Soft textual cartography based on topic modeling and clustering of irregular, multivariate marked networks

Mattia Egloff and Raphaël Ceré

Abstract Soft textual cartography is an original approach aimed to study communities on spatially embedded and textually defined complex weighted networks. The present approach relies on the integration of topic modeling and soft clustering procedures. These two aspects can be combined using topic distances, and weighted unoriented networks representing the spatial configuration; their synergy is promising in topic interpretation and geographical information retrieval. This paper proposes an unified formalism, underlining the compatibility of the two aspects, as illustrated on the textual descriptions of the municipalities of the canton of Vaud, Switzerland. It also points to possible extensions and applications of the method, potentially useful for dealing with the ever growing amount of georeferenced textual content.

Key words: Textual Cartography, Community detection, Complex network, Topic-modeling, Soft clustering, Modularity.

1 Introduction

Regional data analysis using complex networks generally involves numerical or categorical information attached to the regions, such as census block values, level intensities or densities. In this paper we use the textual data attached to the regions as geographical information for community detection, an issue in complex network.

These regions can be represented by an irregular or regular unoriented weighted network. The node weights represent the relative size of the regions (e.g. surface of

Mattia Egloff
Department of Language and Information Sciences, University of Lausanne, Switzerland, e-mail: mattia.egloff@unil.ch

Raphaël Ceré
Department of Geography and Sustainability, University of Lausanne, Switzerland, e-mail: raphael.cere@unil.ch

the geographical unit, population or number of words in the description). The spatial relationship between each pair of regions is here represented by a joint probability of selecting a pair of neighbors defining the edge weight and thus are quantifying the importance of the edges of the network.

This paper applies to textual information the method of regional soft clustering proposed by Ceré and Bavaud [4][5], using spatial configuration and features distances in an image segmentation framework (see [20] for a conceptually comparable approach). We characterize the regions by extracting topics using the Latent Dirichlet allocation (LDA) algorithm [3]. Then the topic distances are computed between geographical entities based on the regional probability distribution over the topics.

This contribution proposes an unified and presumably original formalism aimed at analysing spatial configurations endowed with textual information.

We first present our methodology in section 2 by introducing the basic ingredients permitting to extract textual and spatial information, defining in turn topic dissimilarities between regions as well as a weighted spatial network of regions. Then, those two aspects are combined together into a soft clustering problem. Section 3 presents a case study of the application of the method on Wikipedia entries of the municipalities of the canton of Vaud, Switzerland. Section 4 discusses the method and suggests some future developments of interest.

2 Methodology

Our methodology is based upon a combination of topic modeling, topic distance extraction and spatial clustering. The latter has been developed as an unsupervised classification algorithm for marked networks encountered in Geography and in Spatial Econometrics. This section first separately address each of these components, before formalizing their combined use.

The method requires a minimal amount of elements, namely a dataset of n regions with relative weights $f_i > 0$, reflecting their surface, population, or description size (section 3.1). Each region is associated with a text, such as a descriptive document involving a total variety of N words. Finally, the spatial configuration is defined by the adjacency matrix $A = (a_{ij})$ with values 1 iff i and j are neighbors, and else 0.

2.1 Topic modeling

To extract the topics from the texts associated to the regions we use LDA with Gibbs sampling as implemented in the `R` package `topic models` [11]. The main idea behind LDA is that a document is conceived as a random mixture over q latent topics, in other words "each document in a corpus exhibits multiple topics to a different degree" [3]. In our formalism, to express the results of LDA, we define the probability distributions of the regions over the topics as the row-normalized

$(n \times q)$ matrix $R = (r_{ik})$ and of the probability distributions of the term over the topics defined as the row-normalized $(N \times q)$ matrix $C = (c_{lk})$. The latter permits an interpretation of the topics, whereas the R matrix is used to extract topic distances between the regions.

2.2 Topic distances

To extract the $(n \times n)$ topic distances $D = (d_{ij})$ from the previously defined region-topic matrix R there are several possibilities. The most evident is to compute a distance between every pair of regions by computing the χ^2 or cosine distance between their term distributions, i.e. their rows of the R matrix. Another way to extract distances is to focus on a subset of topics judged particularly relevant for the purposes of the research objective, thus justifying the subsequent loss of information.

Standardizing the procedure (and the calibration of the parameter β below) may advocate the use of a rescaled distance $\widehat{D}$. The choice $\max_{ij} \hat{d}_{ij} = 1$ seems too dependent on the value of n. Instead, we will use the alternative

$$\hat{d}_{ij} = \frac{d_{ij}}{\Delta} \qquad \text{where} \qquad \Delta = \frac{1}{2} \sum_{i,j=1}^{n} f_i f_j D_{ij} \ . \tag{1}$$

2.3 Weighted spatial network

The spatial interaction between the n regions can be represented by a $(n \times n)$ symmetric non-negative *exchange matrix* $E(A, f, t) = (e_{ij})$, specifying the joint probability to select the unoriented edge ij as prescribed from the time-continuous Markov diffusive process with jump generator A at time t, with reversible transition matrix $w_{ij}(t) = e_{ij}(t)/f_i$ and stationary distribution f. The diffusive exchange matrix is *weight-compatible* in the sense $e_{i\bullet} = \sum_{j=1}^{n} e_{ij} = f_i$ [2], and constitutes a weighted generalization of the Laplacian diffusion kernel of machine learning [15][10]. Its limit $\lim_{t \to 0} e_{ij}(t) = f_i \delta_{ij}$ depicts a network made of disconnected nodes, while $\lim_{t \to \infty} e_{ij}(t) = f_i f_j$ represents a complete weighted network.

2.4 Soft clustering

We use the soft regional clustering for communities detection proposed by Ceré and Bavaud [4][5]. Soft partitions of n objects into m groups are represented by the non-negative, row-normalized $(n \times m)$ membership matrix $Z = (z_{ig})$, denoting the probability $p(g|i)$ that region i belongs to group g. Good partitions are defined as

local minima of the *generalized discontinuity free energy functional* $\mathscr{F}[Z]$

$$\mathscr{F}[Z] = \mathscr{K}[Z] + \beta \Delta_W[Z] + \frac{\alpha}{2}\mathscr{G}^{\kappa}[Z] \tag{2}$$

where the regularizing entropy term $\mathscr{K}[Z]$, favoring the advent of soft clustering, is the *mutual information* between the n regions and the m groups.

The second term $\Delta_W[Z] = \sum_{g=1}^{m} \rho_g \Delta_g$ is the *within-group inertia* relatively to the topic distances, whose presence supports the constitution of group of regions homogeneous enough relatively to the topic distributions, where [1]

$$\Delta_g = \frac{1}{2}\sum_{ij} f_i^g f_j^g \widehat{D}_{ij} \qquad f_i^g = p(g|i) = \frac{f_i z_{ig}}{\rho_g} \qquad \rho_g = \sum_{i=1}^{n} f_i z_{ig} \tag{3}$$

The third *discontinuity* term $\mathscr{G}^{\kappa}[Z] = \sum_{g=1}^{m} \rho_g^{-\kappa} \varepsilon[z^g]$, where ρ_g is the group weight (3) and $\varepsilon[z^g] = \frac{1}{2}\sum_{ij} e_{ij}(z_{ig} - z_{jg})^2$, insures the spatial continuity of the group memberships. As for $\mathscr{K}[Z]$, the "spatial energy" $\mathscr{G}^{\kappa}[Z]$ favors the constitution of soft clusters, in contrast to the "feature energy" $\Delta_W[Z]$ which favors *hard* partitions obeying $z_{ig} = 0$ or $z_{ig} = 1$ [1].

The parameter $\kappa \in [0,1]$ interpolates between weighted modularity maximization (for $\kappa = 0$) and Ncut (for $\kappa = 1$). The parameter $\beta > 0$ controls the influence of topic distances, while $\alpha = 0$ coincides with the soft K-means algorithm based on spherical Gaussian mixtures.

Minimizing the free energy functional (2) is performed by cancelling the first-order derivative under the conditions $z_{i\bullet} = 1$ and yields

$$z_{ig} = \frac{\rho_g \exp(-\beta \widehat{D}_i^g + \alpha \rho_g^{-\kappa}(\mathscr{L}z^g)_i - \frac{\alpha\kappa}{2}\rho_g^{-\kappa-1}\varepsilon[z^g]\ \ )}{\sum_h \rho_h \exp(-\beta \widehat{D}_i^h + \alpha \rho_h^{-\kappa}(\mathscr{L}z^h)_i - \frac{\alpha\kappa}{2}\rho_h^{-\kappa-1}\varepsilon[z^h])} \tag{4}$$

where D_i^g the squared Euclidean dissimilarity from i to the centroid of group g and $(\mathscr{L}z^g)_i$ is the *Laplacian* of membership z^g at region i, comparing its value to the average value of its neighbors by the matrix W. Values $\kappa > 0$ downscale the latter mechanism for large groups.

Equation (4) is solved iteratively until convergence. The choice of the initial partition Z^0 is made as follows: each region belonging to a set T of m *seeds* is attributed a distinct group $g = 1, \ldots, m$, the remaining *free regions* of the complementary set F being attributed to the background group $g = 0$ (see the examples of section 3)

The softness of the final partition Z^{∞} can possibly be measured by the value of the mutual information $\mathscr{K}[Z^{\infty}]$. More precisely, the pointwise conditional entropy $H(G|i) = -\sum_g z_{ig}^{\infty} \ln z_{ig}^{\infty}$ (where G denotes the variable "group") measures the membership uncertainty of region i, and takes on large values for regions located at the group frontiers. Alternatively, the final partition can be be further hardened by assigning each region i to group $g = \arg\max_h z_{ih}^{\infty}$.

2.5 Spatial autocorrelation: Moran's I

To evaluate the compatibility between E and the uni- or multivariate feature distances D, we use the *spatial autocorrelation* measure which constitutes a weighted, multivariate generalization of Moran's I [2]

$$I \equiv I(E,D) = \frac{\Delta - \Delta_{\text{loc}}}{\Delta} \tag{5}$$

$$\text{where} \qquad \Delta = \frac{1}{2}\sum_{i,j=1}^{n} f_i f_j D_{ij} \qquad \text{and} \qquad \Delta_{\text{loc}} = \frac{1}{2}\sum_{i,j=1}^{n} e_{ij} D_{ij} \tag{6}$$

respectively define the total inertia between all regions and the local inertia between connected regions. I ranges in $[-1,1]$, where a large positive value is expected when the topic distributions between neighbors are close. The standardized test value z serves at testing the null hypothesis H_0 of absence of spatial autocorrelation in the normal approximation (e.g. [2][4]).

3 Textual cartography of municipalities

To illustrate the method, we considered the English Wikipedia pages of the $n = 309$ municipalities of the canton of Vaud, Switzerland. The size and content of the pages are widely varying, yet the proposed methodology can cope with this heterogeneous and irregular data.

To obtain the data we developed a script that finds the municipalities on Dbpedia[7], from there extracts the administrative number fo the municipality and the link to its Wikipedia and downloads its textual content (without HTML tags). Three municipalities were not detected automatically and their links where added by hand. Also, we excluded the old municipalities which have been merged together but still have a Wikipedia[18] page by using the most recent data[17]. From this second source we got the spacial data, namely the `shapefile` of the canton of Vaud and the adjacency matrix A of the municipalities.

To construct the $(N \times n)$ term-document, or term-region matrix on which to perform LDA, we performed the following operations on the documents associated with the regions using the `tm` package in `R` ([8]: we removed punctuation, and put all the words in lower-case. Then we removed the classical English stop-words, further stop-words and topic general words [19] of our choice (e.g. "new", "most", "residents", "completed", "meter", etc.), and the names of the municipalities themselves. We also removed sparse terms of relative frequency less than 0.01, those terms are too specific and thus have little or none influence on topic modeling [13]. After this treatment $N = 983$ words (types) for a total of $46'022$ occurrences where retained. Finally, we extracted the weights f of the regions (figure 1) being the normalized sum of the columns of the term-region matrix.

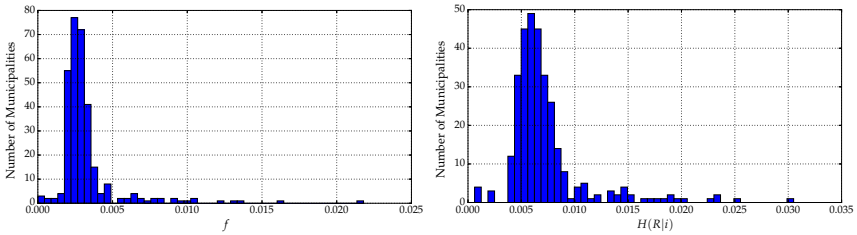

Fig. 1 Number of municipalities in function of *left*: f, *right*: the condition entropy of topics $H(R|i)$ showing municipality-topic probability distribution diversity.

3.1 Topic modeling

To extract the topics we used the `LDA` function of `R` package `topicmodels` [11], using the Gibbs sampling method with $q = 12$ topics[1]. The results are illustrated in figure 2. Interpreting all the topics exceeds the scope of this contribution.

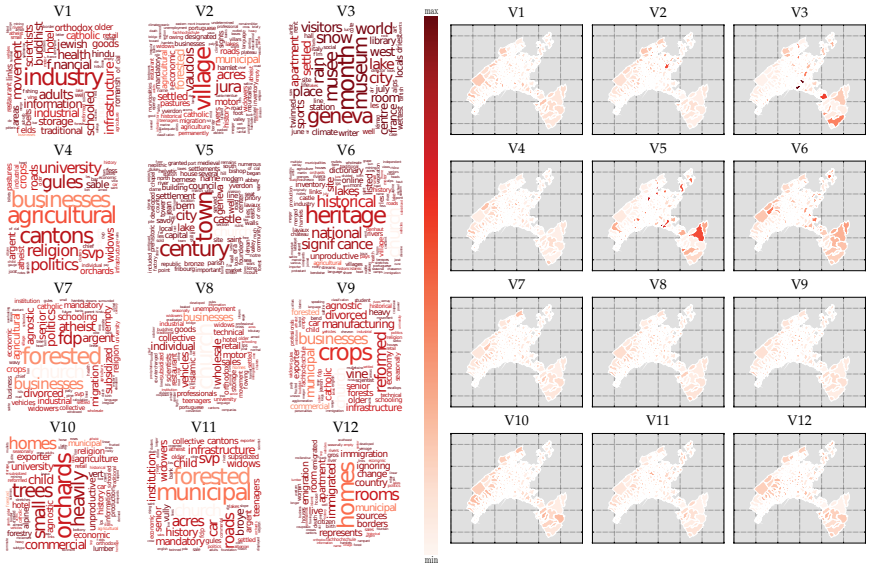

Fig. 2 *Left* : 12 topic wordclouds [9] obtained from the C matrix for topic model with parameters: $k = 12$, *burnin* $= 4'000$, $2'000$ iteration, *thin* $= 500$, *seeds* $= \{2'003, 5, 63, 100'001, 756\}$, *nstart* $= 5$ and *best* $=$ true. The colour scale form is the inverse of the frequency of the word in the whole dataset (e.g. more specific words are darker than common words) whereas their size represents their importance in the topic. *Right* : related maps with the probabilities of the municipalities $\in [0.006, 0.628]$ over each topic (Map base source: FSO).

[1] The value of q has been selected after computing several topic models with q varying from 9 to 36, With $q = 12$ we had a relatively small number of interesting topics apt to illustrate our method.

On one hand, one observe in figure 2 that the topics seem to capture different semantic traits of the regions. On the other hand, the spatial maps underline the spatial autocorrelation of some of the topics. Table 1 gives the values of the corresponding Moran's I for each topic, as well as their standardized test value z.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
I	0.57	0.65	0.64	0.55	0.66	0.63	0.51	0.53	0.50	0.56	0.61	0.75
z	8.64	12.49	12.00	7.64	12.93	11.38	5.58	6.65	5.11	7.74	10.15	16.94

Table 1 Moran's I for the 12 topics and its test values z. A large positive I indicates large communities with similar features, whereas a large negative I indicates regional contrast.

3.2 Topic distances

We used the chi-square distance to compute the topic distance between regions:

$$d^{\chi}_{ij} = \sum_{k=1}^{q} \frac{(r_{ik} - r_{jk})^2}{R_k} \qquad \text{where} \quad R_k = \sum_{i=1}^{n} f_i r_{ik} \quad \text{is the topic weight.} \tag{7}$$

As noted in the literature, weighted multidimensional scaling on d^{χ} amounts to the correspondence analysis (CA) of the associated contingency table (figure 3).

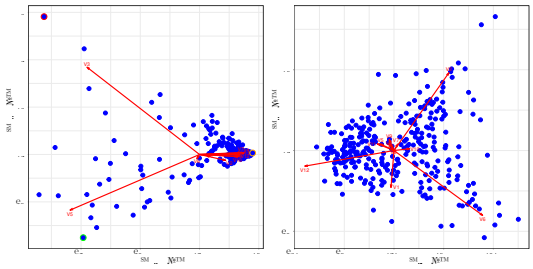

Fig. 3 CA of region-topic matrix (Dimensions 1 and 2 on the left, 3 and 4 on the right) obtained with the `R` package `FactoMineR` [12].

	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
V1	0.25	0.03	0.05	12.58	36.25
V2	1.29	0.25	14.12	48.36	3.80
V3	26.96	63.72	0.13	0.44	0.01
V4	3.88	0.02	0.05	0.19	4.44
V5	49.32	35.76	1.34	0.54	0.05
V6	0.01	0.12	36.87	33.75	19.23
V7	3.18	0.03	0.07	0.34	0.12
V8	2.90	0.00	0.00	0.02	5.94
V9	3.18	0.03	0.11	1.06	1.70
V10	2.75	0.03	1.17	0.03	0.10
V11	2.62	0.00	1.32	0.36	1.61
V12	3.66	0.00	44.76	2.35	26.78

Table 2 CA of region-topic matrix: contributions of the topics to the dimensions.

As made apparent in figure 3 and table 3.2, the different topics have varying contributions to the dimensions of the CA. It is interesting to notice that the first dimension discriminates between city and rural municipalities. Looking more precisely at topics 3 and 5 (that contribute the most to the first dimension) in figure 2, this interpretation seems to be confirmed by the terms describing them. Hence, a CA on the region-topic distance provides some help for interpreting the topics themselves.

3.3 Community detection

To illustrate community detection on the complex weighted network, we first restrict ourselves to the univariate feature consisting of the topic V5 only. The CA results in the table 3.2 indicates that this topic is explaining largely dimension 1 (49.32%). The region initial seeds (for $m = 3$ groups with a background group $g = 0$) are consequently determined as corresponding to the maximum, the minimum and the median values on this dimension. The continuity of this feature within the spatial weighted network is demonstrated by the highly significant Moran's $I = 0.66$, yielding a standardized normal test value $z = 12.93$ (figure 6). Figure 4 depicts the hardened partition Z^{∞} after convergence of the iterative procedure (figure 5). The entropies for each region are show in the figure 7.

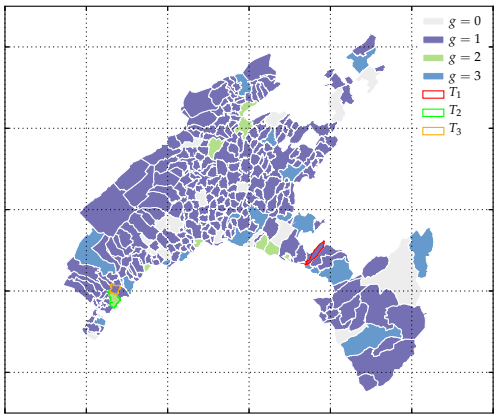

Fig. 4 *Municipalities soft clustering on the topic V5* depicts the unsupervised hard assignment obtained from the initial strokes $T_1 = \{5884\}$ border colored in red (Corsier-sur-Vevey; group 1 colored in purple), $T_2 = \{5724\}$ border colored in neon green (Nyon; group 2 colored in green) $T_3 = \{5487\}$ border colored in orange (Duillier; group 3 colored in blue) and group 0 as the background group colored in grey after 100 iterations.

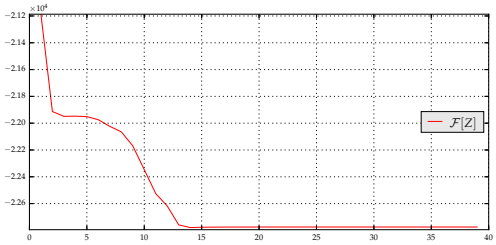

Fig. 5 *One topic* : decrease of the free energy $\mathscr{F}[Z]$ during the iteration with $\kappa = 0.0$, $\beta = 50.0$ and $\alpha = 1.0$.

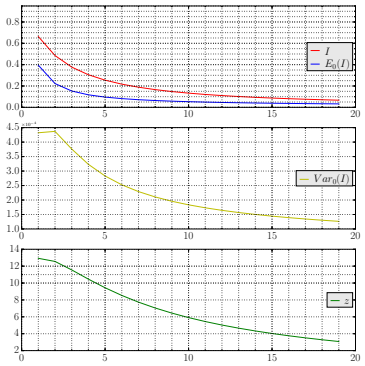

Fig. 6 *Municipalities spatial autocorrelation of the topic V5*: Moran's I and standardized z normal test value, as a function of the free parameter $t \in [1,20]$ of the diffusive exchange matrix specification $E(A, f, t)$.

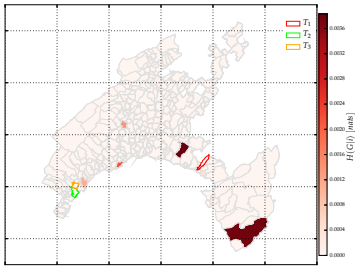

Fig. 7 Conditional pointwise entropy $H(G|i)$ for the conditions soft clustering on the topic V5. A large value of $H(G|i)$ denotes a large uncertainty in the membership of region i.

As a second, multivariate illustration, taking into account all the $q = 12$ topics also demonstrates the coherence between the network and the textual features (Moran's $I = 0.64$ with standardized normal test value $z = 11.86$).

On the CA biplot, three main zones are observed (see figure 3, which also indicates the seeds of the partition Z^0). Figure 8 illustrates the hardened partition Z^∞ after convergence (see figure 9). The entropies for each region are shown in the figure 11.

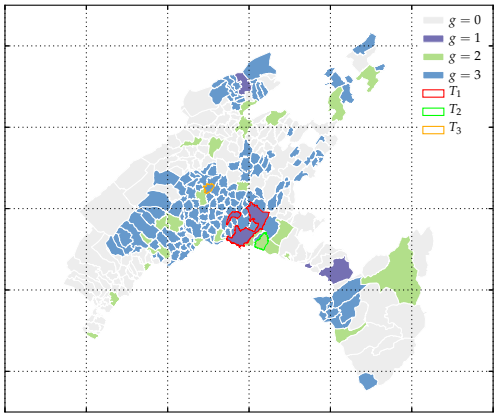

Fig. 8 *Municipalities soft clustering on all the topics* depicts the unsupervised hard assignment obtained from the initial strokes $T_1 = \{5586\}$ border colored in red (Lausanne; group 1 colored in purple), $T_2 = \{5606\}$ border colored in neon green (Lutry; group 2 colored in green) $T_3 = \{5487\}$ border colored in orange (Lussery-Villars; group 3 colored in blue) and group 0 as the background group colored in grey after 100 iterations.

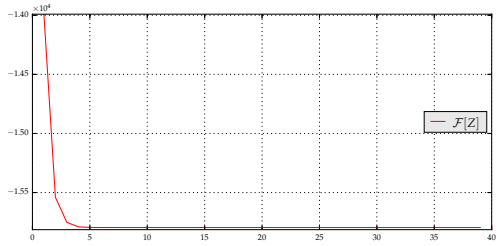

Fig. 9 *All topics* : decrease of the free energy $\mathscr{F}[Z]$ during the iteration with $\kappa = 0.0$, $\beta = 50.0$ and $\alpha = 1.0$.

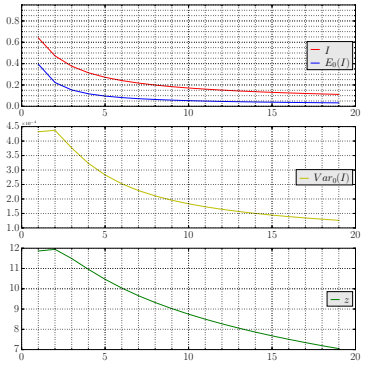

Fig. 10 *Municipalities spatial autocorrelation all the topics*: Moran's I and standardized z normal test value, as a function of the free parameter $t \in [1,20]$ of the diffusive exchange matrix specification $E(A, f, t)$.

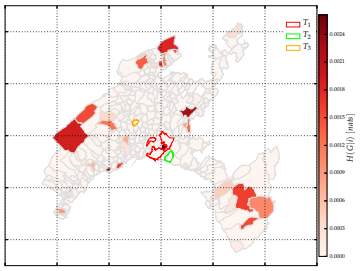

Fig. 11 Conditional pointwise entropy $H(G|i)$ for the conditions soft clustering on all topics. A large value of $H(G|i)$ denotes a large uncertainty in the membership of region i.

In figure 4 we can detect a rural ($g = 1$) versus urban cluster ($g = 2$ and $g = 3$), the latter distinguishing two urban subcategories. Figure 8 reveals a centrality effect (generally urban versus suburban). Furthermore, the conditional pointwise entropy (figures 7 and 11) helps detecting participation to many clusters. Also, the spatial autocorrelation of topics can be analysed at various geographical scales by varying the freely adjustable parameter t, as illustrated by figures 6 and 10.

4 Discussion

We have presented a general method for community detection based on a complex weighted network, made of a spatial network enriched by textual information. The proposed formalism can tackle and integrate both aspects, as illustrated on the real dataset consisting of the text descriptions of the municipalities of the canton of Vaud.

Although limited by the quality and size of the textual dataset, and by the question of how many topics should be selected, our case study seems to open promising perspectives regarding the possibility to interpret the topics (a still debated issue [6]) through CA and community detection. Those aspects could presumably be improved significantly by possessing a larger dataset, of better textual quality - a point to be addressed in future work.

Naturally, it is possible to use the algorithm on a finer spatial regular grid (to overcome the irregularities imposed by arbitrary administrative boundaries or creating textually contiguous regions [16]) and to incorporate numerical features (such as socio-economical data on hectometric census blocks). In the opposite direction, the question of how to attribute a topic distribution for even more irregular datasets, including for instance regions with missing textual data, can already be handled by the soft clustering algorithm presented here.

Also, CA on the region-topic matrix can bring some guidance to determination of the number of communities - an ever-lasting issue in unsupervised clustering in general.

Finally, the exchange matrix can also be used to describe a social network, rather than a geographical network. In this perspective, and based on proper textual data, our method can be used as a collaborative filtering [14], for instance apt to recommend items similar to a given item of interest.

References

[1] Bavaud, F.: Aggregation invariance in general clustering approaches. Advances in Data Analysis and Classification **3**(3), 205–225 (2009)
[2] Bavaud, F.: Testing spatial autocorrelation in weighted networks: the modes permutation test. Journal of Geographical Systems **3**(15), 233–247 (2013)
[3] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003). URL http://dl.acm.org/citation.cfm?id=944919.944937
[4] Ceré, R., Bavaud, F.: Multi-labelled image segmentation in irregular, weighted networks: A spatial autocorrelation approach. In: GISTAM 2017 - Proceedings of the 3rd International Conference on Geographical Informa-

tion Systems Theory, Applications and Management, Porto, Portugal, 27-28 April, 2017., pp. 62–69 (2017). DOI 10.5220/0006322800620069. URL https://doi.org/10.5220/0006322800620069

[5] Ceré, R., Bavaud, F.: Soft image segmentation: on the clustering of irregular, weighted, multivariate marked networks (2017). Accepted for Springer Book of GISTAM 2017: Communications in Computer and Information Science CCIS series.

[6] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Advances in neural information processing systems, pp. 288–296 (2009)

[7] DBpedia: DBpedia (2017). URL https://dbpedia.org/. http://dbpedia.org [Online; accessed 01-September-2017]

[8] Feinerer, I., Hornik, K., Meyer, D.: Text mining infrastructure in r. Journal of Statistical Software **25**(5), 1–54 (2008). URL http://www.jstatsoft.org/v25/i05/

[9] Fellows, I.: wordcloud: Word Clouds (2014). URL https://CRAN.R-project.org/package=wordcloud. R package version 2.5

[10] Fouss, F., Saerens, M., Shimbo, M.: Algorithms and models for network data and link analysis. Cambridge University Press (2016)

[11] Grün, B., Hornik, K.: topicmodels: An R package for fitting topic models. Journal of Statistical Software **40**(13), 1–30 (2011). DOI 10.18637/jss.v040.i13

[12] Lê, S., Josse, J., Husson, F.: FactoMineR: A package for multivariate analysis. Journal of Statistical Software **25**(1), 1–18 (2008). DOI 10.18637/jss.v025.i01

[13] Lu, K., Cai, X., Ajiferuke, I., Wolfram, D.: Vocabulary size and its effect on topic representation. Information Processing & Management **53**(3), 653–665 (2017)

[14] Salah, A., Nadif, M.: Social regularized von mises–fisher mixture model for item recommendation. Data Mining and Knowledge Discovery **31**(5), 1218–1241 (2017). DOI 10.1007/s10618-017-0499-9

[15] Smola, A.J., Kondor, R.: Kernels and regularization on graphs. In: COLT, vol. 2777, pp. 144–158. Springer (2003)

[16] Sui, D.Z., Elwood, S., Goodchild, M.F. (eds.): Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice. Springer, Dordrecht ; New York (2013). OCLC: ocn810987841

[17] Swiss Federal Statistical Office (FSO): STAT-TAB - Interactive tables (2017). URL http://www.bfs.admin.ch. https://www.pxweb.bfs.admin.ch [Online;; accessed 01-September-2017]

[18] Wikipedia: Wikipedia, The Free Encyclopedia (2017). URL https://en.wikipedia.org/. http://en.wikipedia.org [Online; accessed 01-September-2017]

[19] Xu, Y., Yin, Y., Yin, J.: Tackling topic general words in topic modeling. Engineering Applications of Artificial Intelligence **62**, 124 – 133 (2017). DOI 10.1016/j.engappai.2017.04.009. URL http://www.sciencedirect.com/science/article/pii/S0952197617300738

[20] Youssef Mourchid, M.E.H., Cherifi, H.: An image segmentation algorithm based on community detection. In: Complex Networks & Their Applications V Proceedings of the 5th International Workshop on Complex Networks and their Applications (COMPLEX NETWORKS 2016), pp. 821–830. Springer (2017). DOI 10.1007/978-3-319-50901-3_65