An Illustrated Approach to Soft Textual Cartography

Raphaël Ceré and Mattia Egloff

Abstract Soft textual cartography is an approach aimed at extracting clusters of regions taking into account both their spatial relationships and a their textual description within a corpus. The strategy consists in constructing a complex weighted network, reflecting the geographical layout, and whose nodes are further characterized by their thematic dissimilarity, extracted form topic modelling. A soft k-means procedure, taking into account both aspects through expectation maximisation on Gaussian mixture models and label propagation, converges towards a soft membership, to be further compared with expert knowledge on regions. Application on the Wikipedia pages of Swiss municipalities demonstrate the potential of the approach, revealing textual autocorrelation and associations with official classifications. The synergy of the spatial and textual aspects appears promising in topic interpretation and geographical information retrieval, and able to incorporate expert knowledge through the choice of the initial membership.

Key words: Textual Cartography, Complex Network, Topic modelling, Thematic Exploration, Soft Clustering, Text Mining, GIS, Membership Association, Wikipedia.

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

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

1 Introduction

Regional data analysis generally involves numerical or categorical information attached to the regions, such as level intensities or densities provided from census data (e.g. population, socio-economical properties). Another rich information source that should be considered in regional data analysis is "common textual knowledge". Yet, the question of how to exploit this type of data in quantitative methods is generally not trivial. On one hand, textual data may require human interpretation to be used meaningfully and its use in quantitative methods is not straightforward. On the other hand, when evaluating an algorithm, textual data can be useful to provide insight in the results.

In this paper, we first show how it is possible to use textual data in regional geography, and more precisely how to extract textual distances and use them in an adapted clustering algorithm. Secondly, we address the question: how to interpret the clusters obtained from the algorithm in view of, textual and regional characteristics, and using expert knowledge? From a geographical perspective, this second idea follows [14], which argues that fully automated spatial data analysis does not exploit the advantage of the practitioner's input performing the classification. Indeed, a person has to evaluate the results of any automated procedure without knowing exactly how the latter was really performed. Even more, the similarity between administrative entities depends on the points of view. For example a territorial network admits several "valid" classifications corresponding on the nature of the analysis, interest or study objectives. Thus, the knowledge provided from the practitioner can be included by specifying in a clustering task, initial memberships to infer the segmentation in a certain aim with keeping the advantage of automated approach.

Methodologically this paper uses "soft textual cartography", as previously developed in [11]. Textual information is used with the method of regional semi-automated soft clustering proposed by Ceré and Bavaud [6, 7]. That implements the combination of spatial configuration and features distances in an image segmentation framework (see [25] for a conceptually comparable approach) to perform semi-automated regional segmentation.

We improve the results as presented in Egloff and Ceré [11]. Applying the method on a larger dataset composed by all the municipalities of Switzerland. It furthermore, emphasises the role of the initial memberships in the iterative procedure. Also, the analysis of the results is clarified by means of correspondence analysis (CA) between different memberships. For the validation of the obtained memberships we use an official classification provided by the Swiss Federal Statistical Office (FSO).

The paper is structured as follows: Section 2 introduces the basic ingredients necessary for the "soft textual cartography" and the data used for the illustration of the method. Then, in section 3 we introduce the heart of the method explaining:

the extraction of the weighted spatial network, the textual distance obtained from topic modelling on the corpus, the spatial autocorrelation and finally, the clustering algorithm itself. In section 4, different initial memberships used to test the model are described, among which the "gold standard" the official classification. Section 5 presents a method to evaluate membership association and analyses some results obtained by the algorithm and compares it to a classical approach. Finally, section 6 draws some conclusions about the usage of the algorithm.

2 Data

Soft textual cartography requires a minimal amount of elements [11], namely a dataset of n regions with relative weights $f_i > 0$, $\sum_i^n f_i = 1$, reflecting their surface, population, or description size. Also each region has to be associated with a text, such as a descriptive document, involving a total variety of N words. The final element consists in the spatial configuration, which is defined by the binary adjacency matrix $A = (a_{ij})$ of size $n \times n$ with values 1 if i and j are neighbours, and 0 otherwise.

Textual data consists of the Wikipedia pages [24, 9] of the $n = 2068$ municipalities of Switzerland. To keep a spatial continuum, municipalities of Liechtenstein, as well as enclaves (Campione d'Italia and Büsingen am Hochrein) present in the Swiss territory have been included.

Textual sections about important regional personalities as well as external links have been removed. Also, all references to cantons and municipality names have been withdrawn along with the usual stop-words. Finally, low- and high-frequency terms (respectively less than 20 and more than 9000 occurrences) have been also removed. Figure 1 shows the resulting weight-frequency f. This f is reflects the textual volume of information of the Wikipedia pages and defines the relative weight of the municipalities as used in the algorithm.

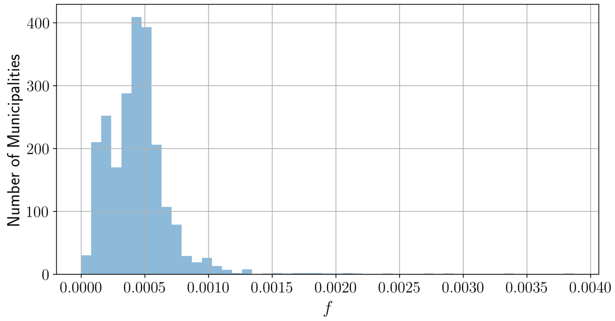

Fig. 1 Number of municipalities in function of *left*: f.

3 Soft Textual Cartography

This section introduces the ingredients involved in the algorithm, in particular the neighbourhood network of the municipalities and the textual distance. Then, with the help of Moran's I, we measure the textual autocorrelation relative to the spatial configuration. Finally, we introduce a particular version of the algorithm leveraging on our previous work [11].

3.1 Weighted spatial network

The spatial connectivity between the n regions is expressed by a $(n \times n)$ symmetric non-negative *exchange matrix* $E(A, f, t) = (e_{ij}^{(t)})$. The latter specifies the joint probability to select the unoriented edge ij as prescribed from the time-continuous Markov diffusive process with jump generator A at time $t > 0$; the so called Laplacian diffusion kernel of machine learning [22, 13] constituted an unoriented unweighed network. Note that the transition matrix $w_{ij}(t) = e_{ij}^{(t)}/f_i$ is reversible and has a stationary distribution f. The weight-compatible $e_{i\bullet} = \sum_{j=1}^{n} e_{ij} = f_i$ [4] diffusive exchange matrix constitutes a weighted generalization of the unweighed approach using diffusive kernel. Its limit $\lim_{t\to 0} e_{ij}^{(t)} = f_i \delta_{ij}$ depict a network made of disconnected nodes, while $\lim_{t\to\infty} e_{ij}^{(t)} = f_i f_j$ represents a complete weighted network.

3.2 Textual distance

There are several possible ways to extract distances between the municipalities from textual data. For the approach illustrated a topic distance is defined as follows. First, we define the $N \times n$ term-municipality matrix as the matrix associating each term with its frequency in the document corresponding to each municipality. In a second step we use the Latent Dirichlet Allocation (LDA) [5] algorithm to extract the latent k topics from the texts, from which the χ^2 distances are finally extracted (see below).

The main idea behind LDA is that a document is conceived as a random mixture over k latent topics and each topic a random mixture over the terms or words. The topics obtained form LDA generally are able to regroup words used in similar contexts (semantically correlated or synonyms) into the same topic (for example: see "city" and "town" in topics V4 in figure 3.2). Consequently, a word possessing more than one sense can belong with a high probability to more than one topic (for example: see "businesses" in topics V2 and V5 in figure 3.2). In this paper we use the Gibbs sampling method to approximate the solution of the LDA to as implemented in the `R` package `topicmodels` [15]).

As the municipalities are in a one to one correspondence with the documents: the probability distributions of the municipalities over the topics is defined as the row-normalized $(n \times k)$ document-topic matrix $R = (r_{iq})$, and the probability distributions of the terms over the topics is defined as the row-normalized $(N \times k)$ term-topic matrix $C = (c_{lq})$. The latter permits an interpretation of the topics, whereas the R matrix is used to extract topic distances between the regions.

To extract the $(n \times n)$ topic-distance $D = (d_{ij})$ from the previously defined municipality-topic matrix R the χ^2 distance $d_{ij}^{\chi} = \sum_{q=1}^{k}(r_{iq} - r_{jq})^2/R_k$ (where $R_q = \sum_{i=1}^{n} f_i r_{iq}$ being the topic weight) is computed between the topic distributions of the municipalities, i.e. the rows of the R matrix. Figures 3.2 and 3.2 depict the topic probabilities of the Swiss municipalities; noticeably the topics extracted seem to be spatially autocorrelated.

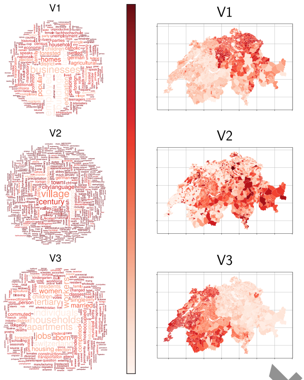

Fig. 2 For $k = 3$ topics, *Left*: topic wordclouds [12] obtained from the C matrix for topic model with parameters: *burning* $= 4'000$, $2'000$ iteration, *thin* $= 500$, *seeds* $= \{2'003, 5, 63, 100'001, 756\}$, *nstart* $= 5$ and *best* $=$ true. The colour scale form is the inverse of the frequency 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 over each topic (Map base source: FSO).

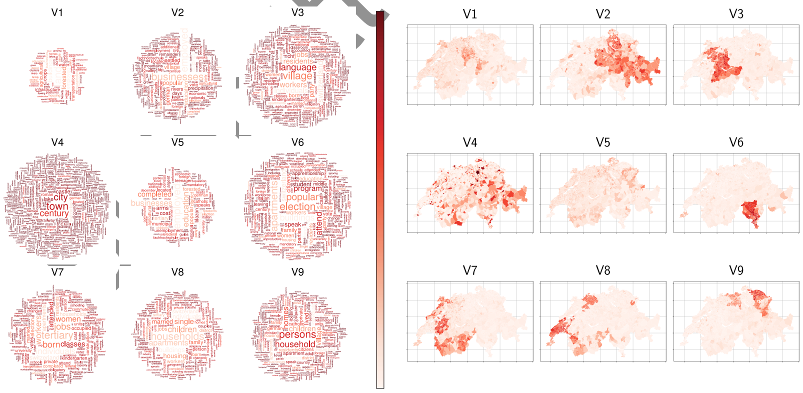

Fig. 3 For $k = 9$ topics, *Left*: topic wordclouds [12] obtained from the C matrix for topic model with parameters: *burning* $= 4'000$, $2'000$ iteration, *thin* $= 500$, *seeds* $= \{2'003, 5, 63, 100'001, 756\}$, *nstart* $= 5$ and *best* $=$ true. The colour scale form is the inverse of the frequency 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 over each topic (Map base source: FSO).

3.3 Spatial autocorrelation

Obviously, the basic spatial statistical analysis or classification of an spatial data set makes sense only if there is a spatial autocorrelation to begin with. The Moran'I provides an index of spatial autocorrelation [1] measuring to which extent the topic-distance D is smaller between spatially close municipalities, as defined by the spatial configuration E. We use here the weighted, multivariate generalization of Moran's I where the spatial autocorrelation significance is evaluated with the standardized test value z (e.g. [4, 6])

$$I \equiv I(E,D) = \frac{\Delta \; - \; \Delta_{\text{loc}}}{\Delta} \qquad \text{with} \qquad z = \frac{|I - E_0(I)|}{\sqrt{\text{Var}_0(I)}} \tag{1}$$

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

respectively define the total inertia between all regions and the local inertia between connected regions. The figure 4 shows the measured I, ranges in $[-1,1]$, where a large positive value is expected when the topic distributions between neighbours are close.

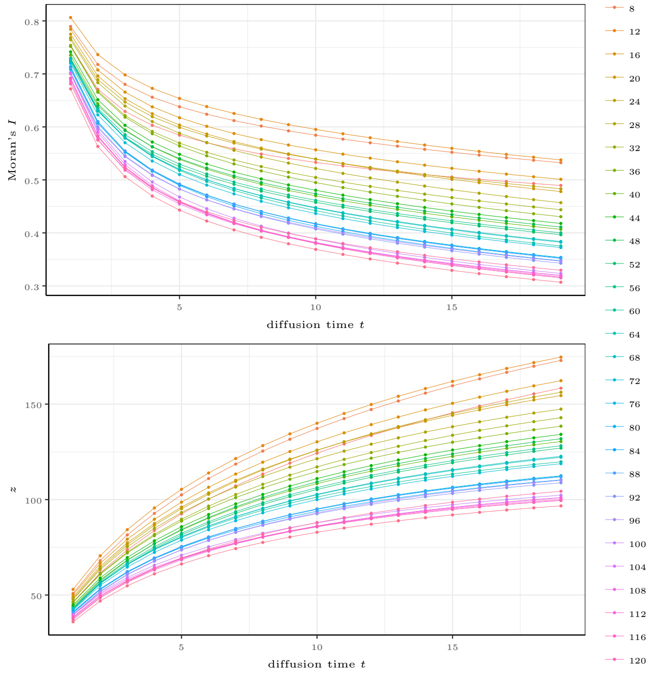

Fig. 4 The figures represent *Moran's I* (*upper*) and z (*bottom*) trough the exchange matrix diffusion process at time $t = 1,\ldots,20$ using various distances extracted from topic models having $k = 8, 12, \ldots, 120$ topics.

3.4 The Algorithm

As a reminder, the soft clustering method already described in [11] is reproduced in this section, with minor adaptations. This approach combines textual information and spatial configuration independently. Notice that the initial membership or partition Z^0 can be used other information (e.g. expert knowledge).

The assignment of n objects to m groups is represented by the non-negative, row-normalized $(n \times m)$ membership matrix $Z = (z_{ig})$, where z_{ig} denotes the probability $p(g|i)$ that region i belongs to group g. In the general soft case, $z_{ig} \geq 0$ with $\sum_{g=1}^{m} z_{ig} = z_{i\bullet} = 1$, whereas $z_{ig} = 0$ or $z_{ig} = 1$ in the hard case.

The soft regional clustering for communities detection [6, 7] is initialised with initial membership $Z^0 = (z^0_{ig})$ and is using expectation maximisation to produce the final assignment. Explicitly, a good membership is defined as local minima of the *generalized discontinuity free energy functional* $\mathscr{F}[Z]$ from Z^0:

$$\mathscr{F}[Z] = \mathscr{K}[Z] + \beta \Delta_W[Z] + \frac{\alpha}{2}\mathscr{G}[Z] \tag{3}$$

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

$$\mathscr{K}[Z] = \sum_{ig} f_i z_{ig} \ln \frac{z_{ig}}{\rho_g} \qquad\qquad \rho_g = \sum_{i=1}^{n} f_i z_{ig} \tag{4}$$

where ρ_g is the group weight. 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 [3]

$$\Delta_g = \frac{1}{2}\sum_{ij} f_i^g f_j^g D_{ij} \qquad\qquad f_i^g = p(i|g) = \frac{f_i z_{ig}}{\rho_g} \tag{5}$$

The third *discontinuity* term $\mathscr{G}[Z] = \sum_{g=1}^{m} \rho_g^{-1} \varepsilon[z^g]$ 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" $\mathscr{G}[Z]$ favours the constitution of soft clusters, in contrast to the "feature energy" $\Delta_W[Z]$ which favours *hard* memberships obeying $z_{ig} = 0$ or $z_{ig} = 1$ [3].

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 (3) is performed by cancelling the first-order derivative under the conditions $z_{i\bullet} = 1$ and yields:

$$z_{ig} = \frac{\rho_g \exp(-\beta D_i^g + \alpha \rho_g^{-1} (\mathscr{L} z^g)_i - \frac{\alpha}{2} \rho_g^{-2} \varepsilon[z^g] \ \ )}{\sum_h \rho_h \exp(-\beta D_i^h + \alpha \rho_h^{-1} (\mathscr{L} z^h)_i - \frac{\alpha}{2} \rho_h^{-2} \varepsilon[z^h])} \tag{6}$$

where D_i^g the standardised[1] 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 neighbours as defined by the matrix W - an ingredient typical of *label propagation models*.

Equation (6) is solved iteratively until convergence. The choice of the initial membership matrix Z^0 is discussed in section 4. The hardness of the final membership matrix Z^∞ can possibly be measured by the value of the mutual information $\mathscr{K}[Z^\infty]$. Also, the point-wise 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 membership matrix can be further hardened by assigning each region i to group $G[i] = \arg\max_{g\in\{1,\ldots,m\}} z_{ig}^\infty$.

4 Parameter choice and initial conditions

To illustrate the algorithm and study the influence of the initial membership Z_0, the following parameter choices were made. First, parameter k (the number of topics) was chosen to be the same as the number of groups m, thus $k = m$. In turn, m was chosen to correspond to the numbers of groups presented in the three official municipality classifications issued by the FSO, namely $m = 3$, $m = 9$ and $m = 25$. The value for parameters β and α of the soft clustering algorithm have been tuned by numerical experimentation. The free parameter β, which can be interpreted as the inverse temperature in statistical mechanics, controls the hardness of the classification. The free parameter α controls the extent to which the spatial configuration is taken into account. Finally, the parameter t controls the age of diffusive process: a low t limits the interactions to the nearest neighbours.

To use the clustering algorithm proposed in section 3.4 an initial membership matrix Z^0 is required. To study the impact of the initial membership we went beyond the method proposed in [11], where pre-selected municipalities were used based on their atypicality in the correspondence analysis over the topics (their distance towards the mean profile). Hence, three different initial membership attributions:

- three official classifications, $m \in \{3, 9, 25\}$, from the FSO based on a urban-rural model, see figure 5,
- two random memberships (soft and hard) for each municipality i to the group g, $m = k$, where the number of the topics is $m \in \{3, 9, 25\}$, see figure 6,
- and three hard memberships, $m \in \{3, 9, 25\}$, obtained from the k-means algorithm on the generalised χ^2 distance see subsection 7 obtaining from the region-document matrix, represented in figure 7.

[1] D has been divided by $\Delta = \frac{1}{2}\sum_{i,j=1}^{n} f_i f_j D_{ij}$ which amounts to recalibrate the value of the free parameter β.

4.1 Official classifications

The official municipalities classifications, $m = 3, 9, 25$, of Switzerland [26] (version 2017) is based on the delimitation of the urban space in 2012 based upon morphological (density) and functional (commuting flows) conditions. The $m = 9$ categories include the size and the accessibility of the municipalities. The so called rural-urban typology $m = 3$ depicts the “Urban (1)”, “Intermediary (2)”, “Rural (3)” municipalities which is based on the classification $m = 9$. The $m = 25$ categories distinguishes by socio-economic conditions in municipalities. The details of how those typologies have been determined are not further investigated here; those typologies are used here as the “gold standard” to compare the results further obtained.

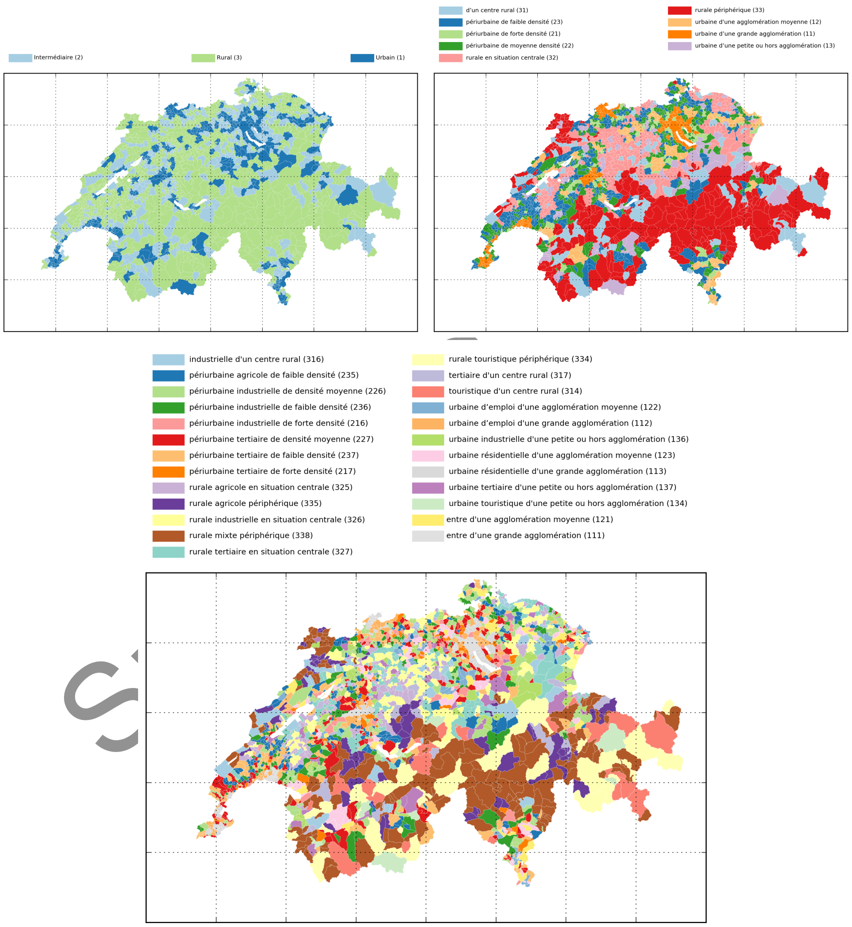
Intermédiaire (2)
Rural (3)
Urbain (1)
d’un centre rural (31)
périurbaine de faible densité (23)
périurbaine de forte densité (21)
périurbaine de moyenne densité (22)
rurale en situation centrale (32)
rurale périphérique (33)
urbaine d'une agglomération moyenne (12)
urbaine d’une grande agglomération (11)
urbaine d’une petite ou hors agglomération (13)
industrielle d'un centre rural (316)
périurbaine agricole de faible densité (235)
périurbaine industrielle de densité moyenne (226)
périurbaine industrielle de faible densité (236)
périurbaine industrielle de forte densité (216)
périurbaine tertiaire de densité moyenne (227)
périurbaine tertiaire de faible densité (237)
périurbaine tertiaire de forte densité (217)
rurale agricole en situation centrale (325)
rurale agricole périphérique (335)
rurale industrielle en situation centrale (326)
rurale mixte périphérique (338)
rurale tertiaire en situation centrale (327)
rurale touristique périphérique (334)
tertiaire d'un centre rural (317)
touristique d'un centre rural (314)
urbaine d'emploi d'une agglomération moyenne (122)
urbaine d'emploi d'une grande agglomération (112)
urbaine industrielle d'une petite ou hors agglomération (136)
urbaine résidentielle d'une agglomération moyenne (123)
urbaine résidentielle d'une grande agglomération (113)
urbaine tertiaire d'une petite ou hors agglomération (137)
urbaine touristique d'une petite ou hors agglomération (134)
entre d'une agglomération moyenne (121)
entre d'une grande agglomération (111)

Fig. 5 Illustration of the maps of the official classifications, form *left top* to *bottom* the parameters are: $m = 3$; $m = 9$; $m = 25$

4.2 Random memberships

For further testing, we first create random memberships where each region is uniformly assigned to groups $g = 1, \dots, m$. Three of them are illustrated in figure 6.

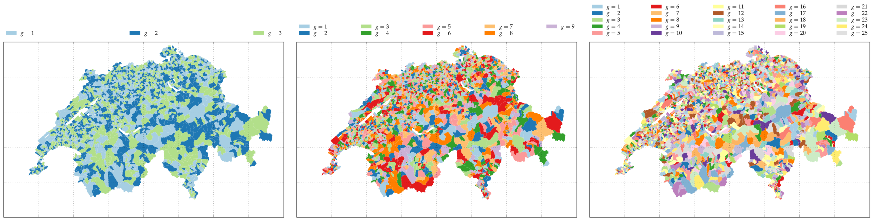

Fig. 6 Illustration of the maps of the groups obtained by uniform random attributions, form left to right the parameters are: $m = 3$; $m = 9$; $m = 25$.

4.3 Membership based on word-frequency

To test the algorithm further we compute another initial membership based on the term frequencies: we first define a distance based on the term-municipality matrix (defined in 3.2). To do this, we used the generalised χ^2 distance (see appendix 7) to compute the distance between the municipalities with respect to their word frequency profile. Figure 7 depicts three examples of groups obtained by submitting the distance obtained to an MDS to which we applied a hard k-means clustering [16] with the `R` package `stats` [20].

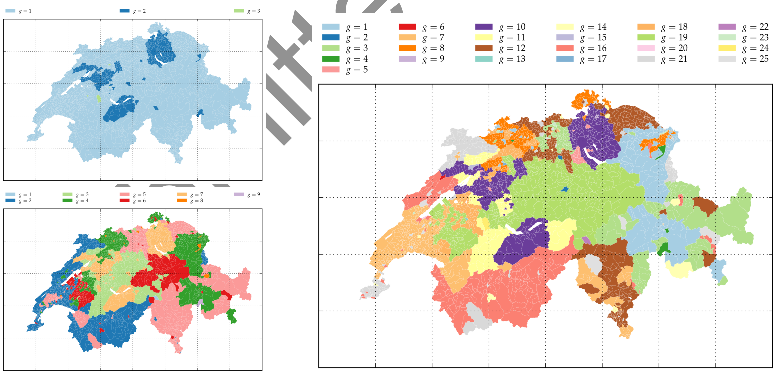

Fig. 7 The maps of the groups obtained by hard k-means clustering on the MDS over the generalised χ^2 distance between municipality profiles in the term-municipality matrix. Form *left top* to *right* the parameters are: $m = 3$, $\theta = 1.01$; $m = 9$, $\theta = 0.5$; $m = 25$, $\theta = 1.5$.

As shown in figure 7 this type of clustering has a tendency, depending on the value of θ, to create patches of municipalities that either have frequent or rare words in their Wikipedia page. It is not self evident that these patches should be spatially contiguous.

5 Results

In this section, we introduce membership association between two memberships, which is later used to compare the results of the algorithm with the official classifications. Then, for each initial membership discussed in section 4 we briefly analyse some results. Finally, we compare the present soft textual cartography approach to two classical approaches based on a network obtained from an affinity matrix.

5.1 Membership association

Starting with the initial membership Z^0, the iterative algorithm (6) converges towards a *local minimum* Z^∞ of the free energy. Z^∞ constitutes a soft membership, which can be further hardened for interpretation purposes, by entirely assigning each municipality i to group $G[i] = \arg\max_{g\in\{1,\ldots,m\}}(z^\infty_{ig})$. On one hand, the iterative process, depending only on the weighted geographical network as well as the the topic-induced distances, should erase in large part the initial attribution Z^0 of municipalities to groups. On the other hand, procedures such as the k-means, soft k-means and their variants are well-known to exhibit sensitive dependence on initial conditions, that is the local minimum Z^∞ does in general depend on the initial membership Z^0.

To compare two classifications, Z with m groups (such as the result of the clustering, hardened or not) and Y with $\tilde{m}$ group (such as the official classification), one can first define the $m \times \tilde{m}$ *overlap matrix* $\mathscr{T} = (\tau_{gh})$

$$\tau_{gh} = \sum_{i=1}^{n} f_i z_{ig} y_{ih} \tag{7}$$

whose margins give by construction the group weights $\rho_g = \tau_{g\bullet}$ and $\pi_h = \sum_i f_i y_{ih} = \tau_{\bullet h}$. The matrix $\mathscr{T}$ constitutes a normalized version of the contingency table $N\mathscr{T}$ (where N is the total number of terms in the corpus), whose chi-square attests, expectedly and in all the instances encountered in this work, a very significative dependence between both classifications. Their association can be further investigated by performing a CA on $\mathscr{T}$, the resulting biplots (figures 8 to 14) permitting to identify which groups $g = 1,\ldots,m$ of Z possibly correspond to which groups $h = 1,\ldots,\tilde{m}$ of Y, and to which extent.

5.2 Random initial membership

Starting with random memberships as illustrated in 4.2 permits to test how the algorithm behaves when there is not any preliminary information available on the groups. As depicted in figures 8 and 9 the algorithm produced groups which match surprisingly well the official classifications. This result could imply that the different types of municipalities (in the case of $m = 3$, the official groups being: “Urban (1)”, “Intermediary (2)”, “Rural (3)”) are reflected by the topics present in the text of their Wikipedia page. For $m = 9$ and $m = 25$, the match between the official classification and the detected ones is thinner: it could be the case that some types of official groups are less reflected in the topics that the three broad categories of $m = 3$, for example “urban of a big agglomeration (11)” and “urban of a mean agglomeration (12)”: those categories make sense from a classification perspective, as they correlate to population and density, but are harder to extract from the Wikipedia description.

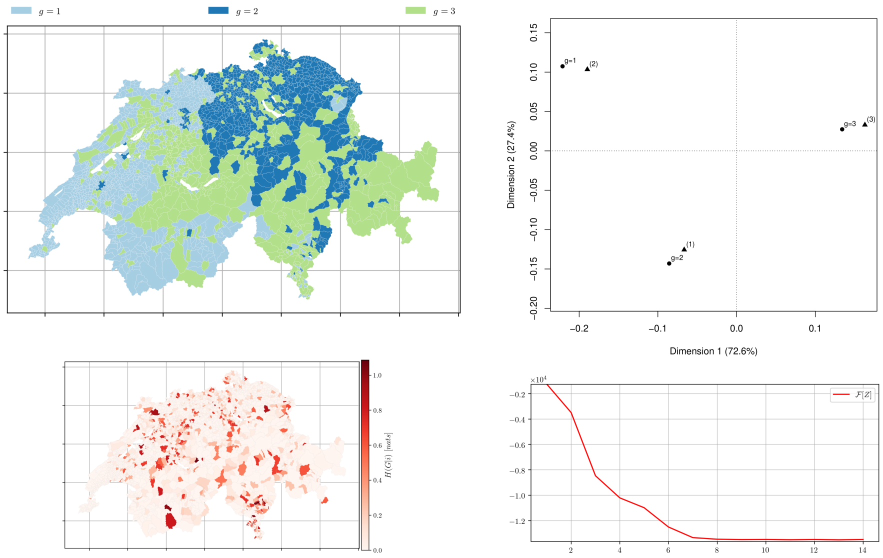

Fig. 8 *Municipalities soft clustering on all the topics* depicts the semi-supervised hard assignment obtained from a random membership matrix Z^0 for $m = 3$ groups using distance matrix D obtained from topic modelling with $k = 3$ after 14 iterations. *Left top* Hard membership. *Left bottom* the conditional entropy of topics $H(R|i)$ showing municipality-topic probability distribution uncertainty. *Right top* the CA biplot between Z^∞ (illustrated by •) and the official classification (illustrated by ▲) with $m = 3$. *Right bottom* The free-energy plot: decreases as the number of iterative steps increases ($\beta = 5, \alpha = 7$).

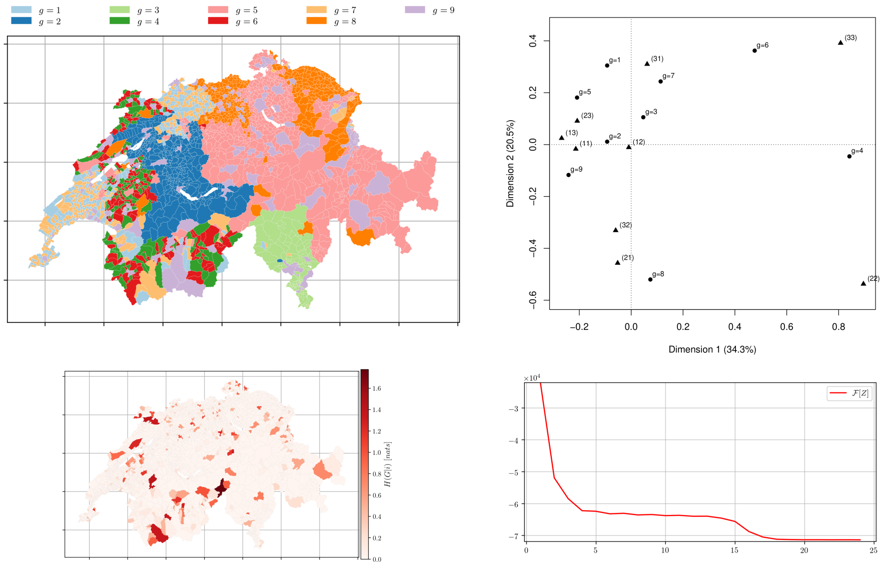

Fig. 9 *Municipalities soft clustering on all the topics* depicts the semi-supervised hard assignment obtained from a random membership Z^0 for $m = 9$ groups using distance matrix D obtained from topic modelling with $k = 9$ after 24 iterations. *Left top* Hard membership. *Left bottom* the conditional entropy of topics $H(R|i)$ showing municipality-topic probability distribution uncertainty. *Right top* the CA biplot between Z^∞ (●) and the official classification (▲) with $m = 9$. *Right bottom* The free-energy plot ($\beta = 20, \alpha = 10$).

5.3 *Official groups as initial membership*

To test if the algorithm minimizes correctly given ideal initial memberships representing the practitioner's knowledge or an official classification, and to verify the intuition that some official categories are more difficult retrieve from the textual description of the municipalities, the initial membership was set to correspond to the official one. For $m = 3$ this initial membership yields, as expected, a better result than the random initial membership. For $m = 9$ and $m = 25$, the choice of initial memberships is less crucial, and the intuition that some groups proposed by the FSO are harder to recover in the corpus of Wikipedia pages is thus confirmed.

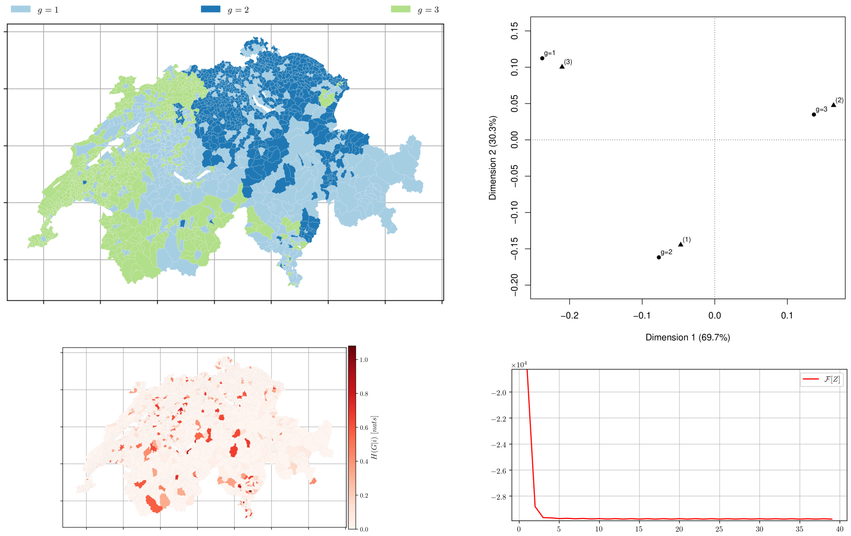

Fig. 10 *Municipalities soft clustering on all the topics* depicts the semi-supervised hard assignment obtained from a official classification Z^0 for $m = 3$ groups using distance matrix D obtained from topic modelling with $k = 3$ after 39 iterations. *Left top* Hard membership. *Left bottom* the conditional entropy of topics $H(R|i)$ showing municipality-topic probability distribution uncertainty. *Right top* the correspondence analysis between Z^∞ (●) and the official classification (▲) with $m = 3$. *Right bottom* The free-energy plot ($\beta = 10, \alpha = 10$).

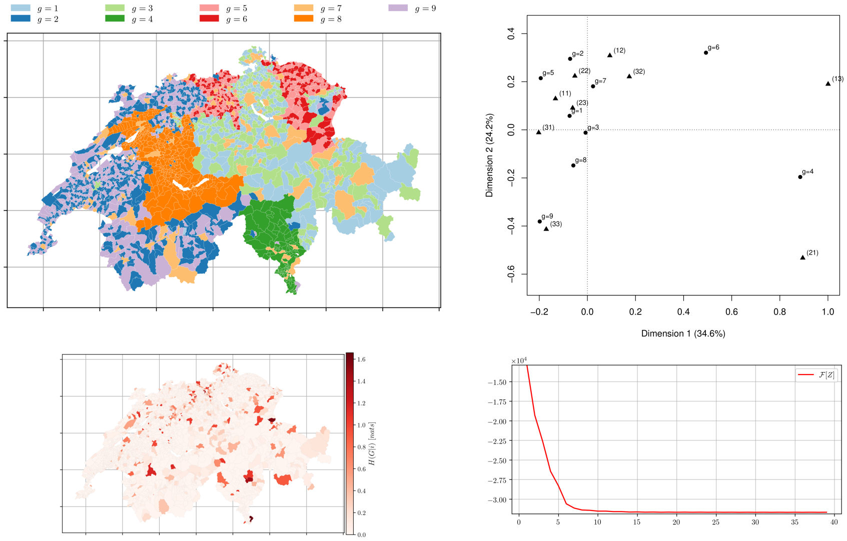

Fig. 11 *Municipalities soft clustering on all the topics* depicts the semi-supervised hard assignment obtained from a official classification Z^0 for $m = 9$ groups using distance matrix D obtained from topic modelling with $k = 9$ after 39 iterations. *Left top* Hard membership. *Left bottom* the conditional entropy of topics $H(R|i)$ showing municipality-topic probability distribution uncertainty. *Right top* the CA biplot between Z^∞ (●) and the official classification (▲) with $m = 9$. *Right bottom* The free-energy plot ($\beta = 10, \alpha = 10$).

5.4 Initial membership based on word frequency

We explored another approach using memberships obtained by using the hard k-means algorithm on the generalized χ^2 distance (see 7) of the municipalities in the term-document matrix. This choice of the initial memberships constitutes an intermediate case between randomness and complete information, and inherits its initial memberships form a distance where the terms can be over-weighted using parameter θ. Initial memberships reflect common usage of rare or frequent words (respectively using $\theta < 1$ or $\theta > 1$) which can be interpreted as a partial knowledge on the textual similarity between municipalities. The results are consistent with the two cases previously observed (see sections 4.2 and 5.3).

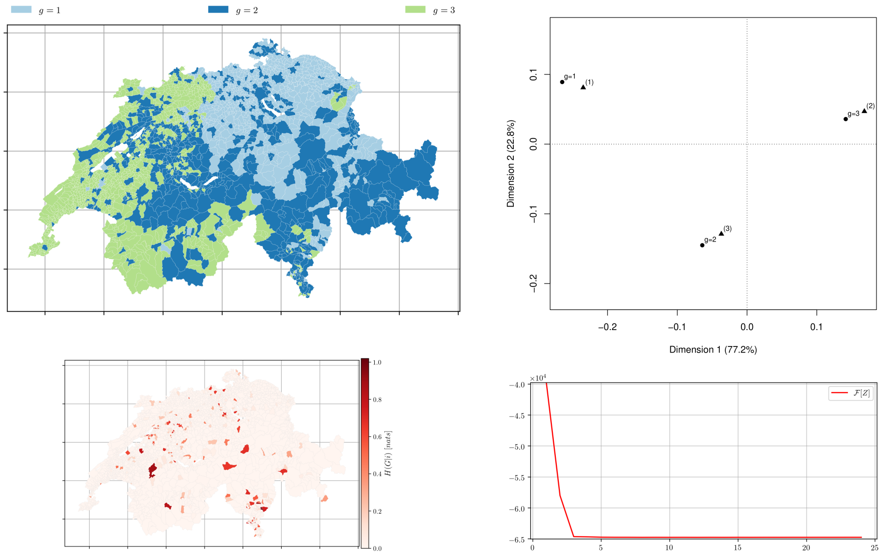

Fig. 12 *Municipalities soft clustering on all the topics* depicts the semi-supervised hard assignment obtained from the k-means performed over the therm-frequency distances with $\theta = 1.01$ on the Z^0 for $m = 3$ groups using χ^2 distances D obtained from topic modelling with $k = 3$ after 39 iterations. *Left top* Hard membership. *Left bottom* the conditional entropy of topics $H(R|i)$ showing municipality-topic probability distribution uncertainty. *Right top* CA biplot between Z^∞ (●) and the official classification (▲) with $m = 3$. *Right bottom* The free-energy plot ($\beta = 20, \alpha = 10$).

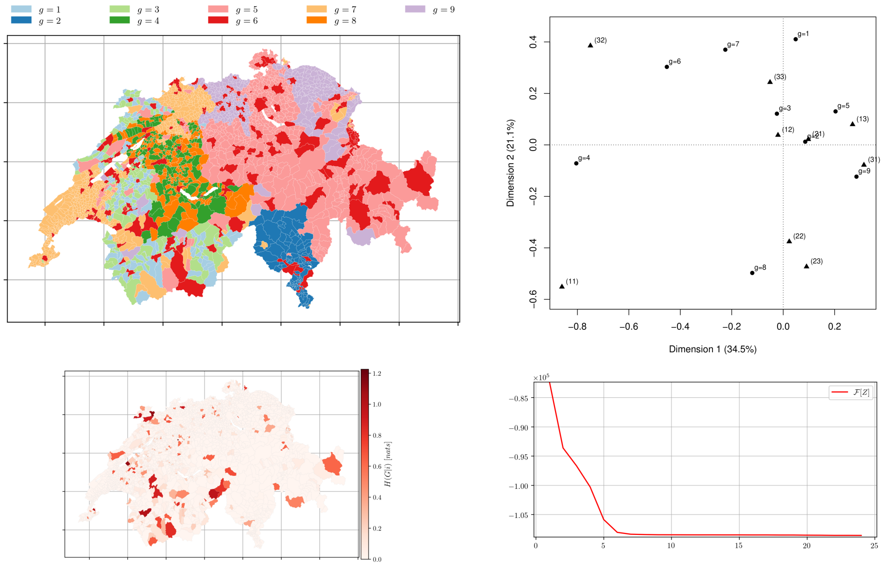

Fig. 13 *Municipalities soft clustering on all the topics* depicts the semi-supervised hard assignment obtained from the k-means performed over the therm-frequency distances with $\theta = 0.99$ on the Z^0 for $m = 9$ groups using χ^2 distances D obtained from topic modelling with $k = 9$ after 24 iterations. *Left top* Hard membership. *Left bottom* the conditional entropy of topics $H(R|i)$ showing municipality-topic probability distribution uncertainty. *Right top* CA biplot between Z^∞ (●) and the official classification (▲) with $m = 9$. *Right bottom* The free-energy plot ($\beta = 30, \alpha = 10$).

5.5 Comparison with a classical approach

How to combine the spatial configuration E of the regions with their textual distances D in order to build a *complex* network on which clustering or boundary detection are then applied is not a trivial question.

An alternative, more classical approach is to combine the textual dissimilarity D_{ij} with the spatial proximity $e_{ij}^{(t)}$ used in graph image segmentation [17, 23] which yields the pairwise region affinity $S = (s_{ij})$ as in:

$$s_{ij} = \frac{e_{ij}^{(t)}}{f_i f_j} \exp(-\lambda\, D_{ij}) \tag{8}$$

where the spatial component $e_{ij}^{(t)}/f_i f_j$ compares the spatial interaction of order t between regions i and j to its expected value under independence. The free parameter $\lambda > 0$ controls the pairwise similarity. The higher s_{ij}, stronger is the interaction along the edge ij.

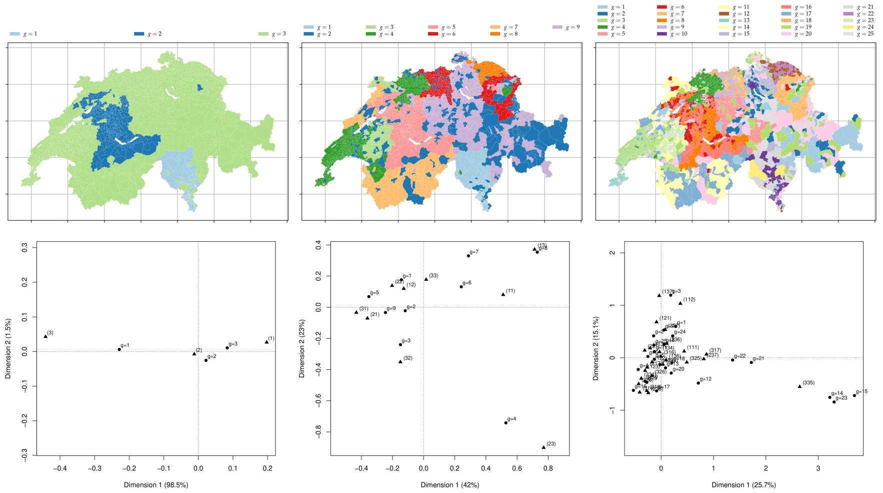

Fig. 14 *Top* Municipalities assignments from the spectral clustering with the parameters, respectively left-right, $m = 3, 9, 25$ for 100 iterations applied on the affinity matrix S (with: $r = 1.0$, $\lambda = 1.0$). *Bottom* the corresponding CA biplot.

For a general comparison we used the well known community detection algorithm Infomap [21], from the `igraph` python package [8] on this network, which turned out to detect $n/2$ communities, irrespectively of the values of parameters. This result could be expected as S yields a complete network and the degrees of municipalities are more or less the same.

Another classical community detection algorithm is spectral clustering [18]. We used the python package `scikit-learn` [19] to perform it on the affinity matrix S. Figure 14 shows interesting results, where the correspondence between memberships obtained form spectral clustering and the official classification are already quite good.

6 Conclusions

This paper exposes and explores the application of the soft clustering algorithm to the exploration of a spatial and thematic corpus based on the Wikipedia pages of Swiss municipalities. We focused the analysis on the impact of differing initial memberships on the results, in order to explore the robustness of the algorithm; the matching of the latter to the official classifications, permitting to incorporate the practitioner's knowledge in the analysis, namely the socio-economical and geographical categorisation of municipalities.

This study has permitted, on one hand, to show that the algorithm strongly depends on the textual or topic distances in use, but is otherwise less sensitive to the initial memberships. On the other hand, the association of the groups computed by the algorithm with the official classification of the municipalities is surprisingly high. Finally, the results demonstrate that the Wikipedia pages of the municipalities constitute a corpus that is both spatially and thematically correlated.

References

[1] Anselin, L.: Local Indicators of Spatial Association-LISA. Geographical Analysis **27**(2), 93–115 (2010)

[2] Bavaud, F.: Generalized factor analyses for contingency tables. In: Classification, Clustering, and Data Mining Applications, pp. 597–606. Springer (2004)

[3] Bavaud, F.: Aggregation invariance in general clustering approaches. Advances in Data Analysis and Classification **3**(3), 205–225 (2009)

[4] Bavaud, F.: Testing spatial autocorrelation in weighted networks: the modes permutation test. Journal of Geographical Systems **3**(15), 233–247 (2013)

[5] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)

[6] 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 Information Systems Theory, Applications and Management, Porto, Portugal, 27-28 April, 2017., vol. 1, pp. 62–69. SciTePress (2017)

[7] Ceré, R., Bavaud, F.: Soft image segmentation: on the clustering of irregular, weighted, multivariate marked networks (2018). Accepted for Springer Book of GISTAM 2017: Communications in Computer and Information Science CCIS series.

[8] Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal **Complex Systems**, 1695 (2006). http://igraph.org [Online; accessed 27-March-2018]

[9] DBpedia: DBpedia (2017). http://dbpedia.org [Online; accessed 27-March-2018]

[10] Deza, M.M., Deza, E.: Encyclopedia of Distances. Springer-Verlag, Berlin Heidelberg (2009)

[11] Egloff, M., Ceré, R.: Soft Textual Cartography Based on Topic Modeling and Clustering of Irregular, Multivariate Marked Networks. In: C. Cherifi, H. Cherifi, M. Karsai, M. Musolesi (eds.) Complex Networks & Their Applications VI, pp. 731–743. Springer (2018)

[12] Fellows, I.: Wordcloud: Word Clouds (2014). R package version 2.5

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

[14] Grady, L., Funka-Lea, G.: Multi-label Image Segmentation for Medical Applications Based on Graph-Theoretic Electrical Potentials. In: M. Sonka, I.A. Kakadiaris, J. Kybic (eds.) Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis, pp. 230–245. Springer (2004)

[15] Grün, B., Hornik, K.: topicmodels: An R Package for Fitting Topic Models. Journal of Statistical Software **40**(13), 1–30 (2011)

[16] Hartigan, J.A., Wong, M.A.: Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) **28**(1), 100–108 (1979)

[17] Lézoray, O., Grady, L. (eds.): Image processing and analysis with graphs: theory and practice. Digital imaging and computer vision series. Taylor & Francis, Boca Raton, FL (2012)

[18] von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing **17**(4), 395–416 (2007)

[19] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[20] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2017). https://www.R-project.org/ [Online; accessed 27-March-2018]

[21] Rosvall, M., Axelsson, D., Bergstrom, C.T.: The map equation. The European Physical Journal Special Topics **178**(1), 13–23 (2009)

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

[23] Solem, J.E.: Programming Computer Vision with Python - Tools and algorithms for analyzing images. O'Reilly (2012)

[24] Wikipedia: Wikipedia, The Free Encyclopedia (2018). http://en.wikipedia.org [Online; accessed 27-March-2018]

[25] 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)

[26] Zecha, L., Kohler, F., Goebel, V.: Niveaux géographiques de la Suisse. Typologie des communes et typologie urbain-rural 2012. Tech. rep. (2017)

7 APPENDIX A: Generalised chi square distance and term-document distance

The generalised χ^2 distance defined in (11) provides a parameter θ which enables to control if the distance should be more sensible to high or low frequencies in the distributions. To define this distance let $U = (u_{il})$ be the $(n \times N)$ document-term matrix, counting the number of occurrences of term l in document i. The relative document-weights f, term-weights v and quotients η are

$$f_i = \frac{u_{i\bullet}}{u_{\bullet\bullet}} \qquad v_l = \frac{u_{\bullet l}}{u_{\bullet\bullet}} \qquad \eta_{il} = \frac{u_{il}\, u_{\bullet\bullet}}{u_{i\bullet}\, u_{\bullet l}} \tag{9}$$

The χ^2 distance between documents i and j is

$$d_{ij} = \sum_l v_l (\eta_{il} - \eta_{jl})^2 \; . \tag{10}$$

And the generalised χ^2 distance is defined as:

$$d_{ij} = \sum_l v_l (\varphi(\eta_{il}) - \varphi(\eta_{jl}))^2 \text{ where } \varphi(\eta) \text{ is any increasing function.} \tag{11}$$

by construction d_{ij} defines a squared Euclidean distance between documents i and j, thus Multidimensional Scaling (MDS) [2] can be performed.

For instance consider $\varphi(\eta) = \eta^\theta$ with $\theta \geq 0$. The case $\theta = 1$ yields the usual χ^2 distance. $\theta > 1$ overweights the contribution of frequent terms, and $\theta < 1$ overweights the contribution of rare terms. The case $\theta = 1/2$ yields the so-called *Hellinger distance* [10], and $\theta \to 0$ yields the *presence-absence dissimilarity*:

$$\lim_{\theta \to 0+} d_{ij}^{(\theta)} = V_{ij^c} + V_{i^c j} \tag{12}$$

where $V_{ij^c} = \sum_{l; l \in i, l \notin j} v_l$ is the total weight of terms present in i but not in j, and $V_{i^c j}$ is defined analogously.