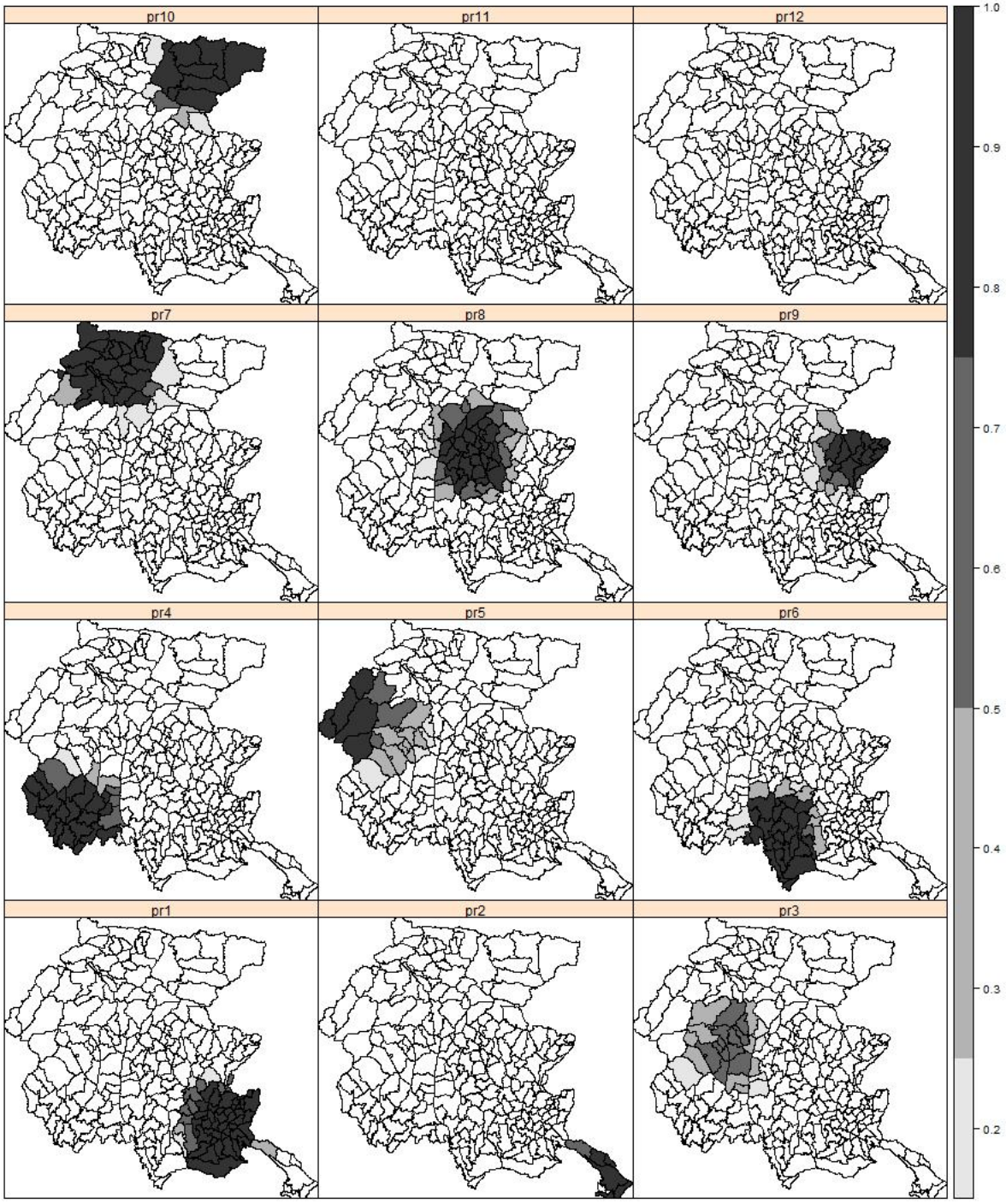

Figure 5: Probability of belonging to groups

Figure 6: Values of the (latent) third coordinate

tion the groups apparently including only one and only four municipalities are actually parts of cross-regional SLLs; also other groups in the right panel of figure 7 include municipalities from other regions (not shown).

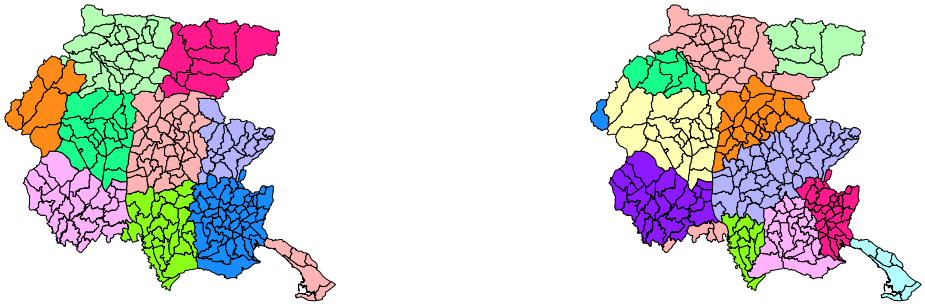

Figure 7: Partitions obtained by the mixture model (left panel) and SLL obtained by ISTAT (right panel)

We compare the two partitions based on self-containment measures, in particular we define, for each group $\mathcal{G}$ including $\#\mathcal{G}$ municipalities

$$c(\mathcal{G}) = \frac{\text{Average internal flows}}{\text{Average external flows}} = \frac{\binom{\#\mathcal{G}}{2}^{-1} \sum_{i,j\in\mathcal{G}, i<j} y_{ij}}{(\#\mathcal{G}(N-\#\mathcal{G}))^{-1} \sum_{i\in\mathcal{G}} \sum_{j\notin\mathcal{G}} y_{ij}}$$

These are pictured in figure 8 as a function of group dimension (number of municipalities) for both the groups obtained by the mixture model (red) and the SLL (blue).
We also compute measures of membership of each municipality to its group, let $\mathcal{G}_i$ be the group to which the i-th municipality belong, then define

- $IF_i = \sum_{j\in\mathcal{G}_i\setminus\{i\}} y_{ij}$ flows to/from i-th municipality from/to municipalities of the same group;

- $EF_i = \sum_{j \notin \mathcal{G}_i} y_{ij}$ lows to/from i-th municipality from/to municipalities of the other groups.

the above two measures are computed for both the partition based on the mixture model, IF^{MM}, EF^{MM} and the SLL, IF^{SLL}, EF^{SLL} and compared in figure 9.
According to this comparison, the partition obtained with the MM model realizea an improvement in terms of all measures of self containment with respect to ISTAT SLL.

6 Concluding remarks

Recently many alternative approaches to obtain geographical partitions which take into account the relationships between territorial units measured by the commuting flows have been considered.
Most of these approaches go beyond the classical deterministic methods based simply on, more or less, efficient agglomerative clustering procedures and introduce a proper (stochastic) statistical model. A modelling strategy that explicitly takes into account the spatial configuration has been proposed (see Chakraborty et al. [2013]). Other approaches are aimed at using computational intensive methods (like genetic algorithm) to efficiently explore the space af all possible spatial partitions in order to find those partitions that maximize a given criterion ([Martínez-Bernabeu et al., 2012]).
The approach presented here relies on a model based procedure and seems very promising. The quality of the results can be measured in terms of goodness of fit of the model. 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. In future work, possible extension of the proposed model (for instance taking into account directed flows) and an application to a wider area using data collected in Italy in the last population census will be considered.
A comparison between these new approaches to identify their relative merits and limitations could be interesting with reference to some specific applications such as the identification of Local Labour Market in Italy with data on the commuting flows collected in the 2011 Population Census.

References

A Chakraborty, MA Beamonte, Alan E Gelfand, MP Alonso, Pilar Gargallo, and Manuel Salvador. Spatial interaction models with individual-level data for explaining labor flows and developing local labor markets. *Computational Statistics & Data Analysis*, 58:292–307, 2013.

Galina Daraganova, Pip Pattison, Johan Koskinen, Bill Mitchell, Anthea Bill, Martin Watts, and Scott Baum. Networks and geography: Modelling community network structures as the outcome of both spatial and network processes. *Social Networks*, 34(1):6 – 17, 2012. ISSN 0378-8733. doi: 10.1016/j.socnet.2010.12.001. URL `http://www.sciencedirect.com/science/article/pii/`

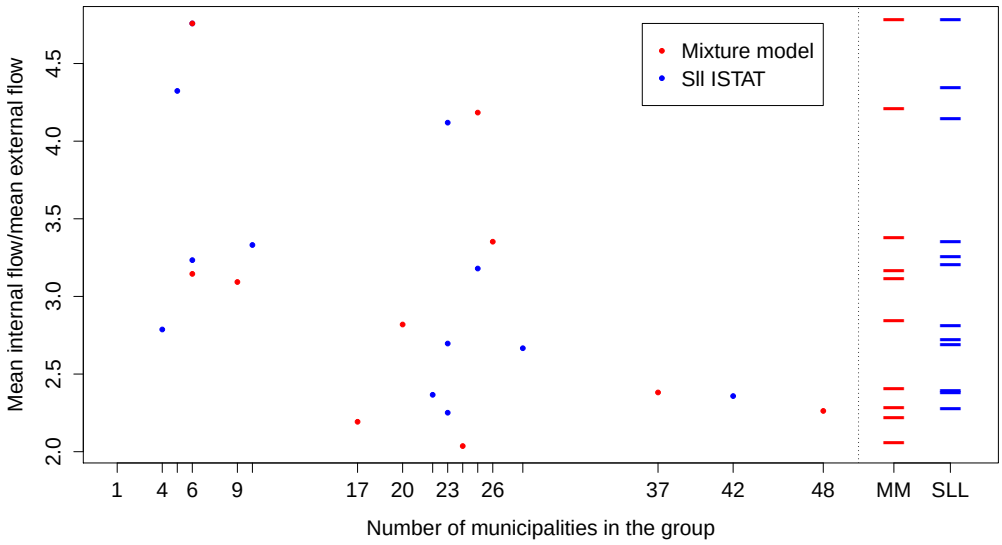

Figure 8: Group self containment measures for the partition obtained by the mixture model (red) and the SLL (blue)

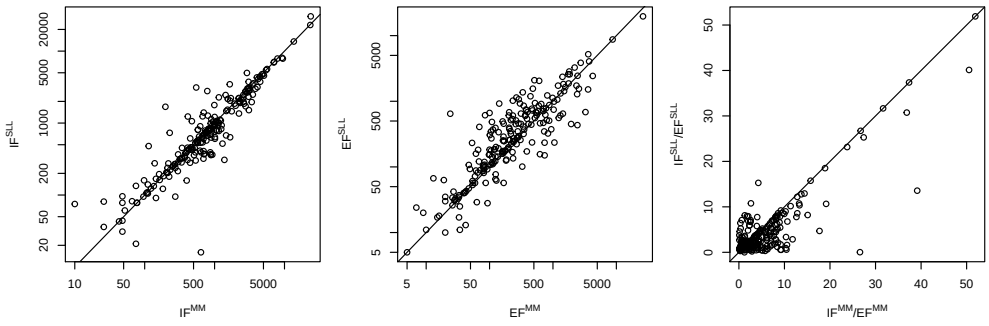

Figure 9: Measures of membership for each municipality in the two partitions

S0378873310000614. ¡ce:title¿Capturing Context: Integrating Spatial and Social Network Analyses¡/ce:title¿.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2004. ISBN 9781584883883. URL http://books.google.it/books?id=TNYhnkXQSjAC.

Mark S. Handcock, Adrian E. Raftery, and Jeremy M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007. ISSN 1467-985X. doi: 10.1111/j.1467-985X.2007.00471.x. URL http://dx.doi.org/10.1111/j.1467-985X.2007.00471.x.

Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):pp. 1090–1098, 2002. ISSN 01621459. URL http://www.jstor.org/stable/3085833.

ISTAT. *I sistemi locali del lavoro 1991.* Istituto Poligrafico e Zecca dello Stato, 1997.

Lucas Martínez-Bernabeu, Francisco Flórez-Revuelta, and José Manuel Casado-Díaz. Grouping genetic operators for the delineation of functional areas based on spatial interaction. *Expert Systems with Applications*, 39(8):6754–6766, 2012.

Martyn Plummer. Jags: A program for analysis of bayesian graphical models using gibbs sampling, 2003.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL http://www.R-project.org/. ISBN 3-900051-07-0.