Geographical Exploration and Analysis Extended to Textual Content

Raphaël Ceré
Department of Geography and Sustainability, University of Lausanne, Switzerland
Raphael.Cere@unil.ch

Mattia Egloff
Department of Language and Information Sciences, University of Lausanne, Switzerland
Mattia.Egloff@unil.ch

François Bavaud
Department of Language and Information Sciences & Department of Geography and Sustainability, University of Lausanne, Switzerland
Francois.Bavaud@unil.ch

Abstract

Textual and socio-economical regional features can be integrated and merged by linearly combining the between-regions corresponding dissimilarities. The scheme accommodates for various squared Euclidean socio-economical and textual dissimilarities (such as chi2 or cosine dissimilarities derived from document-term matrix or topic modelling). Also, spatial configuration of the regions can be represented by a weighted unoriented network whose vertex weights match the relative importance of regions. Association between the network and the dissimilarities expresses in the multivariate spatial autocorrelation index δ, generalizing Moran's I, whose local version can be cartographied. Our case study bears on the Wikipedia notices and socio-economic profiles for the 2251 Swiss municipalities, whose weights (socio-economical or textual) can be freely chosen.

2012 ACM Subject Classification Mathematics of computing → Probability and statistics, Information systems → Clustering

Keywords and phrases Spatial autocorrelation, Weighted spatial network, Document-term matrix, Multivariate features, Soft clustering

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.23

Category Short Paper

1 Introduction

Spatial analysis deals with notions of "*where*" (the spatial configuration of regions), "*what*" (the regional features) and "*how much*" (the relative importance of regions, as given by their surface, the population size or terms size). *The aim of this contribution is to propose a formalism and a case study showing how to* ***directly incorporate textual information****, in the frequent situation where each region is described by a text.* In a nutshell, both socio-economic and textual features can be encoded in a dissimilarity matrix between regions, and linearly combined in a flexible way, producing new dissimilarities mixing both kind of features. The latter can be further used for multidimensional scaling, or distance-based clustering.

Socio-economic features can be spatially auto-correlated, and so are the textual features. Section 2 presents a general formalism for assessing and testing spatial autocorrelation and its local indicators, able to deal with multivariate features. Its application requires the

10th International Conference on Geographic Information Science (GIScience 2018).
Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 23; pp. 23:1–23:7
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

dissimilarities to be squared Euclidean, which leaves open many possibilities and variants, in particular regarding the information retrieval processing of the document-term matrix.

The formalism also represents the spatial configuration of the regions as an unoriented weighted network, where the node weights represent the importance of regions, and the edge weights is a measure of accessibility, larger between spatially close regions. Requiring the sum of the edge weights associated to a region to equal the regional weight is natural and mathematically convenient. Among various possible choices, we adopt here the *diffusive weighted specification*, yielding a family of weight-compatible networks index by a single parameter $t > 0$, the diffusion time. The regional weights themselves can be chosen as proportional to the residential population, or proportional to the document sizes, and this choice has a deep impact on the behaviour of the quantities under consideration, as illustrated in the case study presented in section 3.

2 Formalism and definition

We consider a set of n regions, characterized by textual descriptions, as well as by socio-economic features. The former are typically specified by a $n \times v$ document-term matrix X^{text}, giving, after the usual textual pre-processing, the number of occurrences of term $w = 1, \ldots, v$ in the document describing region $i = 1, \ldots, n$. The latter are specified by a $n \times p$ matrix X^{se} contains the p socio-economic features of interest, such as the proportions of inhabitants belonging to specific ages, nationalities, professional types, the proportions of buildings of a given type, etc.

Regions differ by their importance, as specified by relative weights $f_i > 0$ with $\sum_{i=1}^{n} f_i = 1$. Regional weights can be chosen as reflecting the document sizes $f_i^{\mathtt{text}}$, or the population share f_i^{se} as in standard socio-economic geographic analysis. Finally, the spatial configuration of the n connected regions is specified by a binary $n \times n$ adjacency matrix $A = (a_{ij})$.

2.1 A general framework for spatial autocorrelation

Dissimilarities between regional features may, on average, be smaller between spatially close regions, and this precisely constitutes the issue of spatial autocorrelation. A general framework, permitting to attribute differing weights to regions, whose spatial proximity is modelled by a weighted unoriented network, and whose features can be multivariate, relies on two ingredients :

1. a $n \times n$ symmetric joint probability matrix $E = (e_{ij})$, referred to as the *exchange* matrix, giving the probability to select a pair ij of regions, the idea being that e_{ij} is proportional to the relative weights f_i and f_j of the regions, and decreasing with their spatial distance.
2. a $n \times n$ symmetric dissimilarity matrix $D = (D_{ij})$, where $D_{ij} = \|\vec{x}_i - \vec{x}_j\|^2$ is a squared Euclidean distance between suitably normalized multivariate regional features $\vec{x}_i$ and $\vec{x}_j$.

In addition, and crucially, the exchange matrix is required to be *weight compatible*, that is its margins yield the regional weights, that is $e_{i\bullet} = \sum_{j=1}^{n} e_{ij} = f_i$, where f_i can be interpreted as the probability to select region i.

The global inertia, respectively local inertia, measures the average dissimilarity between randomly selected regions, respectively between neighbours. Their comparison provides an autocorrelation index δ which constitutes a multivariate generalization of *Moran's I*. They read, in order,

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

The values of δ range in $[-1, 1]$, and its standardized value $z = (\delta - E_0(\delta))/\sqrt{\mathrm{Var}_0(\delta)}$ can be tested in the normal approximation [2, 3].

Regional dissimilarities D_{ij} can, as in spatial econometrics and quantitative geography, reflect their socio-economic profiles, but also, and more originally, the textual content of their description, or a mixture of both. All the involved similarities should be squared Euclidean, and this constitutes a necessary and sufficient condition for the application of the formalism. For comparison sake, they should also be preliminary standardized as $D_{ij} \leftarrow D_{ij}/\Delta$.

Local multivariate indicators of spatial autocorrelation. Local multivariate indicators of spatial autocorrelation [1], measuring the average scalar product of the deviations at a region and at its neighbours, can be constructed as

$$\delta_i = \frac{(WB)_{ii}}{\Delta} \qquad \text{with} \quad W = \mathrm{diag}(1/f)E \quad \text{and} \quad B = -\frac{1}{2}HDH' \text{ , where } H = I - \mathbf{1}f' \quad (2)$$

and satisfy $\sum_i f_i \delta_i = \delta$. Here W is the row-standardized $n \times n$ matrix of spatial weights, and constitutes the transition matrix of a reversible Markov chain with stationary distribution f. Also, $B = (B_{ij})$ is the $n \times n$ matrix of scalar products $B_{ij} = (\vec{x}_i - \bar{x})'(\vec{x}_j - \bar{x})$ corresponding to the dissimilarities $D_{ij} = \|\vec{x}_i - \vec{x}_j\|^2$, where $\bar{x} = \sum_i f_i \vec{x}_i$.

Spatial configuration: weighted spatial network. In practice, the weight compatible exchange E matrix, specifying the spatial configuration of regions under the form of weighted spatial network, must be constructed from the given regional weights f (which may be taken as f^{text} or f^{se}) and the adjacency matrix A. That is, $E \equiv E(f, A)$, and among differing possibilities, we adopt here the *diffusive kernel construction*, which essentially consists in considering a *time-continuous Markov process* whose infinitesimal generator is given by the Laplacian of the adjacency matrix (e.g. [11, 9]). Imposing weight-compatibility $E\mathbf{1} = f$, as detailed in [2, 3, 4] yields a time-dependent exchange matrix $E(t) = E(f, A, t)$ with limits $\lim_{t\to 0} e_{ij}^{(t)} = f_i \delta_{ij}$ (reducible network made of n disconnected regions) and $\lim_{t\to\infty} e_{ij}^{(t)} = f_i f_j$ (complete weighted network, free of distance-deterrence effects).

Socio-economic dissimilarities between regions. Socio-economic dissimilarities between regions can be obtained as $D_{ij} = (x_i - x_j)^2$, for numerical univariate features x, or as generalized chi-squared dissimilarities $D_{ij} = \sum_{l=1}^{m} \rho_l (q_{il}^{\theta} - q_{jl}^{\theta})^2$ for categorical features with m modalities, where ρ_l is the proportion of modality l, q_{il} the ratio of observed cross-counts to their expected value under independence, and $\theta > 0$ a distortion factor overweighting for $\theta > 1$ (respectively $\theta < 1$) the contribution of high (respectively low) region-modality associations. In any case, all those dissimilarities are squared Euclidean, and so are their p-variate mixtures $D_{ij}^{\mathsf{se}} = \sum_{k=1}^{p} \alpha_k D_{ij}^{(k)}$, where $D^{(k)}$ is the standardized dissimilarity for the k-variable, and $\alpha_k \geq 0$ the freely adjustable corresponding contribution, thus allowing the generation of flexible socio-economic dissimilarities adapted for particular contexts.

Textual dissimilarities between regions. Each region is described by a document, such as historical or geographical notices; or political or administrative documents; or, in our case study, Wikipedia English articles on Swiss municipalities. After usual textual preprocessing (see e.g. [10]), the resulting document-term matrix X^{text}, serves in turn to the generation of textual dissimilarities between regions :

- as straightforward chi-square dissimilarities on N, possibly generalized (see above)

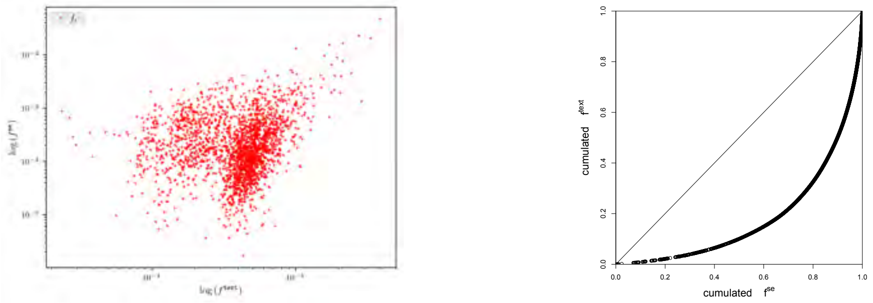

Figure 1 Left: logarithmic scatter plot of the weights f^{se} versus f^{text} illustrates the disparity between population and textual weights. Right: Lorentz curve associated to the Gini coefficient $G = 0.63$ between f^{se} and f^{text}.

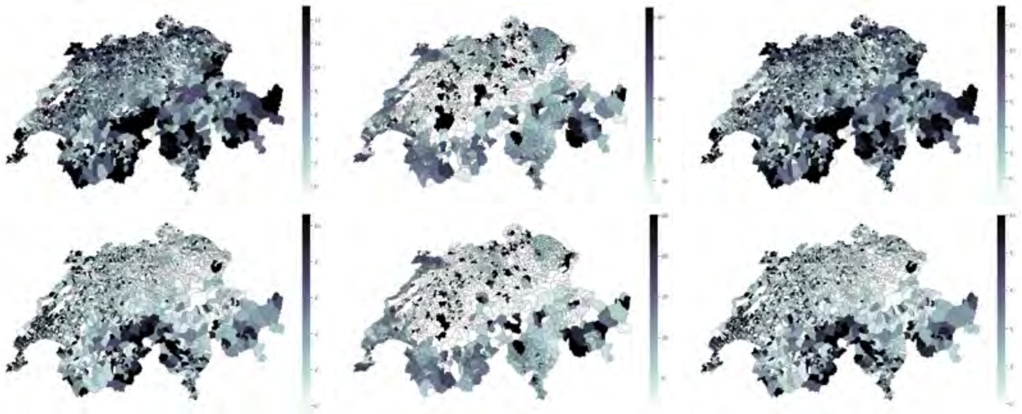

Figure 2 Local indicators of spatial autocorrelation $\delta_i(t)$ of equation (2) for the Swiss municipalities at diffusive time $t = 1$. *Top left* to *right*: dissimilarities are respectively D^{se}, $D^{\chi^2_{\theta=1}}$, and $(D^{\text{se}} + D^{\chi^2_{\theta=1}})/2$, with f^{se} as the reference weight. *Bottom left* to *right*: the resulting local indicators with the same dissimilarities and reference weight f^{text}. A large δ_i indicates strong and parallel feature deviations between municipality i and its neighbours. The notable pattern differences between top and bottom maps reveals the influence of the weight choice.

- from *topic modelling* (see e.g. [5]) on N, yielding in turn membership probabilities (of documents relatively to the topics), on which generalized "topic" chi-square dissimilarities can again be computed.

Socio-economic and textual dissimilarities can be combined as mixtures $\lambda D^{\text{se}} + (1-\lambda) D^{\text{text}}$, where $\lambda \in (0,1)$, which are still squared Euclidean. They can serve at implementing soft k-means clusterings detailed in [6, 7], and extended to textual content in [8].

3 Case study

We illustrate our general approach for spatial autocorrelation upon the $n = 2251$ Swiss municipalities in 2016, exploring the balance between socio-economical and textual features.

Socio-economic dissimilarities of Swiss municipalities: the $p = 6$ socio-economical features X^{se} bearing on sex, age, nationality and civil status of the permanent population (defining the socio-economic weights f^{se}), as well as the count of houses and buildings, constitute census values provided by the FSO. After standardization, their corresponding chi-squared dissimilarities contribute in equal parts to the overall socio-economic dissimilarity $D^{\text{se}} =$

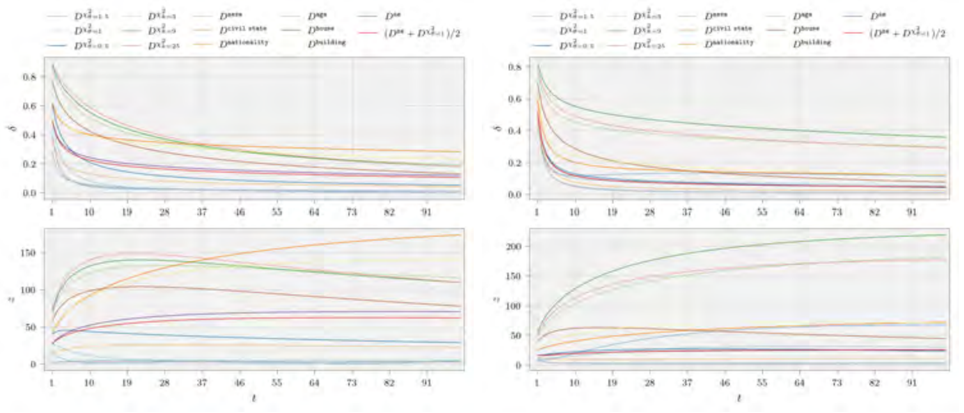

Figure 3 Spatial autocorrelation $\delta(t)$ of equation (1) measured for all Swiss municipalities, at diffusive times $t = 1, \ldots, 99$, for various dissimilarities, with weights f^{se} proportional to the population (left) and f^{text} proportional to the number of terms (right).

$(D^{\text{sex}} + D^{\text{age}} + D^{\text{nationality}} + D^{\text{civil status}} + D^{\text{house}} + D^{\text{building}})/6$.

Textual dissimilarities of Swiss municipalities: for each municipality, we use the Wikipedia pages obtained through the Federal Statistical Office (FSO) number. They are further georeferenced and textually pre-processed (see [8] for more details). Two dissimilarities will be investigated (see section 2): $D^{\chi^2_k}$, resulting from topic modelling with $k = 3, 9, 25$ topics, and $D^{\chi^2_\theta}$, the generalized chi-squared dissimilarity on the original document-term matrix.

Combination of Socio-economic dissimilarities and Textual dissimilarities: the autocorrelation index $\delta(t)$ and its standardized value $z(t)$ are depicted in figure 2 for differing diffusion times $t > 0$, after preliminary choice of the weights f, well contrasted (figure 1), and whose large influence on the analysis is apparent.

Figure 2 depicts the disparate values of the local indicators $\delta_i(t = 1)$, whose range is much larger for the chi2 textual document-term dissimilarities under socio-economic weights, and whose values can be negative, indicating a strong spatial contrast yet to be fully understood. Finally, figure 3 depicts the contrasted behavior of $\delta(t)$ and $z(t)$ for various diffusion times, various dissimilarity choices, and for the two set of weights. Although differing by order of magnitudes, the associated spatial autocorrelations are always significant at level 5% (that is $|z(t)| > u_{.95} = 1.96$), with the exception of D^{sex} and D^{age} for $f = f^{\text{text}}$, and $D^{\chi^2_{\theta=1}}$ for $f = f^{\text{se}}$, which loose their significance for t large.

The "spatial+feature" clustering. The "spatial+feature" clustering method introduced in [6, 7] and extended to textual content in [8] attempts to create clusters containing nodes both strongly connected (as in network clustering) and similar regarding their features (as in distance-based clustering), and does so by running an iterative procedure, decreasing at each step the *free energy* $F[Z]$ (a generalized negative log-likelihood) of the *soft membership matrix* $Z = (z_{ig})$, given the probability that region i belongs to group $g = 1, \ldots, m$. Starting from an initial membership Z^0, the iteration converges to a final membership Z^∞, which constitutes a local minimum of the free energy, and constitutes a generalized soft k-means procedure (spherical Gaussian mixtures) taking into account the spatial configuration of the objects to be clustered.

Figure 4 depicts the final clustering, made hard by assigning each region i to group $G[i] = \arg\max_{g \in \{1,\ldots,m\}} z^\infty_{ig}$, with $m = 9$ groups. In all four cases, the initial membership Z^0

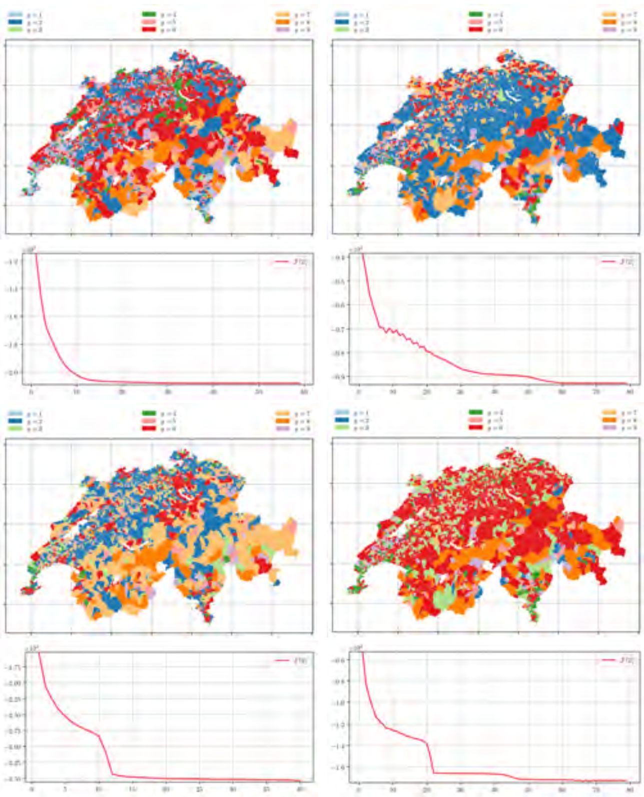

Figure 4 Hard assignment of the final soft attribution Z^{∞} for all Swiss municipalities at diffusive time $t = 1$, for $m = 9$ groups, and decrease of the free energy. Left: socio-economic dissimilarities D^{se} with parameters $\beta = 8, \alpha = 0.1$ (see [7, 8]), and weights f^{se} (top) and f^{text} (bottom). Right: mixed dissimilarities $(D^{\mathsf{se}} + D^{\chi^2_{\theta=1}})/2$ with parameters $\beta = 1.4, \alpha = 0.1$, and weights f^{se} (top) and f^{text} (bottom).

consists of an official attribution of the $n = 2251$ Swiss municipalities in $m = 9$ urban-rural categories, provided by FSO, and updated in 2017 [12].

In guise of conclusion. As illustrated by the case study, the proposed formalism sets up a general methodology able to incorporate directly textual content in the characterization of regions, on equal footing with more usual geographical information such as socio-economic features. A crucial step is the systematic use of squared Euclidean dissimilarities, which can be freely linearly combined. The regional weights can also be chosen as reflecting the population or area regional importance; or, more originally, the regional textual importance – a choice better adapted for e.g. destination image and impressions in tourism studies.

References

1 Luc Anselin. Local indicators of spatial association - LISA. *Geographical analysis*, 27(2):93–115, 1995.

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

3 François Bavaud. Spatial weights: Constructing weight-compatible exchange matrices from proximity matrices. In M. et al. Duckham, editor, *Geographic Information Science*, pages 81–96, Cham, 2014. Springer.

4 François Bavaud, Maryam Kordi, and Christian Kaiser. Flow autocorrelation: a dyadic approach. *Springer Nature 2018*, 2018.

5 David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

6 Raphaël Ceré and François Bavaud. 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*, volume 1, pages 62–69, 2017.

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

8 Mattia Egloff and Raphael Ceré. Soft Textual Cartography Based on Topic Modeling and Clustering of Irregular, Multivariate Marked Networks. In C et al. Cherifi, editor, *Complex Networks & Their Applications VI*, pages 731–743. Springer, 2018.

9 François Fouss, Marco Saerens, and Masashi Shimbo. *Algorithms and models for network data and link analysis*. Cambridge University Press, 2016.

10 Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

11 Alexander J. Smola and Risi Kondor. Kernels and regularization on graphs. In *COLT*, volume 2777, pages 144–158. Springer, 2003.

12 Laurent Zecha, Florian Kohler, and Viktor Goebel. Niveaux géographiques de la Suisse. Typologie des communes et typologie urbain-rural 2012. Technical report, Office fédéral de la statistique (OFS), 2017.