Some reflections about time, durations and transitions when mining complex data. A statistical perspective.

Madalina Olteanu, Université Paris

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Habilitation à diriger des recherches
Université Paris 1 Panthéon Sorbonne

Discipline : Mathématiques appliquées et applications des mathématiques

Soutenue publiquement le 10 décembre 2019, par :

Madalina Olteanu

Some reflections about time, durations and transitions when mining complex data

A statistical perspective

Devant le jury composé de :

Jean-Marc	**Bardet**	Garant	Professeur, Université Paris 1
Christophe	**Biernacki**	Examinateur	Professeur, Université Lille 1
Mark	**Handcock**	Rapporteur	Professeur, UCLA
Céline	**Lévy-Leduc**	Examinatrice	Professeur, AgroParisTech
Mathilde	**Mougeot**	Examinatrice	Professeur, ENSIIE
Fabrice	**Rossi**	Examinateur	Professeur, Université Paris Dauphine
Anne	**Ruiz-Gazen**	Rapportrice	Professeur, Université Toulouse 1

A mes enfants et à mes parents, mes quatre points cardinaux ...

Remerciements

" *Pero uno llama azar*
a su imaginación insuficiente."
I. Vitale

Au delà d'être un document synthétisant des travaux de recherche, ce manuscrit est avant tout le fruit de nombreuses rencontres scientifiques et humaines, de beaucoup de pas en avant et parfois de quelques pas en arrière, d'un cheminement où finalement je pense que le hasard n'intervient que dans la formalisation mathématique. Et puisque tout se remplit de sens lorsqu'on re-trace et qu'on remonte le temps, au commencement il se doit d'y avoir la parole: merci!

Tout d'abord, je voudrais remercier Jean-Marc Bardet et Fabrice Rossi, qui m'ont encouragée et soutenue dans les démarches de cette habilitation, depuis l'idée même de l'envisager et jusqu'à la soutenance. La gentillesse et l'optimisme de l'un, l'humour et le pragmatisme de l'autre, la bienveillance sans faille des deux, m'ont été bien précieux et je leur suis reconnaissante.

Je voudrais remercier très chaleureusement Anne Ruiz-Gazen et Mark Handcock d'avoir accepté et d'avoir pris le temps d'évaluer ce manuscrit. J'ai pris beaucoup de plaisir à l'écrire l'été dernier, sur une île grecque et au bord du Danube, et j'espère vous avoir transmis cela aussi entre les lignes. Un très grand merci également à Céline Lévy-Leduc, Mathilde Mugeot et Christophe Biernacki d'avoir accepté de faire partie du jury et d'être à mes côtés pendant cette épreuve professionnelle importante.

Les résultats résumés dans ce document ne sont les miens qu'en très modeste partie, derrière chaque ligne publiée il y a un travail d'équipe. Je suis reconnaissante et fière d'avoir travaillé avec mes nombreux co-auteurs, que je remercie vivement. Puisqu'il peut être long et délicat de les citer tous, je ne donnerai pas de nom ici, ils se retrouveront et retrouveront nos exploits dans les pages à suivre! Je voudrais néanmoins remercier ici les étudiants - et surtout collègues - m'ayant choisie comme encadrante pour faire un mémoire de M2 ou une thèse. Merci pour votre confiance et pour votre patience, j'apprends à vos côtés! Pour m'avoir (re)mise dans le droit chemin et pour m'avoir éclairée sur la vraie passion que je nourris pour les sciences humaines et sociales, je remercie tout particulièrement les collègues du PIREH.

Durant ces nombreuses et belles années, j'ai eu le bonheur d'être hébergée par deux maisons formidables!

Ma résidence principale, bien située en ville, fut le SAMM. J'ai passé beaucoup de temps entre ses murs et en compagnie de ses locataires, en façonnant ma recherche, ainsi que mes bonnes manières à la française! Je suis heureuse d'avoir partagé toutes ces années avec vous, toutes les discussions scientifiques, les rigolades et les coups de gueule, et aussi toutes les aventures vécues ensemble, du Périgord aux routes cubaines, en passant par les 14 juillet en lointaine banlieue. Si aujourd'hui j'ai un second passeport et deux nationalités, c'est aussi, et surtout, grâce à vous. Un très grand merci tout particulièrement à Marie Cottrell, ma camarade et mère spirituelle, quoique j'espère avoir coupé le cordon depuis le temps! Un grand merci aussi à toi, Marie-Michèle, sans ton investissement et ta bonne humeur, je n'aurais jamais réussi à coordonner un diplôme en apprentissage tout en ayant un semblant de vie personnelle par ailleurs! J'ai une pensée émue et nostalgique pour Christine et Yvonne, les fées des machines informatiques à Paris 1, j'aurais tant aimé pouvoir les inviter aujourd'hui ...

Quant à ma résidence « à la campagne », quel bonheur de respirer de temps en temps de l'air purifié des nombreuses charges administratives, surcharge d'enseignement, et blocages à répétition. J'ai passé deux années en délégation au sein de l'unité MaIAGE et de l'équipe Dynenvie, dans un environnement bien stimulant scientifiquement et très chaleureux – ainsi que très gourmand et joueur. Je remercie bien vivement mes collègues pour leur accueil. Je remercie tout particulièrement Elisabeta Vergu, avec qui j'ai eu le bonheur de travailler dans ma langue maternelle, en nous attaquant à des données massives et difficiles.

Merci à Andrei, tout simplement ...

Contents

Chapter 1

Introduction

" *The saddest aspect of life right now is that it gathers knowledge* (and data!)
faster than society gathers wisdom."
I. Asimov

"So, tell me what's your story in a couple of words!", challenged me the other day my colleague Prof. W. Clark, as we were driving back to Paris from a quantitative geography conference in Luxembourg. This manuscript is an attempt to take up this challenge, while expanding a little on the mathematical and methodological details. I will take this opportunity of summarizing more than ten years of work for also contextualizing it with respect to a more general setting, and for finding the common thread to the many collaborations and projects I was involved in during this time. I will also try to sketch my own perspective on our profession, as shaped over the years, while formulating some personal mid-term projections and expectations.

Let me start with the context and the interesting times we've been living in these last two decades. On the one hand, the big-data blast, and our whole world being overwhelmed by massive amounts of data everywhere. We became, as I heard I. Thomas saying lately, "censors and sensors", and I am not the first nor the last to say that the new sources of data, the new technologies, and the omnipresence and/or omnipotence (?) of artificial intelligence are not only deeply (sic!) changing our research field, but also our behaviors as social beings, at all scales and at all levels of the society. On the other hand, the massive digitization of historical data, combined with the new sources of data from the new media, the social networks etc, and also with the new analytical tools, which challenged the established epistemologies across the social sciences and humanities, and engendered paradigm shifts. Whereas the new masses of data may enhance our understanding of humanity activities across time and space, one may reflect upon the question of a data-driven science, rather than a knowledge-driven one. Will the data *speak for itself* and unveil patterns and precious information without any prior research question, any hypothesis, any theory? One may want to go back to Francis Bacon while looking for an answer!

I have had the chance to be a witness and to experience the delicate process triggered by the massive arrival of data in our common lives and in our scientific research. Scientific and also less scientific debates bring daily in the spotlight the issues related to the fairness of algorithmic decisions, the validity of the algorithmic inferences, the reproducibility of results, the causality, or the privacy and the public interest. Big and/or complex data created a stimulating intellectual bridge between mathematics - and more particularly statistics and machine learning -, and humanities and social sciences, leading to challenging research questions focused on the history of human society, as well as on its present state.

It is in this both complex and inspiring context that I carried out my work, during and after my PhD. My research interests situate at the frontier between computational statistics and machine learning, with a particular interest in analyzing complex data having a temporal component. Although, over the years, I have been involved in many interdisciplinary projets, and studied data coming from various fields, in this manuscript I will only focus on data stemming from humanities and social sciences and on my contributions to analyzing it. I will deliberately skip several aspects of my work, such as the collaborations with the industry – for instance, the methodology developed during Cynthia Faure's PhD, defended last September and that I co-supervised –, or the collaboration with the epidemiologists and mathematicians of MaIAGE, INRA, who hosted me twice, first in 2007, and during a sabbatical these last two years. Although these collaborations are very important and topical to me – I will get back to them at the end of the manuscript in the perspectives section –, I will only single out the case studies I worked on related to humanities and social sciences. The reasons for doing so are multiple.

The first is related to my interest in complex temporal data, the field of humanities and social sciences being amazingly rich from this point of view. The data one has to deal with is not necessarily big, and in some cases rather the contrary. When it comes to historical data, one may be more preoccupied with the quality of the primary sources and hence the quality of the data, the sparsity or the scarcity of it, or with the uncertainty of the temporal instants. Whether the data is small - historical archives - or big - social media -, there is always a variety of sources for complexity in humanities and social sciences, which I find particularly interesting. Focusing the manuscript on case studies in this field will allow me to present several different examples in terms of practical applications, each of them associated to a specific algorithm or methodology that were derived starting from the data itself. How does one use or develop statistical or machine learning algorithms for mining complex data in humanities and social science will be the common thread of this manuscript.

The second reason is more personal and has been stated in the above epigraph. To me, it has to do both with the big data *gold rush*, and the need for more perspective that statistics and machine learning communities are facing these days. As A. Finkel stressed in a Nature opinion column earlier this year, scientists worldwide are "on a treadmill, churning out papers", because of "financial and career incentives". With the revival of artificial intelligence, this is even more so the case in the field of statistical learning. There are, in my opinion, too many *black-box* algorithms and too many *Rube Golberg machines* in our field, and I strongly believe that developing methods designed for handling data from humanities and social sciences, trying to go beyond simply improving a prediction score, understanding mechanisms and causalities, will give one more pause for thought and slow her down a little. I obviously do not claim for the universality of this statement, it is merely an outcome of my collaborations with historians and social scientists over the years, but I do acknowledge that this led me to being more reflexive in my modeling approach.

The following chapters summarize the work published after my PhD, and contain theoretical and methodological results developed for complex temporal data, with applications in humanities and social sciences. Each chapter opens with an introduction giving the context of the work, and summarizing my contributions. Definitions and notations are introduced when needed, as well as references to the related literature. In order to keep the manuscript light and easy to read, I will not go into the technical details of the proofs, they are available in the published papers (a complete list of references, as well as a CV may be found in the Appendix). The four chapters of the manuscript may be read independently.

Chapter 2 is related to the field of computational statistics and time series analysis, and introduces several hidden Markov models and autoregressive switching Markov models for time series of integers, and for bounded real-valued time series. The motivation behind these models was the historical question of the temporality of the Savoy Duchy during the XVIth and XVIIth centuries, and more specifically of the rhythms at which legislation on military logistics was being issued. For each model, an EM estimation procedure was derived, and tested on simulated examples before being trained on the real data.

Chapter 3 uses more theoretical statistics tools and tackles the question of model selection for mixtures and hidden Markov models, and more particularly for regime switching models with autoregressive components. We derive the distribution of the likelihood ratio-test statistic and prove the consistency of a penalized likelihood criterion for selecting the number of regimes, under some general hypothesis. We check these hypothesis in some particular cases, useful for practical applications, such as mixtures of linear autoregressive models or mixtures of multilayer perceptrons.

Chapter 4 takes its root in the field of machine learning, and contains my contributions to the analysis of complex data using kernel and relational self-organizing maps. As I am and was particularly interested in clustering categorical time series and networks, relational algorithms proved to be a very convenient tool for this task. The main contributions address the questions of dimensionality, which is solved using various sparse versions and approximations, as well as that of multiple sources of information, for which we proposed an adapted version. I illustrate the various algorithms using essentially a large longitudinal dataset on career paths for young high-school graduates with a ten year follow-up.

Eventually, Chapter 5 deals with the exploratory analysis of spatial data. Starting from various fine-grained real data (social housing in Paris, ethnic mixing in Los Angeles area, migrant residential distribution in eight European countries), I addressed the issue of residential segregation, and proposed two methodologies for assessing it. The first is based on self-organizing maps and aims at integrating a multidimensional perspective, the second starts from the notion of individual trajectories in the city – in analogy with time series – and introduces new concepts and indices, such as *focal distances* and *distortion coefficients*, which allow to quantify the individual perception of segregation, independently of the scale. This latter chapter contains very recent work, and its mathematical formalization and study are currently undergoing.

Chapter 2

Time segmentation: a matter of events and transitions

" *La science sociale a presque horreur de l'événement.*
Non sans raison : le temps court est la plus capricieuse, la plus trompeuse des durées."
F. Braudel

2.1 Introduction

Let me start this journey with a work on historical time series, carried out mainly in collaboration with Julien Alerini, historian at Université Paris 1 Panthéon Sorbonne. What I had at first considered as an *impossible duet*, turned out to be a challenging and fruitful dialogue, leading eventually to a whole series of results, common research projects and seminars, and involving a whole group of scientists, both from social sciences and from mathematics and computer science. This chapter is thought as a foretaste of the rest of the manuscript, I will mainly focus on the historical questions that one tries to answer starting from the data, the methodological answers the statistician may propose, and eventually the theoretical aspects that have to be settled in order to provide reliable results.

When I met Julien ten years ago, he was investigating the alpine communities in the Duchy of Savoy during the XVIth and XVIIth centuries. These two centuries had been deeply marked by political changes and by several long and intense wars. It had been a period during which the Duchy changed and shaped its structure and its functioning as a state. The historian was interested in characterizing these changes, beyond a chronological listing of events, and, more particularly he wanted to focus on the rhythms at which the Duchy was issuing legislative texts related to military logistics. Empirical observations had shown that periods of conflict were characterized by an important enhancement of the legislative norm, meaning that wars required to increase the mobilisation of financial resources. The state therefore created an important amount of new taxes, as it had to insure accommodation and supplies for the troops. But at which rhythm and which were the steps of this action? After the first discussions, it emerged that using a time series segmentation approach could brig some elements of answer: were there synchronous periods when comparing the production of law with the contexts of war? Or were there discordances between the two, which may inform on the State politics? If the legislative issuance had accelerated before the beginning of a period of belligerence, one may have inferred that the state was preparing for war. Similarly, a prolonged period of an important production of law during peace could have meant either a strengthening of military structures or a long-term transition of politics. Eventually, a period of conflict in which the level of legislative issuance remained low could have meant either a neglect on behalf of the state, or that its military infrastructure was already advanced enough. Therefore, for the historian, the interest in segmenting the data came from the necessity to detach himself from the periodization that is usually offered in the reading of what is traditionally called marking events.

The corpus of data that was to be studied came from the massive work of F-A. Duboin [1] [2], an opus available in the Archives of Torino (Italy), and intended at the restoration of the Savoy legislation, following the Napoleonic age. The work of Duboin is not a compilation of chosen texts, since he collected all records stored in the Piedmontese institutions, in order to establish a legal basis for the restored State. According to [2], this edition would be exhaustive, and few texts would be missing. Hence, we may consider this source as a complete edition of the Sabaudian law, from the XIIIth to the XVIIIth centuries.

Part of following was done during James Ridgway's internship during the summer of 2011, and whom I supervised. Four years later, James brightly defended a PhD in Bayesian statistics in Dauphine University, and he is currently a research associate in a private company.

The results and discussions I will present in this chapter were published as a journal paper [MO5], two peer-reviewed proceedings [MO26, MO36] and one book chapter [MO20].

At this point, I should also provide some insights about the *side effects* of this collaboration. Meeting Julien broadly ment meeting the whole PIREH team, `http://www.pantheonsorbonne.fr/axe-de-recherche/pireh/`, and more particularly Stéphane Lamassé. Almost ten years ago when we started our dialogue and collaboration, speaking broadly about data science and quantitative approaches in history had a blasphemy flavor. Fortunately this is no longer the case nowadays, academic training programs have been created by several universities, and mixing data science and historical research has even become fashionable! I have been quite fortunate to start early enough this collaboration, which brought me to a series of questionings related to how statistical modeling was being used in humanities, how data coming from humanities was being considered by the statisticians, and more specifically questionings related to temporality issues, scale, perception of time, ... Together with Stéphane, we have created a series of half-day interdisciplinary seminars focused on the perception and the modeling of time, `http://samm.univ-paris1.fr/La-temporalite-perceptions-et`, with speakers coming from humanities, mathematical modeling and computer science, and with a case study discussed during each session. This project was funded twice by the Panthéon-Sorbonne scientific board, and several sessions were organized first in 2014, and later on in 2018 and 2019. In close relationship with this, we co-organized with Joseph Rynkiewicz and other colleagues in SAMM the MASHS (Modèles et Apprentissages en Sciences Humaines et Sociales) workshop, in 2014, 2016 and 2018.

Thanks to all these different discussions and seminars, I was brought to collaborate at the writing of a collective book on settlement systems in slow historical time, together with geographers, historians and archeologists from various Parisian institutions. This is the outcome of a working-group hosted by the Labex Dynamite, the book being currently proofread by the editor. Being part of this working group was another opportunity for seizing the diversity and the complexity of problems that humanities may raise for the quantitative science researchers.

Eventually, I should also mention here that time-segmentation techniques proved to be very helpful in other contexts then those related to humanities, and I used them in several industrial projects. Between 2015 and 2019, I co-supervised (with Jean-Marc Bardet) the PhD thesis of Cynthia Faure, aimed at detecting change-points and abnormal behaviors in aircraft health-monitoring. The work was funded by Safran Aircraft Engines, and Cynthia defended her thesis in September 2018. She is currently working as a data scientist for a private company. Starting with October 2018, I am co-supervising (with Fabrice Rossi) a new PhD student, Clément Laroche. This new project is being funded by ANSES, and the goal is to explore a very large and heterogeneous data set containing the levels of contamination of various pesticides and other toxic products in various environments, using techniques related to time-series segmentation, combined with spatial statistics, clustering and other machine-learning related methods. Since I do not wish to present applications in other fields than humanities in this manuscript, I will not detail these projects, the publications related to them are nevertheless being available in the references.

2.2 Time series segmentation: a contextual setting

Throughout this chapter, the modeling issue will consist in segmenting a univariate time series, recorded over a sufficiently large period. This series is supposed to depend, at least partially, on some underlying unobserved process which generates changes, transitions and events, in the observed data. More formally, I shall consider hereafter that $(X_t)_{t\in\mathbb{Z}}$ is a sequence of real valued random variables, either discrete or continuous, possibly with some correlation structure. Two approaches are most commonly used in the literature for segmenting this kind of data, change-point detection (see [3] for a very recent and complete review), and hidden-Markov models (see [4] for a nice and smooth introduction, or [5] for a more formal and rigorous description).

Change-point detection In this case, one assumes that some characteristics of X_t are subject to $K^\star$ abrupt changes occurring at some unknown time instants, $\mathcal{T}^\star = \{t_1^\star < \cdots < t_{K^\star}^\star\}$. Depending on the context, $K^\star$ may be supposed to be known or not, and if the latter it has to be estimated also. Estimating $\mathcal{T}^\star$ amounts to finding the time segmentation $\hat{\mathcal{T}} = \{\hat{t}_1 < \cdots < \hat{t}_K\}$ which minimizes the cost function

$$\mathcal{C}(\mathcal{T} = \{t_1 < \cdots < t_K\}) = \sum_{k=0}^{K} c\left(X_{t_k+1}, \cdots X_{t_{k+1}}\right) \ , \tag{2.1}$$

where $c\left(X_{t_k+1}, \cdots X_{t_{k+1}}\right)$ measures the goodness-of-fit of some given model between t_k and t_{k+1}. If the number of changes K is known, then one has to solve the discrete optimization problem

$$\min_{|\mathcal{T}|=K} \mathcal{C}(\mathcal{T} = \{t_1 < \cdots < t_K\}) \ . \tag{2.2}$$

If the number of changes is unknown, a penalty term allows to achieve a trade-off between complexity and overfitting, and the optimization problem becomes

$$\min_{|\mathcal{T}|\leq K_{\max}} \mathcal{C}(\mathcal{T}) + \text{pen}(\mathcal{T}) \ . \tag{2.3}$$

Without going into the details of different possible cost functions, optimization algorithms, and penalty terms, I would only stress that this approach leads to a crisp segmentation (roughly speaking): no transition between two different regimes is possible, the changes occur at once and one only disposes of the time instant at which the change took place.

Hidden-Markov models (HMM) A hidden Markov model is a particular type of mixture model. In its simplest form, it may be written as a bivariate random process $(S_t, X_t)_{t\in\mathbb{N}}$ such that:

1. S_t is the unobserved *parameter process*, a homogeneous Markov chain, irreducible and aperiodic, valued in a finite state-space $E = \{e_1, ..., e_q\}$ and defined by its transition matrix
$$\Pi = (\pi_{ij})_{i,j=1,...,q} \ , \ \pi_{ij} = \mathbb{P}\left(S_t = e_j | S_{t-1} = e_i\right) \ , \tag{2.4}$$
with $\pi_{ij} > 0$, $\sum_{j=1}^{q} \pi_{ij} = 1$, and by its stationary distribution π^0, $\pi_i^0 = \mathbb{P}\left(S_1 = e_i\right)$, $\forall i = 1, ..., q$;
2. X_t is the observed time series, a real-valued sequence of random variables, such that the distribution of X_t (also called *emission distribution*) depends the current state S_t of the Markov chain only. One usually supposes that, conditionally to S_t, the X_t's are independent. In more complex settings such as autoregressive Markov switching models and variants (see the seminal paper [6] for an example), extra-dependencies at the level of the observed process X_t may be added.

Suppose, for the simplicity and for illustration purposes, that the probability distribution of X_t conditionally to $S_t = e_i$ is $f_{\xi_i} \in \mathcal{F}$, where $\mathcal{F} = \{f_\xi, \xi \in A \subset \mathbb{R}^d\}$ is a parametric family of distributions, and A is a convex set. If the number of states of the Markov chain q is known, then the parameter space of the model is

$$\Theta = \{\theta = (\boldsymbol{\xi}, \Pi) \in A^q \times]0,1[^{q^2}, \forall i \in \{1 \cdots q\}, \sum_{j=1}^{q} \pi_{ij} = 1\} \ , \tag{2.5}$$

where $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_q)$.

If one disposes of a sample of the observed series $(X_1, \cdots, X_T)$ and wishes to infer the parameters, one common approach is to maximize the likelihood through an EM (expectation-maximization) procedure [7]. Indeed, since the path $S_1, \cdots S_T$ is not available, one iteratively optimizes the expected complete log-likelihood, conditionally to the observed data and a given value of the parameter, and obtains an update for the parameter which is then used to feed the algorithm. Eventually, the procedure converges towards a local maximum of the likelihood.

Since I will use the EM procedure several times in this chapter, I will briefly recall here its philosophy and main steps. One starts by writing the complete likelihood conditionally to a given value of the parameter θ:

$$\mathcal{L}(X_1^T, S_1^T; \theta) = \prod_{t=1}^{T} \prod_{i=1}^{q} f_{\xi_i}(X_t)^{\mathbf{1}_{e_i}(S_t)} \prod_{t=2}^{T} \prod_{i,j=1}^{q} \pi_{ij}^{\mathbf{1}_{e_i,e_j}(S_{t-1},S_t)} \times C \ , \tag{2.6}$$

where $X_1^T = (X_1, ..., X_T)$, $S_1^T = (S_1, ..., S_T)$ are the observed series and an associated possible path of the Markov chain, and C is the likelihood of the initial state of the Markov chain. Since the S_t's are not available, one then computes the expected value of the conditional likelihood, for a fixed value $\theta^\star$ of the parameter:

$$\mathbb{E}_{\theta^\star}\left(\ln \mathcal{L}(X_1^T, S_1^T; \theta) | X_1^T\right) = \sum_{t=1}^{T} \sum_{i=1}^{q} \omega_t(e_i) \ln f_{\xi_i}(X_t) + \sum_{t=2}^{T} \sum_{i,j=1}^{q} \omega_{t-1}(e_i, e_j) \ln \pi_{ij} + C \ , \tag{2.7}$$

where $\omega_t(e_i) = \mathbb{P}_{\theta^\star}(S_t = e_i | X_1^T)$ and $\omega_t(e_i, e_j) = \mathbb{P}_{\theta^\star}(S_t = e_i, S_{t+1} = e_j | X_1^T)$. The latter quantities may be tracked thanks to the *forward-backward procedure*, introduced in [8], and allowing to write

$$\omega_t(e_i) = \frac{\mathbb{P}_{\theta^\star}(S_t = e_i, X_1^T)}{\sum_{j=1}^{q} \mathbb{P}_{\theta^\star}(S_t = e_j, X_1^T)} = \frac{\alpha_t(e_i)\beta_t(e_i)}{\sum_{j=1}^{q} \alpha_t(e_j)\beta_t(e_j)} \tag{2.8}$$

and

$$\omega_t(e_i,e_j)=\frac{\mathbb{P}_{\theta^\star}(S_t=e_i,S_{t+1}=e_j,X_1^T)}{\sum_{i,j=1}^q\mathbb{P}_{\theta^\star}(S_t=e_i,S_{t+1}=e_j,X_1^T)}=\frac{\alpha_t(e_i)\beta_{t+1}(e_j)\pi_{ij}^\star f_{\xi_j^\star}(X_{t+1})}{\sum_{i,j=1}^q\alpha_t(e_i)\beta_{t+1}(e_j)\pi_{ij}^\star f_{\xi_j^\star}(X_{t+1})}\ , \tag{2.9}$$

with $\alpha_t(e_i)=\mathbb{P}_{\theta^\star}(S_t=e_i,X_1^t)$ and $\beta_t(e_i)=\mathbb{P}_{\theta^\star}(X_{t+1}^T|S_t=e_i)$, $t=1,...,T$ and $i=1,...,q$. The forward probabilities $\alpha_t(e_i)$ are computed recursively from the conditions:

1. $\alpha_{t+1}(e_i)=f_{\xi_i^\star}(X_{t+1})\times\sum_{j=1}^q\pi_{ji}^\star\alpha_t(e_j)$, $\forall i=1,...,q,t=1,...,T-1$;
2. $\alpha_1(e_i)=f_{\xi_i^\star}(X_1)\pi_i^{\star,0}$;

whereas the backward probabilities $\beta_t(e_i)$ verify:

1. $\beta_t(e_i)=\sum_{j=1}^q\pi_{ij}^\star\beta_{t+1}(e_j)f_{\xi_j^\star}(X_{t+1})$, $\forall i=1,...,q,t=1,...,T-1$;
2. $\beta_T(e_i)=1$.

In practice, forward and backward probabilities tend to zero or infinity exponentially fast in the recursions, so that one needs to apply some kind of normalisation (make them sum to one over i, for example). The EM procedure alternates two steps, repeated until convergence:

E-step: for a given $\theta^\star$, compute $Q(\theta|\theta^\star)=\mathbb{E}_{\theta^\star}\left(\ln\mathcal{L}(X_1^T,S_1^T;\theta)|X_1^T\right)$ with the *forward-backward* procedure.
M-step: find θ maximizing $Q(\theta|\theta^\star)$ and feed it as a new $\theta^\star$ in the *E-step*.

In general, for well-known classes of distributions, the *E-step* is quite straightforward, as well as the *M-step*, where analytical expressions of the updates of θ may be computed. If the optimisation step does not lead to an exact solution, one may use numerical algorithms to approach it. These will generally slow down the algorithm which is by default very slow (EM usually converges at a linear rate), but the results will remain consistent. I will note here that an alternative and well performing approach to deal with hidden Markov model is to use a Bayesian framework and MCMC algorithms. Since this out of scope for the subsequent, I will not go into any details, but a very nice comparison of the two approaches, with all the pros and cons may be found in [9].

Eventually, the goal is to use hidden Markov models in the perspective of time series segmentation. Once the parameters of the model have been estimated, the next step is to compute the *optimal* sequence of states having produced the observed series. Since there are several possible optimality criteria, the choice that was made here was to use Viterbi's algorithm [10]. The procedure builds upon the best scores (probabilities) along a single path, at time t, and which accounts for the first t observations and ends in state e_i:

$$\delta_t(i)=\max_{S_1,...,S_{t-1}}\mathbb{P}_{\theta^\star}\left(S_1,...,S_{t-1},S_t=e_i,X_1,...,X_t\right)\ , \tag{2.10}$$

which verify the recursion formula

$$\delta_t(j)=\left[\max_{i=1,...,q}\delta_{t-1}(i)\pi_{ij}^\star\right]f_{\xi_j^\star}(X_t)\ . \tag{2.11}$$

The probabilities $\delta_t(i)$ being computed, the *optimal* path (along with its associated probabilities) may be computed by backtracking:

$$S_t^\star=\psi_{t+1}\left(S_{t+1}^\star\right)\ ,\ t=T-1,T-2,...,1, \tag{2.12}$$

where $\psi_t(j)=\arg\max_{i=1,...,q}\delta_{t-1}(i)\pi_{ij}^\star$, $t=2,...,T$ and $j=1,...,q$.

So once a hidden Markov model has been trained, one may either look at the crisp segmentation given by the *optimal* path $\{S_1^\star,...,S_T^\star\}$, or at the estimated probabilities of being in one state or another, which allows, particularly in the context of historical data, to characterize transitions and estimate their duration. This property which allows one to characterize with more details the interpretability of results made the use of hidden Markov models much more appealing for our time series, when compared to change-point detection approaches.

The historical data As mentioned in the introduction of this chapter, the corpus of data was extracted from the work of F.-A. Duboin [1], a compilation of ancient law-texts from the thirteenth century until 1798. Studying this corpus allows one to appreciate the legislative activity of the Duchy in developing infrastructure and logistics administration. Between 1559 and 1661, there were 5 775 texts issued by the Duke, the councils, the supreme courts or their agents. The texts related in one way or another to the movement, the supply and the accommodation of military troops were tagged as related to military logistics. The final military logistics

data consisted of 472 documents, representing 8.17% of the whole legislation. The corpus of documents was represented by the historian as a bivariate time series, a first component containing the counts of texts related to military logistics, and a second component containing the counts of all legislative texts together. After having considered three different time scales for the analysis (monthly, quarterly, yearly), each of them having pros and cons both from a historical and statistical perspective, we decided to use the monthly representation, as providing a fine representation of the data, although this lead to an inflation of zeros in the series.

When computing some basic descriptive statistics related to the diplomatic situation, as shown in Table 2.1, one may notice significant differences between periods of peace and periods of conflict, and more particularly in the corpus of documents related to military logistics. An analysis of variance clearly indicated that being at peace or being at war had a strong impact on legislative issuance. These results confirmed the thesis in classical historiography, stating that war is an explanatory factor for the expansion of the state. The close relationship between the issuance of law texts related to military logistics, and the state being at war or at peace is obvious. It is in times of war that the state must feed and lodge a maximum of troops, while in a difficult military, economical and political context. Hence, the average amount of documents on military logistics almost doubles between the periods of peace and of war.

	War (45.7% of the data)			Peace (54.3% of the data)		
	Min	Mean	Max	Min	Mean	Max
Entire legislation	0	4.87	17	0	4.05	18
Military logistics	0	0.54	4	0	0.28	4
Ratios	0	0.11	1	0	0.07	1

Table 2.1: Legislative output conditionally to the diplomatic situation. The statistics are computed on the series of the entire legislation, on the series of texts related to military logistics, and on the series of ratios between the two. The mean values are significantly different in all cases (p-$value < 0.05$).

The statistics above show that the activity of the state depends on the belligerence situation. However, this immediate conclusion is not sufficient. In addition to this, one needs to understand which is the temporality of the Sabaudian military logistics, and whether this is synchronous or not with the whole process of producing law. Also, one needs to know whether there are cycles corresponding to those of war and peace and whether these are disconnected or not completely synchronous. The approaches used to our knowledge for studying time series in quantitative history and based on ARIMA-type models, [11] [12] [13] [14] , do not allow to answer the above questions. The main reason is that they cannot take into account the possible existence of irregular cycles or of various regimes in the behavior of the time series. In order to capture these specific features of the data, we shall prefer to use models with Markov-switching regimes instead.

Several new models based on hidden Markov chains were introduced during this collaboration. They were all designed in order to take into account the specificities of the data. We started by studying the univariate time series of counts related to legislation on military logistics, which showed an over-dispersion in zero and a significant auto-correlation structure. For dealing with the increased mass of zero, we first proposed a hidden-Markov model with zero-inflated Poisson distributions. Next, we addressed the dependency issue, by introducing an auto-regressive switching-Markov model based on INAR-type architectures. Later, we were interested in modeling the bivariate time series and we considered the ratios between the series on military logistics and the global series. This lead us to generalize the hidden Markov model with zero-inflated Beta distributions introduced in [15], by considering any (finite) number of hidden states, and then to introduce zero-and-one inflated Beta distributions in the hidden Markov model. In the following sections, I will briefly describe each of these models and their estimation procedure, and provide some illustrations on simulations. A discussion will follow on the insights brought by each of these models on the historical data, and I will conclude by some work in progress and ideas to be developed in the future.

2.3 Zero-inflated Poisson - hidden Markov models (ZIP-HMM)

For segmenting a time series of counts with an important mass of zeros, we designed a hidden Markov model having zero-inflated Poisson laws as emission distributions. To my knowledge, this model was completely new in the literature, except for [16] who had introduced a very particular version of it one year before, a two-state hidden-Markov model, with one component being equal to zero, and the other being distributed according to a Poisson law. In the following, consider $(S_t)_{t\in\mathbb{N}}$ a homogeneous Markov chain, irreducible and aperiodic, valued in a finite state-space and defined according to the Equation 2.4, and $(X_t)_{t\in\mathbb{N}}$ the observed data, an integer-valued time series, such that X_t conditionally to S_t are independent, and such that the distribution of

X_t, conditionally to S_t is a zero-inflated Poisson ZIP(η_i, λ_i):

$$\mathbb{P}\left(X_t = k | S_t = e_i\right) = \eta_i 1_{\{0\}}(k) + (1-\eta_i)\,\frac{e^{-\lambda_i}\lambda_i^k}{k!},\ \forall k \in \mathbb{N}. \tag{2.13}$$

The parameter space associated to this model is:

$$\Theta = \{\theta = (\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\pi}) \in]0,1[^q \times (\mathbb{R}^+)^q \times]0,1[^{q^2}, \forall i \in \{1 \cdots q\}, \sum_{j=1}^{q} \pi_{ij} = 1\}, \tag{2.14}$$

where $\boldsymbol{\eta} = (\eta_1, ..., \eta_q)$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_q)$ represent the parameters characterizing the q ZIP distributions, while $\boldsymbol{\pi} = (\pi_{ij})_{i,j=1,...,q}$ is the transition matrix of the hidden Markov chain.

We introduced this model in [MO36], and later trained it for the historical data described above in [MO5]. Let me mention here that if the number of hidden states q is known, then Proposition 3.1 in [17] ensures that the HMM model with emission distributions defined by Equation 2.13 and with the parameter space defined by Equation 2.14 is *identifiable*, modulo a permutation of the hidden states.

Estimation procedure For estimating θ from an observed sample $X_1^T = (X_1, ..., X_T)$, and under the hypothesis that q is known and fixed, we train an EM procedure. In the context of ZIP-HMM models, the algorithm described in the section above cannot be applied directly, since a ZIP(η_i, λ_i) distribution is a mixture between a Poisson $\mathcal{P}(\lambda_i)$, and a Dirac. Thus, the complete likelihood in Equation 2.6 has to be written in terms of the hidden path $S_1^T = (S_1, ..., S_T)$, but also in terms of a supplementary hidden random process $Z_1^T = (Z_1, ..., Z_T)$, which, conditionally to S_t is Bernoulli distributed, $Z_t | S_t = e_i \sim \mathcal{B}er(\eta_i)$, $t = 1, ..., T$. We suppose that according to the value of Z_t, one gets a *structural* zero, or a Poisson distribution, and that (X_t, Z_t) are independent, conditionally to the hidden process S_t. Under this framework, one may write the complete likelihood as:

$$\mathcal{L}(Z_1^T, X_1^T, S_1^T; \theta) = \prod_{t=1}^{T}\prod_{i=1}^{q} f(X_t, Z_t | S_t = e_i; \theta)^{\mathbf{1}_{e_i}(S_t)} \prod_{t=2}^{T}\prod_{i,j=1}^{q} \pi_{ij}^{\mathbf{1}_{e_i,e_j}(S_{t-1},S_t)} \times C\ , \tag{2.15}$$

where

$$f(X_t, Z_t | S_t = e_i; \theta) = \eta_i^{\mathbf{1}_{Z_t=1}} (1-\eta_i)^{\mathbf{1}_{Z_t=0}} \left(\frac{e^{-\lambda_i}\lambda_i^{X_t}}{X_t!}\right)^{\mathbf{1}_{Z_t=0}}, \tag{2.16}$$

and $C = \prod_{i=1}^{q} \left(\pi_i^0\right)^{\mathbf{1}_{e_i}(S_1)}$ is the likelihood of the initial state of the Markov chain.

Proposition 2.3.1 *E-step: the expected value of the conditional complete likelihood may be expressed as:*

$$\begin{array}{rcl} Q(\theta|\theta^\star) & = & \mathbb{E}_{\theta^\star}\left(\ln \mathcal{L}(X_1^T, Z_1^T, S_1^T; \theta) | X_1^T\right) \\ & = & \sum_{t:X_t>0}\sum_{i=1}^{q} \omega_t(e_i)\{\ln(1-\eta_i) - \lambda_i + X_t \ln(\lambda_i) - \ln(X_t!)\} \\ & + & \sum_{t:X_t=0}\sum_{i=1}^{q} \xi_i^\star \omega_t(e_i) \ln(\eta_i) + (1-\xi_i^\star)\omega_t(e_i)\{\ln(1-\eta_i) - \lambda_i\} \\ & + & \sum_{t=2}^{T}\sum_{i,j=1}^{q} \omega_{t-1}(e_i, e_j) \ln \pi_{ij} + C\ , \end{array}$$

where $\xi_i^\star = \frac{\eta_i^\star}{\eta_i^\star + (1-\eta_i^\star)e^{-\lambda_i^\star}}$, $\omega_t(e_i) = \mathbb{P}_{\theta^\star}(S_t = e_i | X_1^T)$ *and* $\omega_t(e_i, e_j) = \mathbb{P}_{\theta^\star}(S_t = e_i, S_{t+1} = e_j | X_1^T)$ *are computable with the backward-forward procedure, and* C *is a constant standing for the log-likelihood of the initial state* S_1*.*

Proposition 2.3.2 *M-step: by maximizing the expected log-likelihood* $Q(\theta|\theta^\star)$ *above, and with the previous notations, one gets the exact analytical updates,*

$$\hat{\pi}_{ij} = \frac{\sum_{t=1}^{T-1} \omega_t(e_i, e_j)}{\sum_{t=1}^{T} \omega_t(e_i)}\ ;\ \hat{\eta}_i = \frac{\xi_i^\star \sum_{t:X_t=0} \omega_t(e_i)}{\sum_{t=1}^{T} \omega_t(e_i)}$$

$$\hat{\lambda}_i = \frac{\sum_{t:X_t>0} \omega_t(e_i) \times X_t}{\sum_{t:X_t=0} \omega_t(e_i)(1-\xi_i^\star) + \sum_{t:X_t>0} \omega_t(e_i)}.$$

Simulation study The quality of the estimates and the speed of convergence of the algorithm were tested empirically on several simulated examples. For each of the following scenarios, all parameters are kept fixed, except for one of them which is allowed to take values on a grid. For each parameter configuration, and for sample sizes ranging from 500 to 10 000, 10 000 different samples were simulated. In all cases, the mean squared error (MSE) is computed and reported. The results are illustrated in Tables 2.2, 2.3 and 2.4. They were globally very stable, showing relatively low MSE, decreasing with the sample size.

π_{11} / T	0.1	0.3	0.4	0.5	0.6	0.8	0.9
500	0.0247	0.0286	0.0317	0.0350	0.0415	0.0541	0.0550
1 000	0.0021	0.0054	0.0084	0.0131	0.0081	0.0210	0.0260
5 000	0.0003	0.0015	0.0026	0.0058	0.0110	0.0019	0.0008
10 000	0.0001	0.0008	0.0018	0.0050	0.0105	0.0012	0.0008

Table 2.2: MSE ($\pi_{22} = 0.6, \eta_1 = 0.2, \lambda_1 = 0.5, \eta_2 = 0.2, \lambda_2 = 3$)

λ_1 / T	0.1	0.5	1	5	10	14
500	0.0498	0.0417	0.0732	0.0199	0.0028	0.0397
1 000	0.0085	0.0190	0.0320	0.0103	0.0154	0.0280
5 000	0.0133	0.0193	0.0036	0.0018	0.0030	0.0039
10 000	0.0019	0.0094	0.0010	0.0010	0.0010	0.0019

Table 2.3: MSE ($\pi_{11} = 0.4, \pi_{22} = 0.6, \eta_1 = 0.2, \eta_2 = 0.2, \lambda_2 = 3$)

η_1 / T	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
500	0.003	0.006	0.009	0.013	0.012	0.002	0.029	0.052	0.010
1 000	0.001	0.003	0.004	0.006	0.007	0.009	0.013	0.027	0.009
5 000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.016
10 000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.008

Table 2.4: MSE ($\pi_{11} = 0.4, \pi_{22} = 0.6, \lambda_1 = 0.5, \eta_2 = 0.2, \lambda_2 = 3$)

2.4 Integer-valued autoregressive (INAR(p)) - HMM's

Since the data we started from showed a strong dependency structure, we introduced a hidden-Markov model with extra-dependencies in the X_t's, by using an autoregressive architecture. We thus combine a finite-state hidden-Markov model with integer-valued autoregressive models, as defined in [18], [19] and [20]. We introduced and studied this new model in [MO5].

First, let me recall that an INAR(p) process is a sequence of integer-valued random variables $(X_t)_{t\in\mathbb{Z}}$, verifying

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \cdots + \alpha_p \circ X_{t-p} + \varepsilon_t \ , \tag{2.17}$$

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ is an iid sequence of random variables valued in $\mathbb{N}$ and having a finite second order moment. Usually, ε_t is considered to be distributed according to a Poisson or a negative Binomial (we used a Poisson noise in the subsequent). For all $i = 1, ..., p$,

$$\alpha_i \circ X_{t-i} = \sum_{k=1}^{X_{t-i}} \xi_{i,k} \ , \tag{2.18}$$

is the Steutel-van Harn thinning operator introduced in [21] and $\xi_{i,k}$ are independent and distributed according to a Bernoulli distribution of parameter α_i. Hence, conditionally to X_{t-i}, $\alpha_i \circ X_{t-i}$ is a Binomial distribution with parameters X_{t-i} and α_i. Furthermore, it is supposed that the $\xi_{i,k}$ are independent for all i and for all k, and are independent of X_{t-i} and ε_t. With these assumptions, the conditional distribution of X_t with respect

to $X_{t-1} = x_{t-1}, ..., X_{t-p} = x_{t-p}$ may be written as follows:

$$f(x_t|x_{t-p}^{t-1}) = \sum_{i_1=0}^{x_t \wedge x_{t-1}} C_{x_{t-1}}^{i_1} \alpha_1^{i_1} (1-\alpha_1)^{x_{t-1}-i_1} \sum_{i_2=0}^{x_t - i_1 \wedge x_{t-2}} C_{x_{t-2}}^{i_2} \alpha_2^{i_2} (1-\alpha_2)^{x_{t-2}-i_2} \cdots \sum_{i_p=0}^{(x_t - i_1 - \cdots - i_{p-1}) \wedge x_{t-p}} C_{x_{t-p}}^{i_p} \alpha_p^{i_p} (1-\alpha_p)^{x_{t-p}-i_p} \frac{e^{-\lambda} \lambda^{x_t - i_1 - \cdots - i_p}}{(x_t - i_1 - \cdots - i_p)!} \quad , \tag{2.19}$$

where $x_{t-p}^{t-1} = (x_{t-p}, ..., x_{t-1})$.

The model We introduce the hybrid model INAR(p)-HMM as a bivariate process $(S_t, X_t)_{t\in\mathbb{Z}}$ such that $(S_t)_t$ is a latent process and a homogeneous Markov chain, irreducible and aperiodic, defined as in Equation 2.4; and $(X_t)_{t\in\mathbb{N}}$ is the observed time series, valued in $\mathbb{N}$. Furthermore, the observed time-series X_t is supposed to be INAR(p), conditionally to S_t:

$$(X_t|S_t = e_i) = \alpha_{1,i} \circ X_{t-1} + \alpha_{2,i} \circ X_{t-2} + \cdots \alpha_{p,i} \circ X_{t-p} + \varepsilon_{i,t}, \tag{2.20}$$

where $\varepsilon_{i,t} \sim \mathcal{P}(\lambda_i)$, a Poisson distribution with parameter $\lambda_i > 0$. Here, it is supposed that the lag p is identical for all states of the Markov chain. Hence, the parameter space may be written as:

$$\Theta = \{\theta = (\boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\pi}) \in]0,1[^{p\times q} \times (\mathbb{R}^+)^q \times]0,1[^{q^2} \text{ and } \forall i \in \{1 \cdots q\}, \sum_{j=1}^{q} \pi_{ij} = 1\}, \tag{2.21}$$

where $\boldsymbol{\alpha} = (\alpha_{l,i})_{l=1,...,p;i=1,...,q}$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_q)$ represent the parameters characterizing the q INAR(p) models, while $\boldsymbol{\pi} = (\pi_{ij})_{i,j=1,...,q}$ is the transition matrix of the hidden Markov chain. We suppose in the following that q and p are known and fixed parameters.

The estimation procedure is carried out again using an EM algorithm.

E-step For a given time-series $X_{-p+1}^T = (X_{-p+1}, ..., X_T)$ and for a given Markov-chain path $S_1^T = (S_1, ..., S_T)$, and with the notations $\omega_t(e_i) = \mathbb{P}_{\theta^\star}\left(S_t = e_i | X_{-p+1}^T\right)$ and $\omega_t(e_i, e_j) = \mathbb{P}_{\theta^\star}\left(S_t = e_i, S_{t+1} = e_j | X_{-p+1}^T\right)$, the expected conditional complete log-likelihood may be written as

$$\begin{aligned} Q(\theta|\theta^\star) &= \mathbb{E}_{\theta^\star}\left(\ln \mathcal{L}(X_{-p+1}^T, S_1^T; \theta) | X_{-p+1}^T\right) \\ &= \sum_{t=1}^{T} \sum_{i=1}^{q} \omega_t(e_i) \ln f(X_t | X_{t-p}^{t-1}, S_t = e_i; \theta) + \sum_{t=2}^{T} \sum_{i,j=1}^{q} \omega_{t-1}(e_i, e_j) \ln \pi_{ij} + C \end{aligned}$$

where $f(X_t|X_{t-p}^{t-1}, S_t = e_i; \theta)$ is the conditional density defined in Equation 2.19 for the parameter values in Equation 2.20. The technical difficulty here resides in the practical computation of the log-likelihood, and more particularly of the conditional density $f(X_t|X_{t-p}^{t-1}, S_t = e_i; \theta)$, for a general time-lag, p. In the algorithmic implementation, this difficulty was solved by using recursive programming.

M-step Let us remark that the left term of $Q(\theta|\theta^\star)$ depends on the parameters of the INAR(p) models only, the $\alpha_{l,i}$'s and λ_i's, while the right term depends on the transition probabilities π_{ij} only. Hence, as previously, the maximization step can be performed by independently maximizing each term. For the latter, we obtain the usual expressions for the updates:

$$\hat{\pi}_{ij} = \frac{\sum_{t=1}^{T-1} \omega_t(e_i, e_j)}{\sum_{t=1}^{T} \omega_t(e_i)} \quad ,$$

where $\omega_t(e_i, e_j)$ and $\omega_t(e_i)$ are being computed with the Baum-Welch forward-backward algorithm. For the first term, the maximization cannot be carried analytically, because of the complexity of the conditional distribution $f(X_t|X_{t-p}^{t-1}, S_t = e_i; \theta)$. The optimization will be then performed numerically. The constraints on the α's and λ's are first removed by re-parameterizing as follows:

$$\gamma_i = \ln \lambda_i \ , \ \beta_{l,i} = \ln\left(\frac{\alpha_{l,i}}{1-\alpha_{l,i}}\right) \ , \ \forall i = 1, ..., q \ , l = 1, ...p,$$

The maximization is then performed using the Nelder-Mead algorithm. Because of this additional numerical optimization, the EM algorithm is heavier in terms of computational time, but the results on simulations are satisfactory, as shown next.

Simulation study The EM algorithm proposed above for INAR(p) - HMM models is tested next on several simulated examples. For each of the following scenarios, all parameters are kept fixed, except for one of them which is allowed to take values on a grid. Since the implemented algorithm is much slower than the previous one, only 500 different trainings were performed for each scenario. The sample size is either equal to 100 or 500. In all cases, the mean squared error (MSE) is computed and reported.
Scenario A The datais simulated according to a HMM-INAR(1) with two states for the hidden Markov chain. The parameters kept constant are $\pi_{11} = 0.2, \alpha_1 = 0.2, \lambda_1 = 1, \alpha_2 = 0.1, \lambda_2 = 4$. The remaining parameter, the transition probability π_{22} takes values in the interval $]0, 1[$. The results are given in Table 2.5.

π_{22} T	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
100	0.023	0.086	0.028	0.026	0.034	0.036	0.044	0.038	0.081
500	0.005	0.004	0.004	0.007	0.010	0.012	0.014	0.020	0.026

Table 2.5: Mean squared error - scenario A

Scenario B The data is simulated according to a HMM-INAR(1) with two states for the hidden Markov chain. The parameters kept constant are $\pi_{11} = 0.2, \pi_{22} = 0.4, \alpha_1 = 0.2, \lambda_1 = 1, \lambda_2 = 4$. The remaining parameter, α_2 takes values in the interval $[0.1, 0.5]$. The results are given in Table 2.6.

α_2 T	0.1	0.2	0.3	0.4	0.5
100	0.029	0.066	0.045	0.053	0.066
500	0.014	0.011	0.018	0.016	0.008

Table 2.6: Mean squared error - scenario B

Scenario C The data is simulated according to a HMM-INAR(1) with two states for the hidden Markov chain. The parameters kept constant are $\pi_{11} = 0.2, \pi_{22} = 0.4, \alpha_1 = 0.2, , \alpha_2 = 0.1, \lambda_2 = 4$. The remaining parameter, λ_1 takes values in the interval $[0.1, 7]$. The results are given in Tables 2.7 and 2.8.

λ_1 T	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
100	0.026	0.046	0.075	0.144	0.122	0.161	0.258	0.187	0.170
500	0.003	0.006	0.002	0.019	0.024	0.045	0.043	0.057	0.085

Table 2.7: Mean squared error - scenario C

λ_1 T	2	3	4	5	6	7
100	0.330	0.543	0.352	0.492	0.594	0.467
500	0.069	0.074	0.092	0.178	0.299	0.319

Table 2.8: Mean squared error - scenario C

In all scenarios, the MSE decreases with the sample size and the results are globally satisfying. The worst behavior of the algorithm appears in Scenario C, for larger values of λ. This corresponds to a larger variance of the noise, ε_t, thereby the poor results in this case are not surprising. The convergence of the EM algorithm is also globally slower than for the previous model, ZIP-HMM. For example, one initialization of the EM algorithm for Scenario A takes (in average) 2.80 seconds for 100 observations, 6.28 for 500, 11.53 for 1 000 and 95.70 for 5 000.

2.5 Beta-inflated HMM's for time series of proportions

In statistical modeling, the two common approaches for dealing with continuous proportions are, on the one hand, a logistic transformation of the data [22], and, on the other hand, the use of specific probability distributions such as Beta or Dirichlet [23]. However, both of these approaches have a major drawback, since they do not take into account the possibility of an over-dispersion in the limit values, 0 and/or 1. During the last ten years, this issue has been addressed by several authors, who proposed either further transforming the data [24], or introducing specific probability masses in 0 and/or 1, hence using zero-and/or-one Beta Inflated distributions.

The latter approach has been intensively studied during the last five years, mainly in a regression context [25], [26].

In the context of data one wishes to segment and for which several latent regimes are suspected, it is of interest to train a well suited hidden Markov model, built on zero and/or one Beta inflated distributions. These models are not completely new in the literature, [15] recently introduced a two-state hidden Markov model with emission distributions given by zero-inflated Beta's. Our contribution consisted first in generalizing [15]'s proposal to any (finite) number of states [27], and second in replacing zero-inflated Beta's by zero-and-one inflated Beta distributions [MO26]. In the following, I will briefly describe the emission distributions that will be used hereafter and the estimates of their parameters, and afterwards I will introduce the associated hidden Markov models, the estimation procedure, and some empirical results on simulations.

2.5.1 Zero-and/or-one Beta-inflated distributions

Zero-inflated Beta distributions

Consider a Bernoulli latent random variable Y, $Y \sim \mathcal{B}(\eta)$, and X a second random variable such that

$$\left\{ \begin{array}{l} X|(Y=1) \sim \delta_0 \\ X|(Y=0) \sim \mathcal{B}e(\alpha,\beta) \end{array} \right. , \tag{2.22}$$

where $\eta \in]0,1[$, $\alpha, \beta > 0$. Then, the marginal distribution of X is a zero-inflated Beta distribution, $\mathcal{ZIB}(\eta,\alpha,\beta)$. Its density [1] may then written as

$$f_{\mathcal{ZIB}}(x;\eta,\alpha,\beta) = \eta^{\mathbb{1}_{x=0}} \left((1-\eta)\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \right)^{\mathbb{1}_{x\in]0,1[}} . \tag{2.23}$$

Consider now $X_1^T = (X_1, ..., X_T)$ an i.i.d. T-sample of $\mathcal{ZIB}(\eta,\alpha,\beta)$. Since the log-likelihood $\mathcal{L}(X_1^T;\eta,\alpha,\beta)$ may be written as a product $\mathcal{L}_1(X_1^T;\eta)\mathcal{L}_2(X_1^T;\alpha,\beta)$ which components are independent in terms of parameters, one easily gets the estimate of the mixing parameter,

$$\hat{\eta} = \frac{\sum_{t=1}^T \mathbb{1}_{X_t=0}}{\sum_{t=1}^T \mathbb{1}_{X_t=0} + \sum_{t=1}^T \mathbb{1}_{X_t\in]0,1[}} = \frac{T_0}{T}, \tag{2.24}$$

where $T_0 = \sum_{t=1}^T \mathbb{1}_{X_t=0}$. Since the maximization of the second term does not lead to analytical expressions for the Beta parameters, moment estimates may be used instead:

$$\tilde{\alpha} = \tilde{\mu}\tilde{\phi} \ , \ \tilde{\beta} = (1-\tilde{\mu})\tilde{\phi}, \tag{2.25}$$

where

$$\tilde{\mu} = \frac{1}{T-T_0} \sum_{X_t\in]0,1[} X_t, \ \tilde{\phi} = \frac{\tilde{\mu}(1-\tilde{\mu})}{s^2} - 1, \text{ and } s^2 = \frac{1}{T-T_0} \sum_{X_t\in]0,1[} (X_t - \tilde{\mu})^2 . \tag{2.26}$$

Zero-and-one inflated Beta distributions

Next, if the Beta distribution is mixed with a Bernoulli distribution such as to add probability masses both in 0 and 1, one obtains the zero-and-one inflated Beta distribution, $\mathcal{ZOIB}(\eta,\gamma,\alpha,\beta)$, where $\eta,\gamma \in]0,1[$ and $\alpha,\beta > 0$. In this case, if $Y \sim \mathcal{B}(\eta)$ is a Bernoulli latent random variable, then the conditional distribution of $X \sim \mathcal{ZOIB}(\eta,\gamma,\alpha,\beta)$ is such that

$$\left\{ \begin{array}{l} X|(Y=1) \sim \mathcal{B}(\gamma) \\ X|(Y=0) \sim \mathcal{B}e(\alpha,\beta) \end{array} \right. . \tag{2.27}$$

The marginal density[2] of X is

$$f_{\mathcal{ZOIB}}(x;\eta,\gamma,\alpha,\beta) = (\eta\gamma)^{\mathbb{1}_{x=1}} (\eta(1-\gamma))^{\mathbb{1}_{x=0}} ((1-\eta)f_B(x;\alpha,\beta))^{\mathbb{1}_{x\in]0,1[}}, \tag{2.28}$$

[1] the density is taken with respect to the probability measure $\lambda + \delta_0$, where λ is the Lebesgue measure on $[0,1]$, and δ_0 is the Dirac mass in 0.

[2] the density is taken with respect to the probability measure $\lambda + \delta_0 + \delta_1$, where λ is the Lebesgue measure on $[0,1]$, and δ_0 and δ_1 are Dirac masses in 0 and 1.

where $\eta \in]0,1[$ is the mixture parameter, $\gamma \in]0,1[$ is the Bernoulli-distribution parameter and $\alpha, \beta > 0$ are the Beta-distribution parameters.

Following the same approach as above, and for $X_1^T = (X_1, ..., X_T)$ an i.i.d. T-sample of $\mathcal{ZOIB}(\xi)$, one may derive immediately the maximum likelihood estimates for the mixture and for the Bernoulli parameter, and the moment-estimates for the Beta parameters:

$$\hat{\eta} = \frac{\sum_{t=1}^{T} \mathbb{1}_{X_t \in \{0,1\}}}{\sum_{t=1}^{T} \mathbb{1}_{X_t \in \{0,1\}} + \sum_{t=1}^{T} \mathbb{1}_{X_t \in]0,1[}} = \frac{T_{01}}{T}, \tag{2.29}$$

where $T_{01} = \sum_{t=1}^{T} \mathbb{1}_{X_t \in \{0,1\}}$.

$$\hat{\gamma} = \frac{\sum_{t=1}^{T} \mathbb{1}_{X_t = 1}}{\sum_{t=1}^{T} \mathbb{1}_{X_t = 1} + \sum_{t=1}^{T} \mathbb{1}_{X_t = 0}} = \frac{T_1}{T_{01}}, \tag{2.30}$$

where $T_1 = \sum_{t=1}^{T} \mathbb{1}_{X_t = 1}$. The estimates for α and β are almost identical to those in Equations 2.25 and 2.26, except for T_0 which must be replaced by T_{01}.

2.5.2 Zero-inflated Beta hidden-Markov models (ZIB-HMM)

As in the previous sections, let $(S_t)_t$ be a homogeneous Markov chain, irreducible and aperiodic, defined as in Equation 2.4. Consider also $(X_t)_t$, the observed time series, representing continuous proportions, and valued in $[0,1[$. Furthermore, suppose that X_t are independent conditionally to S_t, and that X_t conditionally to S_t are distributed according to zero-inflated Beta distributions, $\mathcal{ZIB}(\eta_i, \alpha_i, \beta_i)$, with $(\eta_i, \alpha_i, \beta_i) \in]0,1[\times]0,+\infty[^2$.

For a fixed number of states q in the hidden Markov chain, the set of possible values for the parameters may then be written as:

$$\Theta = \left\{ \theta = ((\eta_i, \alpha_i, \beta_i)_{i=\overline{1,q}}, \Pi) \in (]0,1[\times(\mathbb{R}_+^\star)^2)^q \times]0,1[^{q^2}, \sum_{j=1}^{q} \pi_{ij} = 1 \right\} \tag{2.31}$$

I will mention here that, although we fixed the number of states, identifiability issues arise in this case, since mixtures of Beta distributions are generally not identifiable. In order to remove this issue, one should fix one of the two parameters in the Beta distribution. This constitutes some work in progress that we are currently doing. Nevertheless, since the results on the historical data were quite meaningful despite this issue, I chose to present and discuss them.

Estimation procedure For a fixed number of states q, the estimation is carried out using the EM algorithm. By denoting $X_1^T = (X_1, ..., X_T)$ and $S_1^T = (S_1, ..., S_T)$ a complete T-sample of data, the complete likelihood may be easily factorized as

$$\mathcal{L}(X_1^T, S_1^T; \theta) = \mathcal{L}_1(X_1^T, S_1^T; \boldsymbol{\eta}) \mathcal{L}_2(X_1^T, S_1^T; \boldsymbol{\alpha}, \boldsymbol{\beta}) \mathcal{L}_3(X_1^T, S_1^T; \Pi) \times C, \tag{2.32}$$

where $\boldsymbol{\eta} = (\eta_1, ..., \eta_q)$, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_q)$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)$, and C is the likelihood of the initial state of the Markov chain, S_1.

Proposition 2.5.1 *Thanks to the factorization of the complete likelihood, the optimization* (M-step) *may be performed by independently maximizing each term of the expected conditional likelihood,*

$$Q(\theta|\theta^\star) = \mathbb{E}_{\theta^\star}\left[\ln \mathcal{L}(X_1^T, S_1^T; \theta) | X_1^T\right] = Q_1(\boldsymbol{\eta}|\theta^\star) + Q_2(\boldsymbol{\alpha}, \boldsymbol{\beta}|\theta^\star) + Q_3(\Pi|\theta^\star). \tag{2.33}$$

With $\omega_t(e_i) = \mathbb{P}_{\theta^\star}(S_t = e_i | X_1^T)$ *and* $\omega_t(e_i, e_j) = \mathbb{P}_{\theta^\star}(S_t = e_i, S_{t+1} = e_j | X_1^T)$ *computed using the forward-backward procedure, one gets exact expressions for* $\boldsymbol{\eta}$ *and* Π*,*

$$\hat{\eta}_i = \frac{\sum_{X_t=0} \omega_t(e_i)}{\sum_{t=1}^{T} \omega_t(e_i)} \ , \ \hat{\pi}_{ij} = \frac{\sum_{t=1}^{T-1} \omega_t(e_i, e_j)}{\sum_{t=1}^{T} \omega_t(e_i)} \ , \tag{2.34}$$

and may numerically optimize Q_2 *for computing the updates of* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\beta}$*.*

In the following, instead of using a numerical optimization for Q_2 as in [15], we prefer to directly plug in moment estimates:

$$\tilde{\alpha}_i = \tilde{\mu}_i\tilde{\phi}_i \ , \ \tilde{\beta}_i = (1-\tilde{\mu}_i)\tilde{\phi}_i, \tag{2.35}$$

where

$$\tilde{\mu}_i = \frac{\sum_{X_t\in]0,1[}\omega_t(e_i)X_t}{\sum_{X_t\in]0,1[}\omega_t(e_i)} \ , \ \tilde{\phi}_i = \frac{\tilde{\mu}_i(1-\tilde{\mu}_i)}{s_i^2} - 1,$$
$$s_i^2 = \frac{\sum_{X_t\in]0,1[}\omega_t(e_i)\left(X_t-\tilde{\mu}_i\right)^2}{\sum_{X_t\in]0,1[}\omega_t(e_i)}. \tag{2.36}$$

When injected in the EM procedure, this approach will avoid some numerical issues. But, at the same time, the convergence of the algorithm will no longer be immediately guaranteed, although we may assess it at least numerically for the moment.

Simulation study For each of the following scenarios and for sample sizes ranging from 500 to 1 000, 100 different trajectories of a two-state ($q = 2$) $\mathcal{ZIB}$-HMM are simulated. The values of the parameters used for the simulations are the following :

$$\Pi = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}, (\alpha_1,\alpha_2) = (1;0.5), (\beta_1,\beta_2) = (1;2),$$

and $(\eta_1,\eta_2) = (\eta_1, 0.8)$, where $\eta_1 \in \{0.1, 0.3, 0.5, 0.7\}$. The results are detailed in Tables 2.9, 2.10, 2.11 and 2.12 below. In each case are reported the mean values of the estimates, as well as their standard errors and medians. We also provide the squared bias and the ratio of errors in the a posteriori identification of the hidden regimes (mean-values, standard errors and medians). According to the simulations, the model is quite well estimated when the proportion of zeros is not too large (less than 50% of the data).

	$T = 500$		$T = 1000$	
$\hat{\Pi}$	0.84(*0.18*)	0.16(*0.18*)	0.84(*0.14*)	0.16(*0.14*)
	0.90	**0.10**	**0.90**	**0.10**
	0.16(*0.16*)	0.84(*0.16*)	0.16(*0.15*)	0.84(*0.15*)
	0.11	**0.89**	**0.10**	**0.90**
$\hat{\alpha}_1, \hat{\alpha}_2$	1.03(*0.25*)	0.69(*0.94*)	0.98(*0.13*)	0.56(*0.16*)
	1.03	**0.56**	**0.99**	**0.52**
$\hat{\beta}_1, \hat{\beta}_2$	1.00(*0.14*)	2.60(*1.94*)	0.99(*0.10*)	2.15(*1.08*)
	0.99	**2.17**	**1.00**	**1.95**
$\hat{\eta}_1, \hat{\eta}_2$	0.13(*0.13*)	0.76(*0.13*)	0.14(*0.13*)	0.76(*0.13*)
	0.09	**0.78**	**0.10**	**0.80**
$Bias(\theta)^2$	1.26(*1.92*)		0.77(*0.88*)	
	0.72		**0.54**	
$\%ERR$	12.8(*16*)		13.3(*16.5*)	
	7.6		**7.2**	

Table 2.9: Simulation results for $\eta_1 = 0.1$ and 100 time-series of length T. Mean, standard error (italics) and median (bold) of the estimates.

2.5.3 Zero-and-one inflated Beta hidden-Markov models (ZOIB-HMM)

Again, let $(S_t)_t$ be a homogeneous Markov chain, irreducible and aperiodic, defined as in Equation 2.4, and consider $(X_t)_t$, the observed time series, representing continuous proportions, and valued this time in $[0,1]$. X_t are supposed to be independent conditionally to S_t, and X_t conditionally to S_t are distributed according to zero-and-one inflated Beta distributions, $\mathcal{ZOIB}(\xi_i)$, with $\xi_i = (\eta_i, \gamma_i, \alpha_i, \beta_i) \in]0,1[^2\times]0,+\infty[^2$.

For a fixed number of states q in the hidden Markov chain, the set of possible values for the parameters may then be written as:

$$\Theta = \{\theta = ((\xi_i)_{i=\overline{1,q}}, \Pi) \in (]0,1[^2\times(\mathbb{R}_+^\star)^2)^q\times]0,1[^{q^2}, \sum_{j=1}^{q}\pi_{ij} = 1, \forall i = \overline{1,q}\} \tag{2.37}$$

	$T = 500$		$T = 1000$	
$\hat{\Pi}$	0.73(*0.21*)	0.27(*0.21*)	0.78(*0.18*)	0.22(*0.18*)
	0.84	**0.16**	**0.88**	**0.12**
	0.24(*0.20*)	0.76(*0.20*)	0.21(*0.18*)	0.79(*0.18*)
	0.14	**0.86**	**0.11**	**0.89**
$\hat{\alpha}_1, \hat{\alpha}_2$	8.15(*69.16*)	0.61(*0.25*)	1.00(*0.21*)	0.57(*0.17*)
	0.98	**0.58**	**1.03**	**0.55**
$\hat{\beta}_1, \hat{\beta}_2$	3.09(*20.46*)	2.04(*1.33*)	1.01(*0.10*)	1.96 (*0.91*)
	0.98	**1.61**	**1.01**	**1.74**
$\hat{\eta}_1, \hat{\eta}_2$	0.33(*0.20*)	0.73(*0.18*)	0.32(*0.15*)	0.76(*0.12*)
	0.30	**0.78**	**0.30**	**0.79**
$Bias(\theta)^2$	8.65(*7.20*)		0.87(*0.61*)	
	1.09		**0.72**	
$\%ERR$	27.60(*19.60*)		22.30(*15.03*)	
	20.80		**15.20**	

Table 2.10: Simulation results for $\eta_1 = 0.3$ and 100 time-series of length T. Mean, standard error (italics) and median (bold) of the estimates.

	$T = 500$		$T = 1000$	
$\hat{\Pi}$	0.65(*0.24*)	0.35(*0.24*)	0.67(*0.21*)	0.33(*0.21*)
	0.65	**0.35**	**0.64**	**0.36**
	0.28(*0.20*)	0.72(*0.20*)	0.30(*0.21*)	0.70(*0.21*)
	0.25	**0.75**	**0.31**	**0.69**
$\hat{\alpha}_1, \hat{\alpha}_2$	1.20(*1.83*)	4.94(*22.33*)	1.60(*7.41*)	2.44(*2.53*)
	0.89	**0.67**	**0.82**	**0.59**
$\hat{\beta}_1, \hat{\beta}_2$	1.06(*0.85*)	19.48(*97.73*)	1.24(*3.20*)	3.97 (*14.29*)
	0.97	**1.42**	**0.98**	**1.52**
$\hat{\eta}_1, \hat{\eta}_2$	0.52(*0.25*)	0.73(*0.20*)	0.47(*0.24*)	0.75(*0.22*)
	0.48	**0.79**	**0.47**	**0.80**
$Bias(\theta)^2$	19.95(*99.91*)		4.76(*20.30*)	
	1.35		**1.26**	
$\%ERR$	37.00(*15.60*)		37.38(*14.86*)	
	32.80		**33.80**	

Table 2.11: Simulation results for $\eta_1 = 0.5$ and 100 time-series of length T. Mean, standard error (italics) and median (bold) of the estimates.

	$T = 500$		$T = 1000$	
$\hat{\Pi}$	0.50(*0.24*)	0.50(*0.24*)	0.53(*0.22*)	0.47(*0.22*)
	0.48	**0.52**	**0.55**	**0.45**
	0.44(*0.22*)	0.56(*0.22*)	0.46(*0.22*)	0.54(*0.22*)
	0.43	**0.57**	**0.48**	**0.52**
$\hat{\alpha}_1, \hat{\alpha}_2$	9.36(*58.31*)	6.41(*42.23*)	1.95(*4.34*)	4.37(*36.77*)
	0.62	**0.61**	**0.80**	**0.55**
$\hat{\beta}_1, \hat{\beta}_2$	1.83(*8.38*)	10.13(*66.59*)	1.23(*1.63*)	8.26 (*63.58*)
	0.91	**1.24**	**0.97**	**1.18**
$\hat{\eta}_1, \hat{\eta}_2$	0.63(*0.32*)	0.67(*0.31*)	0.70(*0.26*)	0.69(*0.26*)
	0.70	**0.78**	**0.75**	**0.75**
$Bias(\theta)^2$	20.50(*97.15*)		10.21(*73.26*)	
	1.46		**1.25**	
$\%ERR$	47.60(*7.00*)		47.98(*6.05*)	
	48.00		**47.80**	

Table 2.12: Simulation results for $\eta_1 = 0.7$ and 100 time-series of length T. Mean, standard error (italics) and median (bold) of the estimates.

Estimation procedure The estimation is carried out again using an EM algorithm. By denoting $X_1^T = (X_1, ..., X_T)$ and $S_1^T = (S_1, ..., S_T)$ a complete T-sample of data, the complete likelihood may be easily factorized

as

$$\mathcal{L}(X_1^T, S_1^T; \theta) = \mathcal{L}_1(X_1^T, S_1^T; \boldsymbol{\eta})\mathcal{L}_2(X_1^T, S_1^T; \boldsymbol{\gamma})\mathcal{L}_3(X_1^T, S_1^T; \boldsymbol{\alpha}, \boldsymbol{\beta})\mathcal{L}_4(X_1^T, S_1^T; \Pi) \times C, \tag{2.38}$$

where $\boldsymbol{\eta} = (\eta_1, ..., \eta_q)$, $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_q)$, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_q)$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)$, and C is the likelihood of the initial state of the Markov chain, S_1.

Proposition 2.5.2 *Thanks to the factorization of the complete likelihood, the optimization* (M-step) *may be performed by independently maximizing each term of the expected conditional likelihood,*

$$Q(\theta|\theta^\star) = \mathbb{E}_{\theta^\star}\left[\ln \mathcal{L}(X_1^T, S_1^T; \theta)|X_1^T\right] = Q_1(\boldsymbol{\eta}|\theta^\star) + Q_2(\boldsymbol{\gamma}|\theta^\star) + Q_3(\boldsymbol{\alpha}, \boldsymbol{\beta}|\theta^\star) + Q_4(\Pi|\theta^\star). \tag{2.39}$$

With $\omega_t(e_i) = \mathbb{P}_{\theta^\star}(S_t = e_i|X_1^T)$ *and* $\omega_t(e_i, e_j) = \mathbb{P}_{\theta^\star}(S_t = e_i, S_{t+1} = e_j|X_1^T)$ *computed with the forward-backward procedure, one gets exact expressions for* $\boldsymbol{\eta}$, $\boldsymbol{\gamma}$, *and* Π,

$$\hat{\eta}_i = \frac{\sum_{X_t \in \{0,1\}} \omega_t(e_i)}{\sum_{t=1}^T \omega_t(e_i)} \ , \ \hat{\gamma}_i = \frac{\sum_{X_t = 1} \omega_t(e_i)}{\sum_{X_t \in \{0,1\}} \omega_t(e_i)} \ , \ \hat{\pi}_{ij} = \frac{\sum_{t=1}^{T-1} \omega_t(e_i, e_j)}{\sum_{t=1}^T \omega_t(e_i)} \ . \tag{2.40}$$

and may numerically optimize Q_3 *for computing the updates of* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\beta}$.

In the following, instead of using a numerical optimization for Q_3 as in [15], we prefer to directly plug moment estimates. This leads to the same updates as in Equations 2.35 and 2.35.

Simulation study In order to empirically test the quality of the estimates and the convergence rate, the algorithm was trained on several simulated examples. For each of the following scenarios and for sample sizes ranging from 500 to 1 000, 100 different trajectories of a two-state ($q = 2$) $\mathcal{ZOIB}$-HMM were simulated. The values of the parameters used for the simulations are the following :

$$\Pi = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}, (\alpha_1, \alpha_2) = (1; 0.5), (\beta_1, \beta_2) = (1; 2), (\gamma_1, \gamma_2) = (0.5; 0.9),$$

and $(\eta_1, \eta_2) = (\eta_1, 0.8)$, where $\eta_1 \in \{0.1, 0.3, 0.5, 0.7\}$. The results are detailed in Tables 2.13, 2.14, 2.15 and 2.16 below. In each case are reported the mean values of the estimates, as well as their standard errors and medians. We also provide the squared bias and the ratio of errors in the a posteriori identification of the hidden regimes (mean-values, standard errors and medians).

	$T = 500$		$T = 1000$	
$\hat{\Pi}$	0.82(*0.18*)	0.18(*0.18*)	0.84(*0.16*)	0.16(*0.16*)
	0.89	**0.11**	**0.89**	**0.11**
	0.19(*0.20*)	0.81(*0.20*)	0.16(*0.16*)	0.84(*0.16*)
	0.11	**0.89**	**0.11**	**0.89**
$\hat{\alpha}_1, \hat{\alpha}_2$	0.99(*0.27*)	0.74(*0.66*)	0.99(*0.14*)	0.54(*0.15*)
	0.96	**0.60**	**1.00**	**0.52**
$\hat{\beta}_1, \hat{\beta}_2$	0.98(*0.11*)	2.70(*1.83*)	1.00(*0.11*)	2.11(*0.75*)
	0.99	**2.35**	**1.00**	**2.09**
$\hat{\gamma}_1, \hat{\gamma}_2$	0.55(*0.22*)	0.88(*0.09*)	0.52(*0.18*)	0.89(*0.07*)
	0.53	**0.90**	**0.51**	**0.90**
$\hat{\eta}_1, \hat{\eta}_2$	0.15(*0.13*)	0.76(*0.15*)	0.15(*0.14*)	0.77(*0.13*)
	0.10	**0.80**	**0.10**	**0.80**
$Bias(\theta)^2$	1.39(*1.65*)		0.71(*0.57*)	
	0.89		**0.53**	
$\%ERR$	14.6(*18.4*)		12.1(*14.8*)	
	6.9		**6.7**	

Table 2.13: Simulation results for $\eta_1 = 0.1$ and 100 time-series of length T. Mean, standard error (italics) and median (bold) of the estimates.

According to these first results on synthetic data, most of the parameters (Π, the η's and the γ's) are generally correctly estimated, even for short time series. However, the quality of the estimated transition matrix diminishes when η_1 has larger values (the ratio of zeros and ones is overriding the ratio of values in $]0, 1[$). The α's and the β's are correctly estimated for small values of η_1 and sufficiently large time series, with a length at least equal

	$T = 500$		$T = 1000$	
$\hat{\Pi}$	0.78(*0.19*)	0.22(*0.19*)	0.84(*0.15*)	0.16(*0.15*)
	0.86	**0.14**	**0.90**	**0.10**
	0.24(*0.24*)	0.76(*0.24*)	0.16(*0.15*)	0.84(*0.15*)
	0.11	**0.89**	**0.11**	**0.89**
$\hat{\alpha}_1, \hat{\alpha}_2$	1.06(*0.78*)	1.12(*3.49*)	1.01(*0.34*)	0.55(*0.14*)
	0.97	**0.56**	**1.00**	**0.54**
$\hat{\beta}_1, \hat{\beta}_2$	1.01(*0.18*)	2.91(*3.30*)	0.99(*0.11*)	2.08(*0.81*)
	0.99	**1.88**	**0.98**	**1.94**
$\hat{\gamma}_1, \hat{\gamma}_2$	0.51(*0.21*)	0.86(*0.15*)	0.50(*0.15*)	0.89(*0.06*)
	0.48	**0.89**	**0.49**	**0.90**
$\hat{\eta}_1, \hat{\eta}_2$	0.34(*0.14*)	0.75(*0.16*)	0.31(*0.12*)	0.78(*0.11*)
	0.29	**0.80**	**0.30**	**0.79**
$Bias(\theta)^2$	2.23(*4.51*)		0.76(*0.62*)	
	0.94		**0.56**	
$\%ERR$	21.2(*16.5*)		15.1(*10.6*)	
	12.2		**11.6**	

Table 2.14: Simulation results for $\eta_1 = 0.3$ and 100 time-series of length T. Mean, standard error (italics) and median (bold) of the estimates.

	$T = 500$		$T = 1000$	
$\hat{\Pi}$	0.69(*0.23*)	0.31(*0.23*)	0.77(*0.19*)	0.23(*0.19*)
	0.76	**0.24**	**0.88**	**0.12**
	0.24(*0.19*)	0.76(*0.19*)	0.24(*0.22*)	0.76(*0.22*)
	0.18	**0.82**	**0.13**	**0.87**
$\hat{\alpha}_1, \hat{\alpha}_2$	13.08(*80.91*)	4.01(*27.24*)	0.94(*0.19*)	0.58(*0.30*)
	0.93	**0.60**	**0.97**	**0.54**
$\hat{\beta}_1, \hat{\beta}_2$	4.30(*20.91*)	7.58(*45.68*)	0.98(*0.12*)	2.27(*2.59*)
	1.00	**1.74**	**0.98**	**1.71**
$\hat{\gamma}_1, \hat{\gamma}_2$	0.50(*0.24*)	0.81(*0.20*)	0.54(*0.21*)	0.83(*0.16*)
	0.48	**0.88**	**0.51**	**0.89**
$\hat{\eta}_1, \hat{\eta}_2$	0.51(*0.22*)	0.75(*0.18*)	0.50(*0.13*)	0.76(*0.15*)
	0.49	**0.79**	**0.50**	**0.79**
$Bias(\theta)^2$	20.31(*98.05*)		1.25(*2.39*)	
	1.34		**0.79**	
$\%ERR$	30.6(*18.7*)		26.4(*16.2*)	
	23.2		**18.2**	

Table 2.15: Simulation results for $\eta_1 = 0.5$ and 100 time-series of length T. Mean, standard error (italics) and median (bold) of the estimates.

to 1 000. However, the algorithm fails in fairly approaching them when η_1 is greater than 0.5 or for time series shorter than 1 000 observations. Furthermore, when comparing the mean values of the estimates with their medians, one may easily see that, if considering the medians, the performances of the algorithm are eventually not bad even in this limit cases. When looking into details, some of the incoherences come from atypical time series in the simulations which potentially rise identifiability issues.

2.6 Segmenting a century of history

2.6.1 Segmenting the series of monthly counts related to military logistics legislation

I will start by presenting the results on the integer-valued series, consisting of the monthly counts related to the military logistics legislation. A more detailed historical analysis may be found in [MO5] and [MO20].

Let me recall here that the data is a monthly time-series recorder between 1559 and 1661. The architectures of the models to be estimated (number of states for the hidden Markov chains, number of lags in the autoregressive

	$T = 500$		$T = 1000$	
$\hat{\Pi}$	0.59(*0.25*)	0.41(*0.25*)	0.61(*0.24*)	0.39(*0.24*)
	0.60	**0.40**	**0.61**	**0.39**
	0.33(*0.21*)	0.67(*0.21*)	0.30(*0.19*)	0.70(*0.19*)
	0.26	**0.74**	**0.27**	**0.73**
$\hat{\alpha}_1, \hat{\alpha}_2$	4.15(*31.39*)	10.58(*53.98*)	0.89(*0.37*)	5.96(*51.40*)
	0.81	**0.56**	**0.86**	**0.54**
$\hat{\beta}_1, \hat{\beta}_2$	2.15(*11.52*)	80.26(*664.83*)	0.97(*0.20*)	5.77(*36.67*)
	0.99	**1.53**	**0.99**	**1.56**
$\hat{\gamma}_1, \hat{\gamma}_2$	0.51(*0.26*)	0.80(*0.19*)	0.53(*0.24*)	0.81(*0.16*)
	0.53	**0.83**	**0.50**	**0.87**
$\hat{\eta}_1, \hat{\eta}_2$	0.64(*0.23*)	0.77(*0.19*)	0.64(*0.19*)	0.78(*0.17*)
	0.69	**0.80**	**0.68**	**0.80**
$Bias(\theta)^2$	85.93(*666.99*)		7.86(*63.01*)	
	1.46		**1.24**	
$\%ERR$	37.8(*15.2*)		30.0(*15.0*)	
	35.8		**31.0**	

Table 2.16: Simulation results for $\eta_1 = 0.7$ and 100 time-series of length T. Mean, standard error (italics) and median (bold) of the estimates.

parts) were determined based on the historian expertise. For instance, the number of states for the hidden Markov chain was a priori chosen to be equal to two. Indeed, the data led us to expect the existence of two regimes, one of them corresponding to an "intense" legislative activity and the other to a "normal" one. These regimes were to be confronted against the diplomatic situation of the state (war *vs.* peace). For the ZIP-HMM model, an architecture with three hidden states was also trained, but the results were not convincing, neither from a model-selection criterion perspective, nor from the estimated values of the parameters and the corresponding partitioning of the data. For example, the BIC criterion was minimized by the two-state estimated model (2066.72 against 2191.91 for the three-state model).

Results with a ZIP-HMM

For the two-states ZIP-HMM, the estimated parameters are

$$\hat{\pi} = \begin{pmatrix} 0.98 & 0.02 \\ 0.04 & 0.96 \end{pmatrix}, \ \hat{\lambda} = \begin{pmatrix} 0.30 \\ 0.80 \end{pmatrix}, \ \hat{\eta} = \begin{pmatrix} 0.26 \\ 0.06 \end{pmatrix}.$$

The transition matrix shows very stable states. Also, the estimated parameters corresponding the first regime suggest a milder activity in producing law related to military logistics when the state enters this regime. The a-posteriori conditional probabilities of the Markov chain being in the second regime were computed and plotted in the second graph of Figure 2.1. Moreover, the values of these probabilities are thresholded at 0.5 (blue dotted line).

A close study of Figure 2.1, combined with the timeline matching the a posteriori probabilities of the hidden Markov regimes with the periods of war and peace in Figure 2.3, allows a detailed analysis of the activity of the state. The Duchy of Savoy experienced a long period of peace between 1560 and 1588, followed by the War of Provence (1588-1601). During this period of peace and even during the most part of the War of Provence, the Sabaudian state appears not to need to legislate more than usually on matters of feeding or lodging the military. For a switch to the second regime, one has to wait until June 1597, when an important series of documents redefining the military logistics of the Duchy starts being issued. The switch is the result of ten years of war experience, of financial and technical difficulties in supplying the troops. This series ends is January 1599, but another one of the same type starts in July 1600 and ends in March 1602, several months after the signing of the Treaty of Lyon in 1601. This latter switch to the second regime may be the response to two purposes: on the one hand, draw the lessons from the War of Provence and, on the other hand, secure a functional system of stopovers to the numerous Spanish troops present on the lands of the Duchy.
The period from March to October 1607 belongs to the second regime also: the state has healed the wounds of the war and is now completely reforming its logistical administration. The Duchy takes advantage of the situation of peace for deciding purposeful actions, being now free of the constraints of military operations and extreme tax tensions. When Charles Emmanuel Ist signs the Treaty of Bruzolo with France, he is actually preparing to go to war, hence the production of law on military logistics between April and December 1610

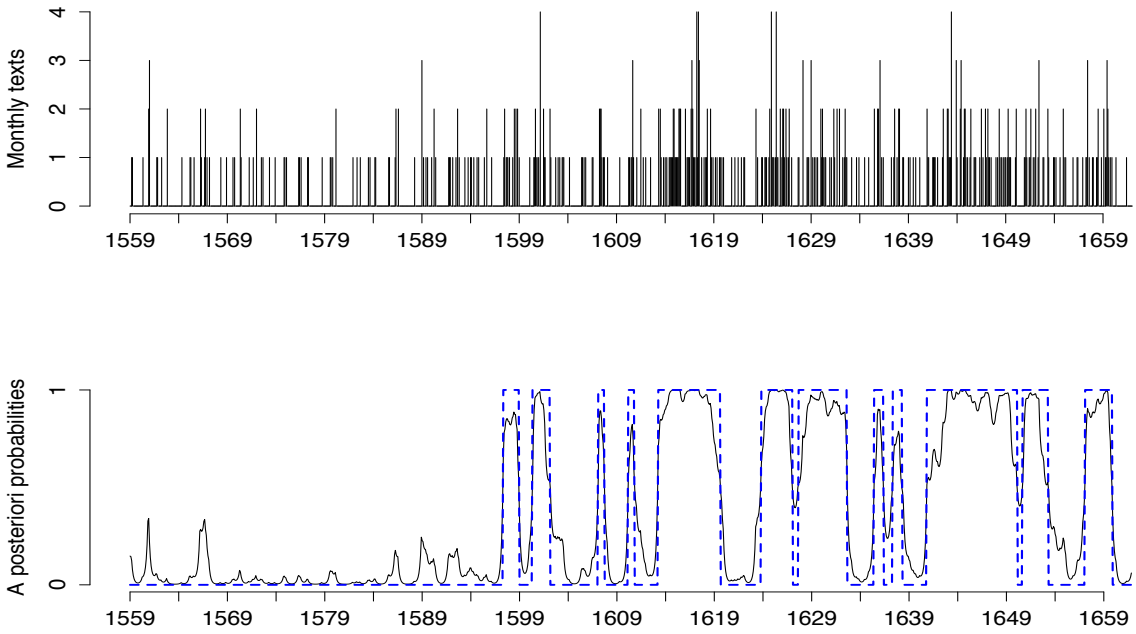


Figure 2.1: Initial time series and a-posteriori probabilities for the second regime of the ZIP-HMM model

follows the second regime. The Duke is thus anticipating the war, which eventually does not start, following the assassination of Henry IV and the changes subsequently intervened in the French politics.
Concerning the rest of the time series, the second regime mostly corresponds to the wars led or sustained by the Duchy, until the Treaty of the Pyrenees. The first War of Montferrat starts with the unexpected invasion of this marquisate on April 22nd, 1613, and ends with the signing of the Treaty of Pavia on October 9th, 1617. The second regime, however, remains the current one until September 1619. During the War of Montferrat, the Sabaudian state doubled its military staff. After the end of the conflict, Charles Emmanuel did not dismiss his troops. On the contrary, he maintained an army two and half times larger than at the beginning of the century. The endeavor to lodge and supply these soldiers continued at the same pace as during the war. Subsequently, the size of the troops decreases gradually until 1623.
Although the war of Valtellina begins in March 1625, the Constable de Lesdiguières and the Duke of Savoy had planned it since 1620. The offensive alliance between Venice, France and the Duchy of Savoy is made official in February 1623 and confirmed in Susa in October 1624. This prewar period is used by Charles Emmanuel for preparing his army and for increasing the size of it. The time series belongs to the second regime from December 1623 until January 1627, time during which the size of the troops goes from 4 500 to 26 600 men. In 1627, the size of the army drops to 5 500 men.
The next switch to the second (January 1628 - September 1632) occurs in the context of the second War of Montferrat and lasts until after the French invasion of the Duchy in 1630 and the Casale assault. Starting with July 1635, the Duchy of Savoy is being constantly at war until 1660, but the time series of documents on military logistics is alternating between the two regimes, the second regime being however the most persistent. Two possible explanations may be hypothesized for this situation. First, the quality of the source may be corrupting the data. For example, the second regime is uninterrupted from September 1640 to February 1650, but afterwards eight consecutive months without any document issuance create a break. Most probably, the state did produce directives for lodging and supplying the troops and for collecting taxes, but the documentation was lost. Second, the long period of war experienced by the Duchy led to various changes and innovations in administrative and tax matters. These changes modified the rhythm of issuing legislation, either by slowing it down or by accelerating it.
With all the above considerations, a first conclusion of this study is that issuing legislation on military logistics is not exactly synchronous with being at war. If periods of conflict logically lead to an increase of the legislative production related to the establishment of rules for lodging and supplying the troops, the temporality of the state may be quite different. The state may respond immediately, in order to face the event, but also it may anticipate future needs or draw the lessons from previous conflicts and act during peacetime.

Results with an INAR(6)-HMM

The series was next segmented using an INAR(6)-HMM model, with two hidden states. The time lag for the autoregressive part was selected according to the historian expertise on the data (the troops were moved according to summer and winter quarters, which generally lasted six months) and according to the information given by the partial-autocorrelation function. The estimated parameters of the model are given below:

$$\hat{\pi} = \begin{pmatrix} 0.999 & 0.001 \\ 0.001 & 0.999 \end{pmatrix} \quad \hat{\lambda} = \begin{pmatrix} 0.306 \\ 0.205 \end{pmatrix}$$

$$\hat{\alpha} = \begin{pmatrix} 0.114 & 0.062 & 0.001 & 0.046 & 0.052 & 0.117 \\ 0.041 & 0.004 & 0.000 & 0.000 & 0.000 & 0.000 \end{pmatrix}.$$

The transition matrix shows very stable regimes, and the expected values of the Poisson distributions are relatively close and quite small. The main difference between the two regimes arises in the values of the binomial coefficients in the autoregressive expressions. According to these results, the first state is strongly dependent on the first and on the sixth lag, while the second state depends at most on the first lag. Hence, the first regime is characteristic to a semestrial regularity of the state in producing legislation on military logistics. The second regime describes a more limited activity of the state in issuing legislation. In Figure 2.2, the time series as well as the a-posteriori probabilities of the first regime are plotted. The values of the probabilities were also thresholded at 0.5 (blue dotted line). According to these plots, there is no alternation between the two regimes, the time series only switches from regime B to regime A once and the transition between the two takes almost ten years. The a-posteriori probability of regime A becomes greater than that of regime B in March 1595, and greater than 0.95 in September 1596.

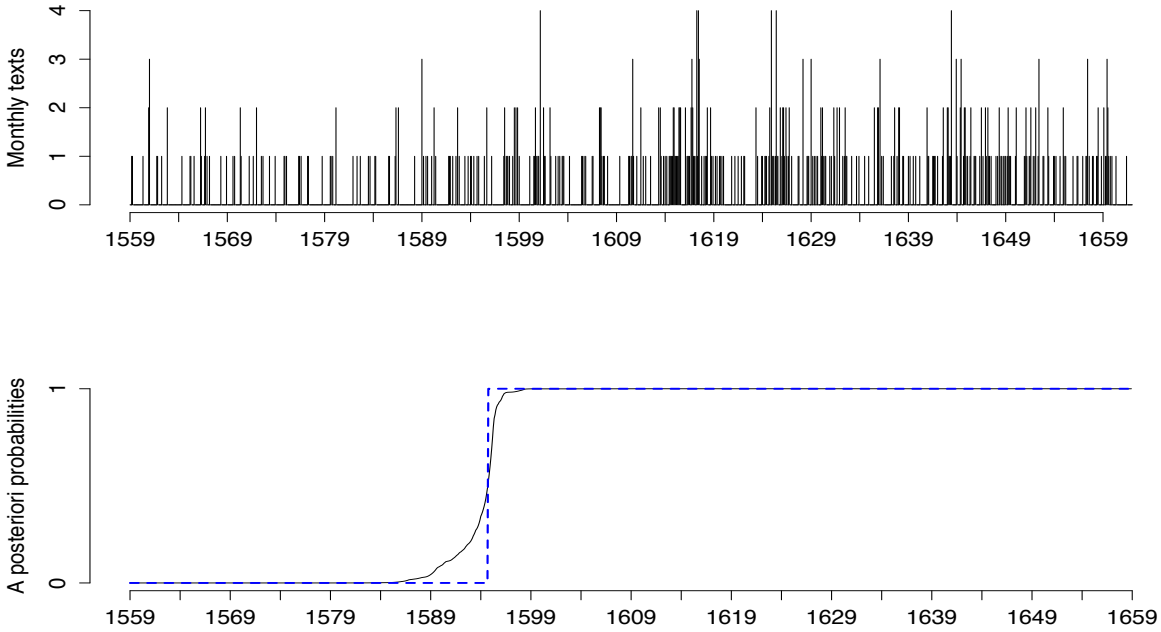


Figure 2.2: Initial time series and a-posteriori probabilities for the first regime of the INAR(6)-HMM model

The two regimes issued from estimating an INAR(6)-HMM model offer a new insight and a new reading for the temporality of the Duchy. The temporal dependency introduced in the autoregressive part of the model leads to a loss of sensitivity in capturing the politico-military situation. But, at the same time, this model brings out the transition from a system yet medieval in logistics administration to an intense and organized state activity during the Iron Century. The interesting point here is that this transition appears not to be linked to the Thirty Years' War, traditionally acknowledged as the key moment in the transformation of the Duchy, but rather to the end of the war against France. In October 1589, the a-posteriori conditional probability of regime B drops below 0.95. This is followed by a long transition between two normative production systems (from October 1589 to August 1596) and which largely corresponds to the War of Provence. During this period of transition, one may witness an increase in the legislative activity of the Duchy. These results converge with the analysis

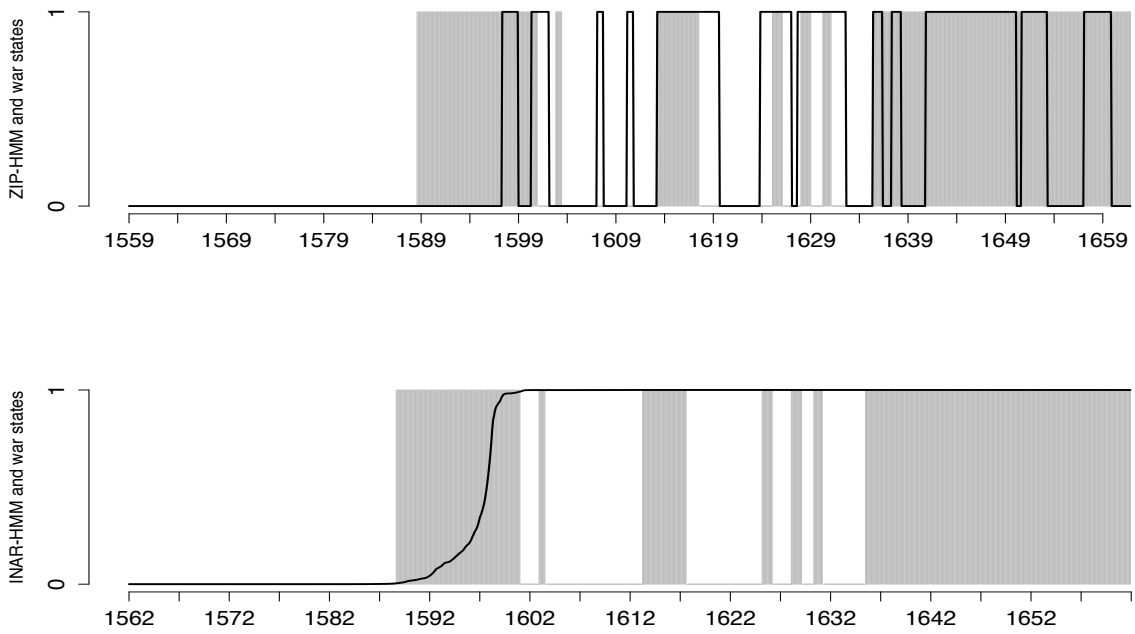


Figure 2.3: Hidden-Markov models segmentations crossed with war (grey) vs. peace (white) periods

of C. Rosso [28] on the development of the Sabaudian bureaucracy starting with the end of 1590, and illustrate to what extent the production of law on military logistics follows a similar temporality with the rest of the institutions of the state and with the rest of legislation issuance.

Eventually, the two models prove to be complementary: the ZIP-HMM model appears as more suitable for highlighting the politico-military situation, with shorter visits in each regime and sudden switches, while the INAR(6)-HMM seems more suitable for pointing out a long ten-year transition between two epochs in the existence of the Duchy of Savoy.

2.6.2 Segmenting the series of ratios

I shall focus here on a different aspect related to the data, which is the relative importance of military logistics in the legislative activity. With this in mind, it is the series of ratios that will be segmented, using Hidden-markov models with Beta-inflated distributions.

A ZOIB-HMM model for the monthly series

The first model, trained on the monthly series of ratios, is a $\mathcal{ZOIB}$-HMM with two states for the hidden Markov chain. The estimated values of the parameters are

$$\hat{\Pi} = \begin{pmatrix} 0.83 & 0.17 \\ 0.26 & 0.74 \end{pmatrix}, \ (\hat{\alpha}_1, \hat{\alpha}_2) = (5.92; 4.11), \ (\hat{\beta}_1, \hat{\beta}_2) = (8.07; 15.83),$$
$$(\hat{\gamma}_1, \hat{\gamma}_2) = (0.01; 0.02), \ (\hat{\eta}_1, \hat{\eta}_2) = (0.87, 0.45). \tag{2.41}$$

The Beta components in the inflated distributions are illustrated in Figure 2.4. According to these and to the rest of the estimated parameters above, the two regimes have quite different behaviors. The first regime contains more than 85% of null values, but rather larger values, associated with a larger variance, in the Beta component. At the same time, less than 45% of the observations in the second regime are zero, and its Beta component is shifted towards the left with respect to the first one. This may signify that the first regime corresponds both to a reduced legislative output and/or to a small proportion of military logistics texts among the whole production, associated with spikes of activity. As for the second regime, it captures mainly the instants where

the importance of military logistics is steadier and more significant. Eventually, the estimated probabilities in the transition matrix hint at a behavior not too persistent in any of the two regimes.

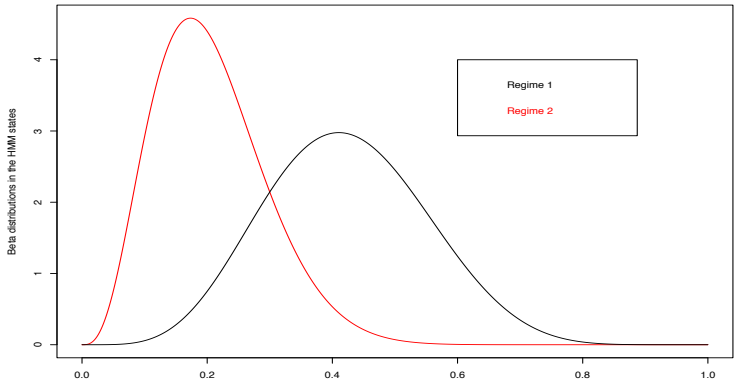


Figure 2.4: Beta components in the emission distributions for the zero-and-one Beta-inflated HMM model (monthly data)

Using the Viterbi algorithm [10], one may track the a posteriori probabilities of the two hidden states and illustrate the estimated trajectory of the hidden Markov chain, as shown in Figure 2.5. The black curve corresponds to the thresholded probabilities of the second regime, conditionally to the data and the estimated parameters. As one may notice and as it was hinted by the estimated transition matrix, the regimes are not stable and keep switching from one to the other. The second regime appears however as more present during the second half of the series. A thorough study should be carried out here by taking advantage of the historian expertise, in order to seize and give meaning to the switches. Although one might think of bridging these results with the "event" temporality highlighted in [MO5], this is not immediate and requires further investigation.

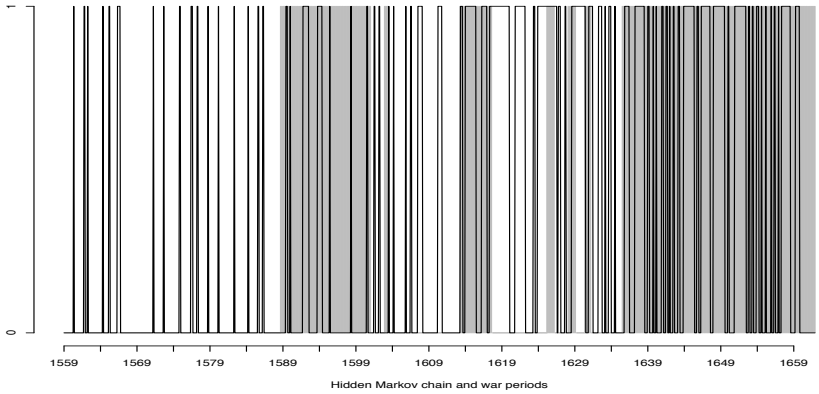


Figure 2.5: A posteriori estimated probabilities of the second regime (zero-and-one Beta-inflated HMM, monthly data). The periods of war for the Duchy are in grey.

A zero-inflated Beta HMM for the quarterly series

Here, instead of modeling the monthly series of ratios, we aggregated it into quarterly observations. For this new series, which contained an important mass of zeros, and no values equal to 1, a two-state hidden-Markov model having zero-inflated Beta distributions as emissions probabilities was trained. The estimated values of the parameters were the following:

$$\hat{\Pi} = \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}, \ (\hat{\alpha}_1, \hat{\alpha}_2) = (2.04; 2.47), \ (\hat{\beta}_1, \hat{\beta}_2) = (8.83; 16.72),$$

and $(\hat{\eta}_1, \hat{\eta}_2) = (0.54, 0.26)$.

The corresponding estimated Beta components in the mixture distributions are illustrated in Figure 2.6. The first regimes contains more zero's than the second one (54% versus 26%), while the Beta components are close in terms of modes; the variance of the first Beta component is larger, favoring larger values on the $[0, 1]$ interval. Hence, the first regimes contains a high proportion of zeros, but also a higher probability mass on the large values, favoring spikes, while the second regime appears as steadier, with less zeros and more concentrated around smaller values. This appears as more consistent with the empirical observations on the time series of ratios. Furthermore, the transition matrix indicates that the two regimes are very persistent.

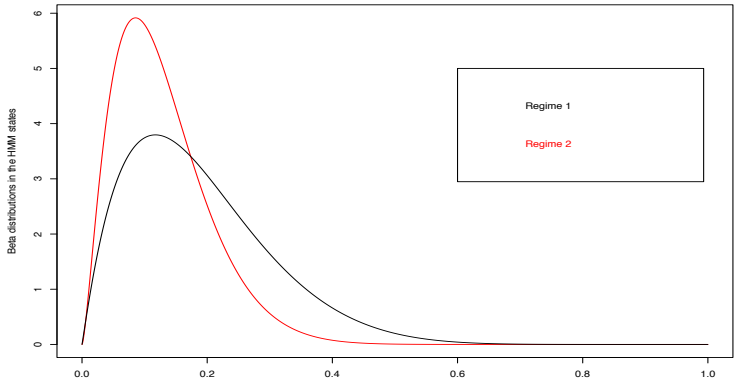


Figure 2.6: Beta components in the emission distributions for the zero-inflated Beta HMM model (quarterly data)

In Figure 2.7, the a posteriori probabilities of the time series being in the first regime of the Markov chain, conditionally to the observed data and the estimated parameters, are illustrated. No thresholding is applied here, the black curve corresponds to the exact probabilities, as computed with the Viterbi filter. The time series switches from one state to the other once only, and this transition lasts about ten years. As anticipated and according to the estimated parameters, the first regime contains more zeros and more spikes, while the second is steadier. Hence, the State appears to have changed its behavior with respect to issuing legislation and also with respect to the importance of military logistics.

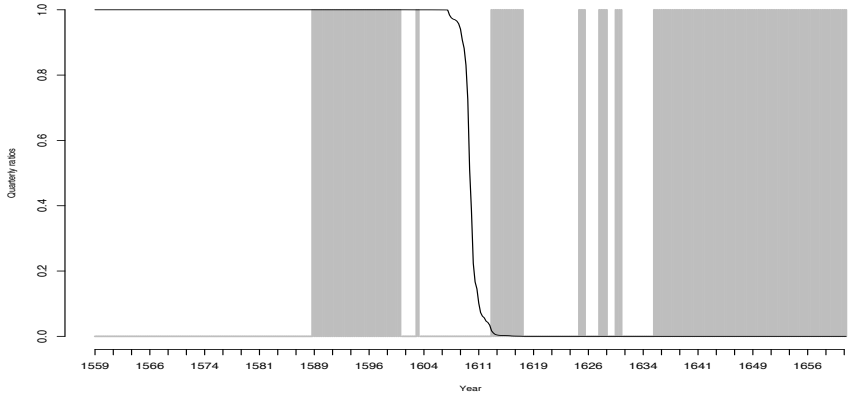


Figure 2.7: A posteriori estimated probabilities of the first regime (zero Beta-inflated HMM, quarterly data). The periods of war for the Duchy are in grey.

The quarterly data segmentation reveals the evolution of the relationship of the State with the situation of war. During the sixteenth century, the legislation on military logistics was issued in an ad hoc manner. Some intermittent edicts and ordinances fixed the main principles of conduct for military logistics: feeding the troops, lodging them, establishing military discipline. The war against France in 1588-1601 changed things. With the military defeat and the financial hardship, the Duke of Savoy became aware of the weakness of the military organization of the Duchy, and of the need to reform it. The main reforms were published between 1602 and 1607. They created a real administration of the military logistics, with specialized officers. This was a period of peace, hence the ratio of texts on military logistics remained low, despite the creation of this administration.

The Markov model detects the transition during the spring of 1610, which remained a period of peace. But this was also the instant where the Treaty of Brussol was signed, on April 25th, when France and Savoy agreed to attack Spain and remove it from Milan and Montferrat. The Duchy prepared for war, he needed to feed and lodge his own 14.000 soldiers, as well as 23.000 French soldiers. He endeavored greatly into military logistics. But, Henri IVth's murder on May 10th changed everything. The regent Marie de Medici refused to go to war against Spain. The Duke had to stand alone against the 30.000 soldiers of Count Fuentes, Governor of Milan. On July 22nd, the Count died and the imminence of a Spanish attack was postponed. During summer and fall, the Duchy of Savoy and Spain negotiated, and the troops were kept mobilized meanwhile. In December, the Duke Charles Emmanuel gave up attacking Spain. However, the risk of war remained high during the next years. The Duke fancied obtaining the Duchy of Montferrat. The ratio of military logistics increased during the preparation for war. The Succession War of Montferrat (1613-1617) opened the "Iron Century". Furthermore, the Thirty Years' War (1618-1648) between France (the Duchy of Savoy is an ally) and Spain tagged the whole time period after 1610 as a "war epoch", despite some peace moments. Edicts and ordinances related to military logistics were being published regularly. The fiscalization of the military logistics increased its importance in the legislative output. The weight of the taxes intended for feeding and lodging the troops reached 50% of the global amount of taxes. The managements of this tax system resulted in a specific and coercitive legislation, at least until the Treaty of the Pyrenees (1659), which also corresponds to the last observations of the available time series.

To summarize, the hidden Markov model on the quarterly data reveals a deep transformation in the State's activity, forced by the necessities of conflict periods. The two regimes detected by the model are not synchronized with periods of peace or war. When the second regimes arrives, the State behaves differently at war. Notwithstanding the peace, the State keeps a high military potential in order to anticipate future conflicts. According to [29] and [30], the State considers itself in state of permanent war after 1610.

2.6.3 Discussion

Why so different results on monthly and quarterly data for the ratios?

The first element for discussion is to understand why the results on monthly data and quarterly data appear to be very different, with a very unequal quality? The hidden Markov model appears as better suited for the quarterly data, and this partly comes from the sensitivity of the quarterly series to two phenomena. The first one is the Chancellery's rhythm of work. If it has a periodic activity during the first part of the observation period, this characteristic fades away as time goes by. If the monthly series quantifies the instants of the issue of legislative texts, the quarterly series would be more related to the continuity of the legislative production. The second phenomenon would be the periodicity of the war. In the Ancien Ŕegime, one would not go to war throughout the year. During the winter quarters, troops rest. In December, the State imposes the local communities to accommodate the soldiers. Other instructions are given to end the winter quarters and to prepare the cavalry's horses by spring. The summer campaigns usually begin with the end of spring. This is an important moment for the State, since it has to organize the return of the troops and solve the issue of tax arrears.

According to the previous, one might conclude that the difference comes from the fact that using different scales changes the level of perception. It is then necessary to reconsider the meaning of the observed human activities. Hence, the scale is not only about the granularity of the data, but also implies a cognitive dimension. Beyond the scale and perception issues, the question of the relevance of the model is also to consider. If the hidden Markov model with Beta-inflated distributions appears as very well fitted for the quarterly data, its performances are more controversial on the monthly series. Identifiability issues may also have an influence here.

Bridging the results on the series of counts and on the series of ratios

Eventually, let us compare the segmentation obtained with the INAR(6)-HMM on the monthly series of counts related to military logistics with that on quarterly ratios. The transitions issued from the two models are illustrated in Figure 2.8. Although the two segmentations bring out similar behaviors in the temporality of the Duchy, there is also a significant time shift in the occurence of the transitions.

The transition computed on the military logistics time series only may be explained as follows. The armies of the Duchy were defeated in 1596-1597. The outcomes of it were important military and taxation reforms, and also an increase of the Chancellery activity. This moment was a crucial one for the State centralization [28]. This transition leads to a steady increase in the monthly production of texts, aimed at ensuring the operational preservation of the army.

The study of the bivariate time series and more particularly of the ratios provides another perspective on the transition. Between 1596 and 1612, the number of texts related to military logistics increased, but the global legislative output increased also. In this case, a different aspect is brought to light, that of the role of military logistics in the construction of the State. The two analyses are not contradictory, they highlight different instants and different phenomena.

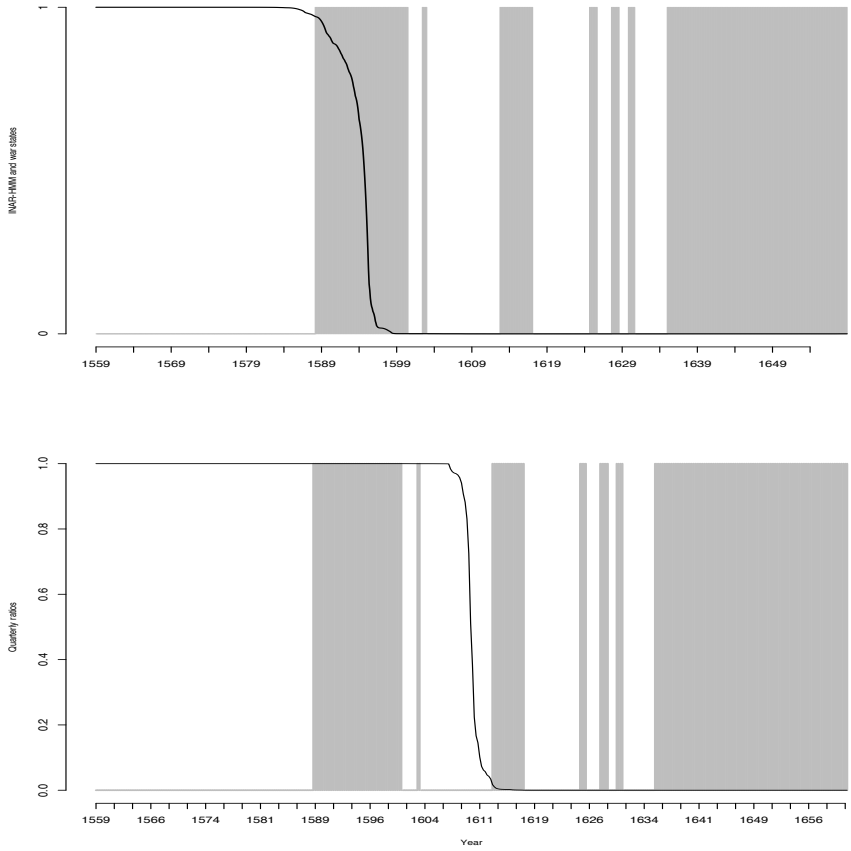


Figure 2.8: Long term transitions captured by the hidden-Markov models. Above: a posteriori probabilities computed with an INAR-HMM model on the military logistics data at a monthly scale. Below: a posteriori probabilities computed with a ZIB-HMM model on the ratios data at a quarterly scale.

2.7 Conclusion, some ongoing work, and some perspectives

This chapter was thought as an illustration of the possible research questions and results stemming from an interdisciplinary collaboration with humanities.

With respect to the historical matter and the data presented above, we are currently implementing a hidden Markov model with a bivariate observed series: a Poisson or a zero-inflated Poisson distribution - for the series of the entire legislation -, coupled with a Binomial conditionally to the Poisson - for the series related to the military logistics. We hope this approach will lead to more consistent and interpretable results on the monthly series, as compared to the model with the Beta distributions. We also consider to train the proposed models on an extended series, especially the INAR(p)-HMM which only switches once on the current data, some additional funding allowing us to hire an intern student for the archival work next summer.

On the theoretical side, most of the work remains to be done. If the identifiability conditions for the ZIP-HMM model are now guaranteed, this is not the case for the Beta-inflated HMM models, and the situation for the INAR(p)-HMM and for the new models we are currently developing needs further investigation. We should also establish formally the properties of the estimates and the model selection issue, when the number of states of the Markov chain is unknown. We are also considering replacing the EM procedure in the INAR(p)-model which is time-consuming by an MCMC-inspired one.

Eventually, and ideally, all these models for integer-valued time series should become available for the community,

in the form of an R-package. This package should also include segmentation procedures based on change-point detection techniques. As I mentioned in the very beginning of this chapter, this kind of data, tagged events with a temporal stamp, could be available not only as aggregated series of counts, but also as series of dates, with the associated tags. One could then be interested in the burstiness of the series, and use models inspired by [31]. As far as I see things, a complete tool designed for the quantitative historians interested in time partitioning should comprise all these algorithms, if possible in a quite user-friendly environnement.

Related to the above, I recently became involved in a new research project with the historian researchers in PIREH, which consists in studying the digital production of knowledge and the rhythms associated to it, by investigating the content of several Wikipedia pages of famous researchers and historical figures. Temporal information on the content of a page, since its creation and with a high temporal resolution, may be queried form Wikipedia: the size of the page, the number of words, and, with more text-mining effort, the table of contents, ... Our idea is to apply time-segmentation techniques (change-point detection in this case) in order to explore the evolution of the pages, identify key-events if any, and compare the evolution of a page in various languages, which constitues an important bias in digital access to information and culture.

Although we received some institutional and financial support from the University scientific board for the work presented here, which allowed us to train several interns and organize several seminars, I realize that this is not sufficient for making things move forward sufficiently fast. My colleagues and I are currently putting together a broader project on the issue of modeling temporality rhythms, that we will submit to forthcoming national and European calls.

I will end this section and this chapter by mentioning that one of the outcomes of this collaboration was creating a Master's degree in digital and quantitative history, which will start in September this year, and which lectures are given both by historians, and mathematicians from SAMM. The biggest challenge associated to the creation of this new degree is most probably to train future PhD candidates in digital humanities, since I am more and more persuaded that a good combination for doing interdisciplinary research is to set up joint PhD projects, involving statistics and humanities candidates.

Chapter 3

The model selection issue

"*If it wasn't for the coffee, I'd have no identifiable personality whatsoever.*"
David Letterman

3.1 Introduction

In the very beginning of the manuscript, I said that I would not organize it chronologically, but rather focus on giving meaning to my strolls among and around time issues in the statistical analysis of data, and particularly data related to humanities and social sciences. The next step brings me to a more theoretical part of my work.

The results I will present next were developed partly during my PhD, and partly in the next years afterwards, in collaboration with Joseph Rynkiewicz, who had co-supervised my PhD. I chose to present this work after the more applied study on Savoy history, because the questions raised in this chapter are more theoretical (and also more philosophical in some sense), while at the same time a natural consequence of the previous modeling approach. Once one gets a *good* model (*good* should be read here as providing meaningful results on the data, from the practitioner's point of view), she may wonder at the reliability of it. And hence get into the more statistical aspects of the question, one of them being of course whether the *true* model having generated the data has been tracked.

The following results were initially motivated by a practical application, where we aimed at modeling some financial time series with a complex autoregressive model, mixing multilayer perceptrons (MLP) – that will be seen here as some parametric nonlinear regression functions –, and hidden Markov chains (HMC). This complex model, that had been introduced in [32], had been proven to be very powerful for modeling nonlinear time series.

Indeed, although linear models have been the standard tool for time series analysis for a long time, and although this is still the case nowadays in some fields of applications in humanities and social sciences, their limitations have been stressed by the statistical community during the past thirty years. Real data often exhibits characteristics that are not taken into account by linear models such as nonlinearities, changes of behavior, ... Financial series, for instance, alternate strong and weak volatility periods, while economic series are often related to the business cycle and switch from recession to growth periods. A large palette of models, including heteroscedastic ARCH or GARCH models [33], [34], multilayer perceptrons [35], autoregressive switching Markov models [6], change-point models [36], [37], etc were proposed to overcome these issues. These are only a few examples, furthermore in the parametric framework, of a very rich literature on time series analysis, with contributions coming from various research fields such as statistics, machine learning, signal processing, statistical physics, ...

We had started by studying a complex hybrid MLP-HMC model, that we had used for modeling a time series related to the French stock market [MO17]. The training of the model had lead to a very meaningful segmentation of the time series. Let me recall here that in the previous chapter also, I illustrated how hidden Markov models, with some well chosen distributions, or combined with an autoregressive process, provided meaningful segmentations of the series, and new ways of reading and interpreting history. Nevertheless, meaningful segmentations although very useful for practitioners, are often not enough, and one needs to know more about the model having generated the data, its architecture etc. It was then only natural to get interested into model selection issues.

In our work, we considered the general case of models which allow the time series to switch between several regimes, with auto-regressive components in each regime. The switches could be independent, distributed

according to a finite-state homogeneous Markov chain, or based on additional or accumulated knowledge as in the case of *gated experts* or *mixtures of experts*, as introduced in [38] and [39]. The autoregressive components were considered either linear, or nonlinear functions, such as for instance multilayer perceptrons.

When the number of regimes or components is fixed, the statistical inference is relatively straightforward using for instance the EM algorithm (see [6] for hidden Markov models with linear autoregressive regimes), and the asymptotic properties of the parameter estimates have been already established (see [40], [41] and [42] for autoregressive processes with Markov switching). When the number of regimes is unknown and has to be selected, the question is far less obvious and difficult to answer. In this case, because of identifiability issues, the Fisher-information matrix is degenerate, and the likelihood ratio test statistic (LRTS hereafter) is no longer convergent *as usual* towards a χ^2-distribution. Some partial answers had been proposed in [43], [44], and [45], where an asymptotic bound for the distribution of the LRTS was derived based on empirical processes techniques, and [46] where the asymptotic distribution of the LRTS had been obtained, but under some very restrictive conditions. In a Bayesian framework, the consistence of the estimate of the number of regimes had been proven in [47].

In the particular case of mixture models, several methods have been proposed to estimate the number of components: nonparametric techniques as in [48], [49] and [50], moment techniques as in [51] and [52], or penalized maximum-likelihood techniques as in [53], [54] and [55]. Furthermore, [56] proved that in the case of hidden Markov models, the number of regimes could be estimated using a penalized marginal-likelihood estimate.

Our work drew its inspiration from the lastly cited papers above and builds on the idea of penalized likelihood criteria. Our contributions extended the results on mixtures and hidden Markov models to the more general case where the mean of the observed process is replaced by a regression function, linear or nonlinear. We proved the consistency of a penalized likelihood criterion for various models (mixtures of linear autoregressive models, mixtures of multilayer perceptrons, mixtures of experts), under some good regularity conditions, and checked that these conditions were easily fulfilled for particular cases of models highly used in practice. The results I will present below were published in three journal articles [MO10], [MO11], [MO12], and three peer-reviewed proceedings [MO37], [MO40], [MO41].

This chapter will be organized as follows. I will start by formalizing the notions and recalling the definitions of autoregressive models with regime switches, independent or Markovian. Next, I will focus on mixtures of autoregressive models (independent regime switches), and state several results which give the limit distribution of the LRTS and the consistency of a penalized log-likelihood criterion for selecting the number of regimes. Next, the assumptions of the consistency result in some particular cases of models highly used in practice, mixtures of linear autoregressive models and mixtures of multilayer perceptrons, are checked. The case of mixtures of experts, where the probabilities of switching between regimes are dependent on the past or on some additional information, is shown to be also suitable for the theoretical framework we proved. Eventually, the possibility of extending the consistency results to autoregressive models Markovian switches is discussed. As the reader will see, the results in this chapter are build using tools mainly related to empirical processes theory.

Since I want to keep this manuscript light and easy to read, I will deliberately not go into the technical details of the proofs, they are fully available in our published papers. I will instead focus on presenting the models, the results, and especially their practicality: how convenient in practice are the hypothesis of the main results, how easily verifiable, to which extent are the models for which the theoretical results hold suited for practical problems and applications.

3.2 LRTS and penalized likelihood criteria for autoregressive mixture models

3.2.1 The general model and some notations

Having started from a practical application, we were first interested in a quite large class of autoregressive models. We would study a sequence of random vectors $(X_t, Y_t)_t$ defined on a probability space $(\Omega, \mathcal{K}, \mathbb{P})$, such that

$$Y_t = F_{\theta_{X_t}}(Y_{t-1}, ..., Y_{t-l}) + \sigma_{X_t}\varepsilon_t \ , \tag{3.1}$$

where

- $\{F_\theta, \theta \in \Theta \subset \mathbb{R}^{d(l)}\}$ is a parametric family of regression functions, and Θ is a compact set;

- X_t is a sequence of random variables valued in the finite set $\{1, ..., p\}$. In the following, X_t may be either an i.i.d. sequence with probability distribution $\pi := (\pi_1, \cdots, \pi_p)$, or a homogeneous Markov chain, irreducible and aperiodic, with transition matrix $\Pi = (\pi_{ij})_{i,j=1,...,p}$, and stationary probability distribution $\pi := (\pi_1, \cdots, \pi_p)$;
- $\sigma_i > 0$ is the standard deviation of the noise term;
- $(\varepsilon_t)_t$ is a sequence of i.i.d. variables, having a positive density with respect to the Lebesgue measure, and such that ε_t is independent of $(Y_{t-k})_{k\geq 1}$. Usually, one considers ε_t as standard Gaussians.

For fixed values of p, the number of regimes, and l, the number of lags, the parameters of the model in Equation 3.1 that one has to estimate are $(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\Pi})$, where $\boldsymbol{\theta} = (\theta_1, ..., \theta_p) \in \Theta^p$, $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_p) \in \mathbb{R}_+^p$, and $\boldsymbol{\Pi}$ being either the mixture distribution $\pi := (\pi_1, \cdots, \pi_p)$ if X_t is an i.i.d. sequence, or the transition matrix Π if X_t is a Markov chain. If p and l are fixed, and under some usual regularity conditions, one may estimate the parameters and derive the asymptotic properties of the estimates. This is more or less straightforward, depending on the form of F_θ, but as I mentioned in the introduction, there is a thorough and well-documented literature available on the topic.

Our main concern was nonetheless the case where the number of regimes was not fixed, or priorly known, which is the most common situation in practical applications. It is this situation that generates identifiability issues in the estimation procedure. I will focus on this question in the following pages. For simplifying the writing, I will suppose in the following that the number of lags l is fixed and equal to 1 (this won't affect the generality of the following results). Before listing some of the results, I will introduce some further notations and definitions.

Throughout the rest of the chapter, I will suppose that $(Y_t)_t$ is strictly stationary, and geometrically β-mixing. This assumption may appear as a strong one, but we will see that it is actually fulfilled by a wide class of processes. Let me briefly recall that, for all $n \geq 1$, the β-mixing coefficients are defined as

$$\beta_n = \beta\left(\mathcal{F}_{-\infty}^0,\, \mathcal{F}_n^\infty\right) \ ,\ \mathcal{F}_{-\infty}^0 = \sigma\left(Y_k,\, k \leq 0\right) \ ,\ \mathcal{F}_n^\infty = \sigma\left(Y_k,\, k \geq n\right) \ , \tag{3.2}$$

where

$$\beta\left(\mathcal{A}, \mathcal{B}\right) = \frac{1}{2} \sup_{(A_i)_{i\in I}, (B_j)_{j\in J}} \sum_{(i,j)\in I\times J} \left|\mathbb{P}\left(A_i \cap B_j\right) - \mathbb{P}\left(A_i\right)\mathbb{P}\left(B_j\right)\right| \tag{3.3}$$

and $(A_i)_{i\in I}$, respectively $(B_j)_{j\in J}$, range over the set of $\mathcal{A}$, respectively $\mathcal{B}$, measurable partitions. The mixing condition ensures an asymptotic independence at a geometrical rate for the observed process, which will allow us to establish the following results.

With the previous definitions and remarks, we consider a parametric framework and hence a parametric family of densities, $\mathcal{P} = \left\{g_\theta, \theta \in \Theta \in \mathbb{R}^d\right\}$, with respect to some positive measure ν, where Θ is a compact finite-dimensional set. For the moment, I will consider the case of mixtures only, hence X_k is supposed to be i.i.d. In this case, for every Y_k, one will suppose that the true density conditionally to Y_{k-1} and marginally in X_k is

$$g^0\left(Y_k \mid Y_{k-1}\right) = \sum_{i=1}^{p_0} \pi_i^0 g_{\theta_i^0}\left(Y_k \mid Y_{k-1}\right) \ , \tag{3.4}$$

where $g_{\theta_i^0} \in \mathcal{P}$, $\pi_i^0 \geq 0$ and $\sum_{i=1}^{p_0} \pi_i^0 = 1$.

The question to answer next is how to appropriately estimate and select p_0? One common approach would consist into looking at the LRTS and derive its asymptotic distribution. For a sample $\{Y_1, ..., Y_n\}$, the likelihood-ratio test statistic is defined as

$$2\lambda_n = 2\left(\sup_{g\in\mathcal{G}} l_n(g) - l_n(g^0)\right) \ , \tag{3.5}$$

where $l_n\left(g\right) = \sum_{k=2}^n \ln\, g\left(Y_k \mid Y_{k-1}\right)$ is the log-likelihood of $(Y_1, \cdots, Y_n)$, conditionally to Y_1. In this definition,

$$\mathcal{G} = \left\{g = \sum_{i=1}^p \pi_i g_{\theta_i}, \pi_i \in [0,1], \sum_{i=1}^p \pi_i = 1, g_{\theta_i} \in \mathcal{P}, p \in \mathbb{N}^\star\right\} \ , \tag{3.6}$$

where $p > p_0$ (if $p < p_0$ there are no identification issues), is the set of all possible conditional densities.

In the classical statistical theory, the approximation of the log-likelihood and the convergence of the LRTS are established using a second-order Taylor expansion in a neighborhood of the true parameter. For example, if the distribution g depends on a parameter θ, and the true parameter of g^0 is θ^0, then one writes

$$l_n(g) \simeq l_n(g^0) + (\theta - \theta^0)\frac{\partial l_n(g)}{\partial\theta}(\theta^0) + \frac{1}{2}(\theta - \theta^0)^2\frac{\partial^2 l_n(g)}{\partial^2\theta}(\theta^0) \ . \tag{3.7}$$

Then, one may derive also that the LRTS converges to a χ^2 distribution.

This well established theory is no longer valid in our case. Indeed, since our model is not identifiable, the true parameter is not unique, and one cannot use the above expansion. Instead, she may define an *extended set of score functions* which allow one to give an approximation of the LRTS in a neighborhood of the set of true parameters (parameters giving the true log-likelihood function).

In order to get such an approximation, one has to control the size of λ_n, which may be done thanks to the empirical processes theory, and more particularly thanks to the functional versions of the law of large numbers and of the central limit theorem. To get the law of large numbers, the considered set of functions has to be not too big, one says *Glivenko-Cantelly* (see [57] for a proper definition), that is that it may be covered by a finite set of balls. The assumption for the asymptotic normality is more restrictive, since the covering number, which depends of the diameter ε of the balls, has now to be of order $\exp^{\frac{1}{\varepsilon^2}}$, when ε goes to zero. One calls such a set of functions a *Donsker class* (see [57]).

3.2.2 A useful approximation of the LRTS

One of our first results, proven in [MO10] and generalizing a theorem in [56], gives an approximation of the LRTS as a supremum over the *limit set of score functions*. This approximation will be very useful for deriving the asymptotics of the likelihood ratio. The main assumption one needs is that the *extended set of score-functions* that I will introduce next is Donsker.

For any $\eta > 0$, one may define the *extended set of score-functions* $\mathcal{S}_\eta$ as

$$\mathcal{S}_\eta = \left\{ s_g = \frac{\frac{g}{g^0} - 1}{\left\| \frac{g}{g^0} - 1 \right\|_{L^2(\mu)}}, g \in \mathcal{G}_\eta \right\}, \ \mathcal{G}_\eta = \left\{ g \in \mathcal{G}, \|g - g^0\|_{L^2(\mu)} \leq \eta \right\}, \tag{3.8}$$

and the *limit set of scores* $\mathcal{D}$ as

$$\mathcal{D} = \left\{ d \in \mathbb{L}^2(\mu) \mid \exists (g_n) \in \mathcal{G}, \| \frac{g_n - g^0}{g^0} \|_{\mathbb{L}^2(\mu)} \xrightarrow[n\to\infty]{} 0, \|d - s_{g_n}\|_{\mathbb{L}^2(\mu)} \xrightarrow[n\to\infty]{} 0 \right\}. \tag{3.9}$$

By putting $g_t = g_n$ for $t \in [0,1]$ and $n \leq \frac{1}{t} < n+1$, one obtains that, for all $d \in \mathcal{D}$, there exists a parametric path $(g_t)_{0\leq t\leq 1}$ such that $\forall t \in [0,1]$, $g_t \in \mathcal{G}$, $t \to \| \frac{g_t - g^0}{g^0} \|_{\mathbb{L}^2(\mu)}$ is continuous, $\| \frac{g_t - g^0}{g^0} \|_{\mathbb{L}^2(\mu)} \xrightarrow[t\to 0]{} 0$ and $\|d - s_{g_t}\|_{\mathbb{L}^2(\mu)} \xrightarrow[t\to 0]{} 0$.

Next, one needs to define $\mathcal{L}_{2,\beta}(\mathbb{P})$ spaces and the notion of bracketing entropy. If $(Y_k)_{k\in\mathbb{Z}}$ is a strictly stationary sequence, β-mixing and such that $\sum_{n\geq 1} \beta_n < \infty$, then

$$\mathcal{L}_{2,\beta}(\mathbb{P}) = \left\{ f, \ \|f\|_{2,\beta} < \infty \right\}, \quad \|f\|_{2,\beta} = \sqrt{\int_0^1 \beta^{-1}(u) \left[Q_f(u) \right]^2 du}, \tag{3.10}$$

where $\beta(u)$ is the *càdlàg extension* of the mixing coefficients β_n by considering $\beta(u) = \beta_{[u]}$ and $\beta_0 = 1$; $\varphi^{-1}(u) = \inf\{t \in \mathbb{R}, \varphi(t) \leq u\}$, if φ is a non-increasing function; and Q_f is the quantile function of $|f(Y_1)|$, that is the inverse of $t \to \mathbb{P}(|f(Y_1)| > t)$.

If one considers now the extended set of score-functions $\mathcal{S}_\eta$, equipped with the norm $\|\cdot\|_{2,\beta}$, and if she defines an ε-bracket as $[l, u] = \{f \in \mathcal{S}_\eta, l \leq f \leq u\}$ such that $\|u - l\|_{2,\beta} < \varepsilon$, then the ε-bracketing entropy of $\mathcal{S}_\eta$ with respect to the norm $\|\cdot\|_{2,\beta}$ is

$$\mathcal{H}_{[\cdot]}\left(\varepsilon, \mathcal{S}_\eta, \|\cdot\|_{2,\beta}\right) = \ln\left(\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{S}_\eta, \|\cdot\|_{2,\beta}\right)\right), \tag{3.11}$$

where $\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{S}_\eta, \|\cdot\|_{2,\beta}\right)$ is the minimum number of ε-brackets necessary to cover $\mathcal{S}_\eta$.

Assumption (B) Assume that $\mathcal{G}$ is Glivenko-Cantelli and that there exists $\eta > 0$ such that

$$\int_0^1 \sqrt{\mathcal{H}_{[\cdot]}\left(\varepsilon, \mathcal{S}_\eta, \|\cdot\|_{2,\beta}\right)} d\varepsilon < \infty. \tag{3.12}$$

According to [58], assumption **(B)** implies that $\mathcal{S}_\eta$ is Donsker. Then, one may state and prove the following theorem, using similar inequalities as those proven in [56]:

Theorem 3.2.1 *Under the assumption* **(B)** ,

$$2\lambda_n = \sup_{d\in\mathcal{D}} \left(\max\left\{ \frac{1}{\sqrt{n}} \sum_{i=2}^{n} d(Y_i, Y_{i-1}); 0 \right\} \right)^2 + o_P(1) \ . \tag{3.13}$$

3.2.3 The asymptotic distribution of the LRTS

The result above being quite general, we were interested next in giving sufficient conditions for assumption **(B)** to hold. Usually, for parametric models, a Lipschitz condition on θ would be sufficient to show that $\mathcal{S}_\eta$ is Donsker. However, since g depends on the parameter θ, the score function $\theta \mapsto s_g = \frac{\frac{g}{g^0}-1}{\left\|\frac{g}{g^0}-1\right\|_{L^2(\mu)}}$ may not be continuous, thus not Lipschitz, in θ_0.

In [MO10], we gave a set of assumptions – verified by many models widely used in practice as we shall see afterwards –, and derived the asymptotic behavior of the LRTS. The following theorem generalizes a previous result of [55].

Consider the following assumptions:

H-1 The set $\mathcal{G}$ is Glivenko-Cantelli and the set of possible parameters:

$$\{\pi_1, \cdots, \pi_p \in [0,1], \theta_1, \cdots, \theta_p \in \Theta\}$$

contains a neighborhood of the parameters defining the true conditional density g^0.

H-2 There exists $\eta > 0$ such that for all $g \in \mathcal{G}$ with $\|g - g^0\|_{L^2(\mu)} < \eta$, $\left\|\frac{g}{g^0} - 1\right\|_{L^2(\mu)} < \infty$

H-3 By denoting $l_{\theta_i} := \frac{g_{\theta_i}}{g^0}$ and, with a slight abuse of notation, $\frac{\partial^q}{\partial \theta_j^q}$ the derivative of order q with respect to all components of θ_j, we assume the existence of a square-integrable function h and of a neighborhood $\mathcal{V}$ of $(\theta_1^0, \cdots, \theta_{p^0}^0)$ such that, for all $(\theta_1, \cdots, \theta_{p^0}) \in \mathcal{V}$,

$$\left|\frac{\partial l_{\theta_j}}{\partial \theta_j}(\theta_j)\right| \le h, \ \left|\frac{\partial^2 l_{\theta_j}}{\partial \theta_j^2}(\theta_j)\right| \le h \text{ and } \left|\frac{\partial^3 l_{\theta_j}}{\partial \theta_j^3}(\theta_j)\right| \le h.$$

H-4 With the notations

$$l_j' := \frac{\partial l_{\theta_j}}{\partial \theta_j}\left(\theta_j^0\right), \ l'' := \frac{\partial^2 l_{\theta_j}}{\partial \theta_j^2}\left(\theta_j^0\right)$$

we assume that for distinct $(\theta_i)_{1\le i\le p}$, $\left\{(l_{\theta_i})_{1\le i\le p}, (l_i')_{1\le i\le p^0}, (l_i'')_{1\le i\le p^0}\right\}$ are linearly independent in the Hilbert space $L^2(\mu)$.

We also define $\Omega : L^2(P) \to L^2(\mu)$ by $\Omega(g) = \frac{g}{\|g\|_2}$, for $g \neq 0$.

Theorem 3.2.2 *Let d be the parametric dimension of the regression functions. Under the assumptions H-1, H-2, H-3 and H-4, there exists a centered Gaussian process $\{W_S, S \in \mathbb{F}\}$ with continuous sample paths and covariance kernel $P(W_{S_1} W_{S_2}) = P(S_1 S_2)$ such that*

$$\lim_{n\to\infty} 2\lambda_n = \sup_{S\in\mathbb{F}} \left(\max(W_S, 0)\right)^2 .$$

The index set $\mathbb{F}$ is defined as $\mathbb{F} = \cup_t \mathbb{F}_t$, with the union running over $t = (t_0, \cdots, t_{p_0}) \in \mathbb{N}^{p_0+1}$ with $0 = t_0 < t_1 < \cdots < t_{p_0} \le p$ and

$$\begin{array}{l} \mathbb{F}_t = \left\{ \Omega\left(\sum_{i=1}^{p_0} \zeta_i l_{\theta_i^0} + \sum_{i=p_0+1}^{p} \zeta_i l_{\theta_i} + \sum_{i=1}^{p_0} \lambda_i^T l_i' + \delta \sum_{i=1}^{p_0} \sum_{j=t_{i-1}+1}^{t_i} \gamma_j^T l_i'' \gamma_j \right), \right. \\ \left. \lambda_1, \cdots, \lambda_{p_0}, \gamma_1, \cdots, \gamma_{t_{p_0}} \in \mathbb{R}^d \ ; \zeta_1, \cdots, \zeta_p \in \mathbb{R}, \theta_{t_{p_0}+1}, \cdots, \theta_p \in \Theta - \left\{\theta_1^0, \cdots, \theta_{p_0}^0\right\} \right\} \end{array}$$

where $\delta = 1$ if there exists a vector $\mathbf{q}$ such that: $q_j \le 0$, $\sum_{j=t_{i-1}+1}^{t_i} q_j = 1$, $\sum_{j=t_{i-1}+1}^{t_i} \sqrt{q_j}\gamma_j^t = 0$ for $i = 1, \cdots, p_0$; and $\delta = 0$ otherwise.

3.2.4 Penalized-likelihood estimate for the number of regimes

The previous results prove to be quite useful for practical applications. Since λ_n is tight, it is quite straightforward to prove the consistency of penalized likelihood criteria.

For some fixed $P \in \mathbb{N}^\star$ sufficiently large, we shall consider the following class of functions

$$\mathcal{G}_P = \bigcup_{p=1}^{P} \mathcal{G}_p \ , \ \mathcal{G}_p = \left\{ g\left(y_k \mid y_{k-1}\right) = \sum_{i=1}^{p} \pi_i g_{\theta_i}\left(y_k \mid y_{k-1}\right), \pi_i \in [0;1], \sum_{i=1}^{p} \pi_i = 1, g_{\theta_i} \in \mathcal{P} \right\}. \tag{3.14}$$

For every $g \in \mathcal{G}_P$ we define the number of regimes as

$$p\left(g\right) = \min\left\{p \in \{1, ..., P\} \ , \ g \in \mathcal{G}_p\right\}. \tag{3.15}$$

With this definition, $p_0 = p\left(g^0\right)$ is the number of regimes of the true model.

The estimate of the number of regimes $\hat{p}$ can now be defined as $p \in \{1, ..., P\}$ maximizing the penalized criterion:

$$T_n\left(p\right) = \sup_{g \in \mathcal{G}_p} \ l_n\left(g\right) - a_n\left(p\right) \ , \tag{3.16}$$

where $a_n\left(p\right)$ is a penalty term.

We proved the following theorem in [MO10]. It holds under the general assumptions in the previous section, while its proof is inspired by similar results for mixtures of distributions in [56] and [54].

Theorem 3.2.3 *Suppose the following assumptions are true:*

- **H-1**, **H-2**, **H-3** *and* **H-4**;
- **(A)**: $a_n\left(\cdot\right)$ *is an increasing function of* p, $a_n\left(p_1\right) - a_n\left(p_2\right) \xrightarrow[n\to\infty]{} \infty$ *for every* $p_1 > p_2$, *and* $\frac{a_n(p)}{n} \xrightarrow[n\to\infty]{} 0$ *for every* p.

Then, $\hat{p}$ *maximizing the penalized criterion defined by Equation 3.16 converges in probability,* $\hat{p} \xrightarrow[n\to\infty]{\mathbb{P}} p_0$.

3.3 Mixtures of autoregressive linear models

Let me recall at this point that the theoretical results above were mainly motivated by a case study that we had started from. It was thus essential for us to make a step beyond the theoretical framework, and more particularly come back to the data and check if the hypothesis above held for practical classes of models. The first we investigated is that of mixtures of autoregressive linear models.

In this section, I shall consider that the process (X_t, Y_t) follows the true model

$$Y_t = a^0_{X_t} Y_{t-1} + b^0_{X_t} + \sigma^0_{X_t} \varepsilon_t \ , \tag{3.17}$$

where

- X_t is an i.i.d. sequence of random variables valued in a finite space $\{1, ..., p_0\}$ and with probability distribution $\pi^0 := \left(\pi^0_1, \cdots, \pi^0_{p_0}\right)$
- for every $i \in \{1, ..., p_0\}$, a^0_i, b^0_i, σ^0_i are real numbers with $\left|a^0_i\right| < 1$ and $\sigma^0_i > 0$. The true global parameter is then $\left(\pi^0_i, \theta^0_i\right)_{i=1,...,p_0}$, with $\theta^0_i = \left(a^0_i, b^0_i, \sigma^0_i\right)$.
- $(\varepsilon_t)_{t\in\mathbb{N}}$ is a sequence of i.i.d. standard Gaussian variables, independent of $(Y_{t-k})_{k\geq 1}$.

3.3.1 Checking the hypothesis for the consistency of penalized likelihood criteria

First, we needed to check the general assumptions of stationarity and ergodicity that we made at the beginning of this chapter. This was done in the following proposition, proven in [MO10].

Proposition 3.3.1 *The process (X_t, Y_t) defined by (3.17) is strictly stationary, geometrically ergodic and, in particular, geometrically β-mixing. Moreover, there exists $\delta > 0$ such that $E_\mu\left(e^{\delta Y_t^2}\right) < \infty$.*

Next, we got interested whether the hypothesis **(H-1)-(H-4)** above were verified. Since the noise is supposed to be Gaussian, the set of possible conditional densities is the following :

$$\mathcal{G} = \left\{ g\left(y_2 \mid y_1\right) = \sum_{i=1}^{p} \pi_i \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2\sigma_i^2}(y_2-(a_i y_1+b_i))^2}, p = 1, ..., P; \theta_i = (a_i, b_i, \sigma_i) \in \Theta; \pi_i \in [0,1], \sum_{i=1}^{p} \pi_i = 1 \right\} \tag{3.18}$$

with $\Theta \subset \mathbb{R}^2 \times \mathbb{R}_+^\star$ a compact set.

The true density is

$$g^0\left(y_2|y_1\right) = \sum_{i=1}^{p_0} \pi_i^0 \frac{1}{\sqrt{2\pi(\sigma_i^0)^2}} \exp\left(-\frac{1}{2(\sigma_i^0)^2}\left(y_2 - \left(a_i^0 y_1 + b_i^0\right)\right)^2\right) \tag{3.19}$$

Within this framework, $\mathcal{G}$ is Glivenko-Cantelli and the assumption **(H-1)** is true. Moreover, the derivatives up to the third order of

$$l_{\theta_i}(y_1, y_2) = \frac{g_{\theta_i}(y_2|y_1)}{g^0(y_2|y_1)} = \frac{\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2(\sigma_i)^2}\left(y_2 - (a_i y_1 + b_i)\right)^2\right)}{\sum_{j=1}^{p_0} \pi_j^0 \frac{1}{\sqrt{2\pi(\sigma_j^0)^2}} \exp\left(-\frac{1}{2(\sigma_j^0)^2}\left(y_2 - \left(a_j^0 y_1 + b_j^0\right)\right)^2\right)} \tag{3.20}$$

exist and are dominated by a square integrable function, hence the assumption **(H-3)** also holds.

Next, we check whether the generalized score functions are well defined (assumption **H-2**):

$$\left\| \frac{g}{g^0} - 1 \right\|_{L^2(\mu)} < \infty, \forall g \text{ such that } \|g - g^0\| \leq \eta.$$

One can prove by direct computations that:

Proposition 3.3.2 $\left\| \frac{g}{g^0} - 1 \right\|_{L^2(\mu)} < \infty$ *if for every $i \in \{1, ..., p\}$, there exists $k \in \{1, ..., p_0\}$such that $\sigma_i^2 < 2\left(\sigma_k^0\right)^2$ and $\left|a_i - a_k^0\right| < \sqrt{\delta\left(2\left(\sigma_k^0\right)^2 - \sigma_i^2\right)}$ for $\delta > 0$ verifying $E\left(e^{\delta Y_t^2}\right) < \infty$.*

Roughly speaking , the sufficient condition in this Proposition states that assumption **H-2** will be verified if the possible models are not too different from the real one. I will discuss later the consequences of this condition, and assume for the moment it is fulfilled.

Eventually, assumption **(H-4)** is fulfilled thanks to the following result:

Proposition 3.3.3 *The family of functions*

$$\left\{ g_{\theta_i}, i = 1, \cdots, p, \frac{\partial g_{\theta_i^0}}{\partial a_i}, \frac{\partial g_{\theta_i^0}}{\partial b_i}, \frac{1}{\sigma_i^0}\frac{\partial g_{\theta_i^0}}{\partial \sigma_i} + \frac{\partial^2 g_{\theta_i^0}}{\partial b_i^2}, \frac{\partial^2 g_{\theta_i^0}}{\partial a_i^2}, \frac{\partial^2 g_{\theta_i^0}}{\partial \sigma_i^2}, \frac{\partial^2 g_{\theta_i^0}}{\partial a_i \partial \sigma_i}, \frac{\partial^2 g_{\theta_i^0}}{\partial a_i \partial b_i}, \frac{\partial^2 g_{\theta_i^0}}{\partial b_i \partial \sigma_i}, i = 1, ..., p_0 \right\} \tag{3.21}$$

are linearly independent.

According to what has been presented above, the theorems in this chapter hold under relatively mild and usual conditions for mixtures of linear autoregressive models. If each linear component is stationary ($|a_i^0| < 1$), the noise is Gaussian, and the possible models are not too different from the *true* one, then the penalized likelihood criterion is consistent and may be properly used for selecting the *size* of the model. Nevertheless, as the true regression function is not actually known, it appears it is impossible to assume that what is being used as possible model is not too far from the true one! If the parameter set is not restricted, we shall see that the LRTS diverges, and the results above do not longer hold. I will illustrate this in the next section on a very simple example.

3.3.2 An example illustrating the divergence of the LRTS

Consider a very simple scenario, with $g^0\left(y_k \mid y_{k-1}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_k^2}$, $\mathcal{P} = \left\{ g_\theta(y_k|y_{k-1}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_k - \theta y_{k-1})^2}, |\theta| < 1 \right\}$, and the possible set of conditional densities,

$$\mathcal{G} = \left\{ g\left(y_k \mid y_{k-1}\right) = \pi g_\theta\left(y_k \mid y_{k-1}\right) + (1-\pi) g^0\left(y_k \mid y_{k-1}\right), \pi \in [0;1], g_\theta \in \mathcal{P} \right\} . \tag{3.22}$$

One is interested in whether the true model is a mixture model (i.e. $\theta \neq 0$ and $\pi \neq 0$), or whether the observations are independent (i.e. $\theta = 0$ or $\pi = 0$). The LRTS, defined as in Equation 3.5, is

$$2\lambda_n = 2\left(\sup_{g\in\mathcal{G}} \ln(g) - \ln(g^0)\right) = 2\sup_{g\in\mathcal{G}} \sum_{k=2}^{n} \ln \frac{\pi g_\theta\left(y_k \mid y_{k-1}\right) + (1-\pi) g^0\left(y_k \mid y_{k-1}\right)}{g^0\left(y_k \mid y_{k-1}\right)} \ . \tag{3.23}$$

Two cases are analyzed, according to whether π is close to 0, or whether there exists $\delta > 0$ such that $\pi \geq \delta$.

In the first case, one may easily prove that the LRTS may be divergent, since it is possible to have $\mathbb{E}_\mu\left(\ln(g) - \ln(g^0)\right) \to 0$ when $\theta \neq 0$. Indeed, if $-\frac{1}{\sqrt{2}} < \theta < \frac{1}{\sqrt{2}}$, the norm

$$\left\| \frac{\pi g_\theta + (1-\pi) g^0}{g^0} - 1 \right\|_{L_2(\mu)} = \pi \left\| \frac{g_\theta}{g^0} - 1 \right\|_{L_2(\mu)} < +\infty \ , \tag{3.24}$$

and the score functions are well defined. The set of limit score functions will contain the following set of functions:

$$\left\{ s_\theta\left(Y_1, Y_2\right) = \frac{\frac{g_\theta(Y_2|Y_1)}{g^0(Y_2|Y_1)}}{\left\| \frac{g_\theta(Y_2|Y_1)}{g^0(Y_2|Y_1)} \right\|_{L^2(\mu)}} \ , \ \theta \in \left]-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right[\right\} \ . \tag{3.25}$$

We've already seen that for a finite number of possible parameters $\theta_1, ..., \theta_m$, the distribution of the LRTS will converge towards the square of a m-dimensional Gaussian distribution. Then, if an arbitrary number of *almost* uncorrelated random variables can be found, λ_n can take an arbitrarily large value since the maximum of m independent samples from a standard Gaussian is approximately $\sqrt{2\ln m}$. Using Theorem 3.2.2 and a result in [59], one may show that by selecting a sequence of parameters $\theta_m = \frac{1}{\sqrt{2}} - \frac{1}{m}$, the sequence of score functions s_{θ_m} converges towards 0 in probability, and the LRTS is divergent.

If $\pi \geq \delta > 0$, then the maximum likelihood estimate $\hat{\theta}$ converges towards $\theta^0 = 0$, otherwise the limit of the LRTS cannot be close to zero. Thus, the model is identifiable in θ and unidentifiable in π. One may also show that

$$s_g = \frac{\frac{g_\theta}{g^0} - 1}{\| \frac{g_\theta}{g^0} - 1 \|_{L^2}} = \frac{Y_1 Y_2 + o(1)}{\| Y_1 Y_2 + o(1) \|_{L^2}}, \tag{3.26}$$

hence the LRTS converges to the classical χ^2 distribution.

This discussion on a particular case was meant to clarify how to deal in practice with mixture models. If mixture weights can be as small as possible, then the LRTS tends to infinity. In order to avoid this divergence, one should constraint the parameters in a neighborhood of the true value, which is senseless for real data where the true model is unknown. But if the mixture weights are bounded from below, then all parameters of the regression functions converge to some true ones.

3.3.3 An illustration on simulations

We illustrated the theoretical consistency result on several simulated data. The examples considered here are mixtures of two linear autoregressive models and are excerpted from [MO12],

$$g^0\left(y_1, y_2\right) = \pi_1^0 f_1^0\left(y_2 - F_1^0\left(y_1\right)\right) + \left(1 - \pi_1^0\right) f_2^0\left(y_2 - F_2^0\left(y_1\right)\right), \tag{3.27}$$

where $F_i^0\left(y_1\right) = a_i^0 y_1 + b_i^0$ and $f_i^0 \sim \mathcal{N}\left(0, \left(\sigma_i^0\right)^2\right)$, for $i = 1, 2$. We denote by $\theta_0 = (\pi_i^0, a_i^0, b_i^0, \sigma_i^0)_{i=1,p_0}$ the global parameter of the model.

For every example, we pick equal standard errors $\sigma_1^0 = \sigma_2^0 = 0.5$, $b_1^0 = 0.5$ and $b_2^0 = -0.5$ and let vary the rest of the coefficients: $\pi_1^0 \in \{0.5, 0.7, 0.9\}$, $a_1^0, a_2^0 \in \{0.1, 0.5, 0.9\}$. We also varied the length of the series, $n = 200, 500, 1000, 1500, 2000$, and simulated 20 samples for each of the possible cases. The upper bound for the number of regimes, P, was fixed equal to 3.

The likelihood was maximized using an EM algorithm. The M-step provides analytical updates for the parameters, given by the following equations:

$$\hat{\pi}_i = \frac{1}{n-1} \sum_{t=2}^{n} h_i^\star(y_t, y_{t-1}) \ , \tag{3.28}$$

$$\hat{a}_i = \frac{\sum_{t=2}^{n}\sum_{t'=2}^{n} h_i^\star(y_t, y_{t-1}) h_i^\star(y_{t'}, y_{t'-1}) y_t (y_{t-1} - y_{t'-1})}{\sum_{t=2}^{n}\sum_{t'=2}^{n} h_i^\star(y_t, y_{t-1}) h_i^\star(y_{t'}, y_{t'-1}) y_{t-1} (y_{t-1} - y_{t'-1})} \ , \tag{3.29}$$

$$\hat{b}_i = \frac{\sum_{t=2}^{n}\sum_{t'=2}^{n} h_i^\star(y_t, y_{t-1}) h_i^\star(y_{t'}, y_{t'-1}) y_t y_{t'-1} (y_{t-1} - y_{t'-1})}{\sum_{t=2}^{n}\sum_{t'=2}^{n} h_i^\star(y_t, y_{t-1}) h_i^\star(y_{t'}, y_{t'-1}) y_{t-1} (y_{t'-1} - y_{t-1})} \ , \tag{3.30}$$

$$\hat{\sigma}_i = \frac{\sum_{t=2}^{n} (y_t - \hat{a}_i y_{t-1} - \hat{b}_i)^2 h_i^\star(y_t, y_{t-1})}{\sum_{t=2}^{n} h_i^\star(y_t, y_{t-1})} \ , \tag{3.31}$$

with the notation

$$h_i^\star(y_t, y_{t-1}) = \frac{\pi_i^\star f_i^\star(y_t - F_i^\star(y_{t-1}))}{\sum_{j=1}^{p} \pi_j^\star f_j^\star(y_t - F_j^\star(y_{t-1}))} \ , \tag{3.32}$$

and where $\pi_i^\star$, $f_i^\star$ and $F_i^\star$ are computed with the current values of the parameter in the EM procedure, $\theta^\star$. To avoid local maxima, the algorithm was initialized several times with different starting values: in our case, ten different initializations provided good results. The BIC penalty was used in the criterion.

The summary of the results is given by Table 3.1. In almost every case, the convergence is reached for the samples containing 2000 inputs. In practice, the results will be then more or less accurate, depending on the size of the sample, but also on the proximity of the components and on their frequency.

	π_1^0	0.5			0.7			0.9		
	n	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$
$a_1^0 = 0.1$	200	20	0	0	20	0	0	20	0	0
$a_2^0 = 0.1$	500	18	2	0	18	2	0	20	0	0
	1000	14	6	0	9	11	0	11	9	0
	1500	6	14	0	4	16	0	5	15	0
	2000	5	15	0	0	20	0	1	19	0
$a_1^0 = 0.1$	200	12	8	0	13	7	0	20	0	0
$a_2^0 = 0.5$	500	11	9	0	6	14	0	18	2	0
	1000	0	20	0	1	19	0	14	6	0
	1500	0	20	0	0	20	0	8	12	0
	2000	0	20	0	0	20	0	7	13	0
$a_1^0 = 0.1$	200	0	20	0	4	16	0	17	3	0
$a_2^0 = 0.9$	500	0	20	0	0	20	0	9	11	0
	1000	0	20	0	0	20	0	9	11	0
	1500	0	20	0	0	20	0	4	16	0
	2000	0	20	0	0	20	0	0	20	0

Table 3.1: Selected number of components in the case where the true model is a mixture of two linear regressions.

3.4 Mixtures of nonlinear autoregressive models

In a subsequent work [MO12], we gave sufficient conditions and proved the consistency of a penalized likelihood criterion in the more general case of mixtures of nonlinear autoregressive models.

In this framework, the *true* model is:

$$Y_t = F^0_{\theta_{X_t}}(Y_{t-1}) + \varepsilon_{\theta_{X_t}}(t) \ , \tag{3.33}$$

where

- X_t is an iid sequence of random variables valued in a finite space $\{1, ..., p_0\}$ and with probability distribution π^0 ;
- for every $i \in \{1, ..., p_0\}$, $F^0_{\theta_i}(y) \in \mathcal{F} = \left\{F_\theta,\ \theta \in \Theta,\ \Theta \subset \mathbb{R}^l \text{ compact set}\right\}$ is the family of possible regression functions. Here, it is supposed that $F^0_{\theta_i}$ are sub-linear: they are continuous and there exist $\left(a_i^0, b_i^0\right) \in \mathbb{R}_+^2$ such that $\left|F^0_{\theta_i}(y)\right| \leq a_i^0 |y| + b_i^0$, $\forall y \in \mathbb{R}$;

- for every $i \in \{1, ..., p_0\}$, $(\varepsilon_{\theta_i}(t))_t$ is an iid noise such that $\varepsilon_{\theta_i}(t)$ is independent of $(Y_{t-k})_{k\geq 1}$. Moreover, $\varepsilon_{\theta_i}(t)$ has a centered Gaussian density $f^0_{\theta_i}$.

As I will illustrate later, the sub linearity condition on the regression functions is quite general, and the consistency for the number of components holds for various classes of functions, including mixtures of multilayer perceptrons.

I will also remark here that the compactness hypothesis is also useful in practice, and not only for proving the theoretical result. Indeed, one usually needs to bound the parameter space in order to avoid numerical problems such as hidden-units saturation for multilayer perceptrons.

Some regularity conditions are further needed, in order to have strict stationarity and geometric ergodicity of the stochastic process Y_t and prove the consistency of the estimate for the number of regimes:

$$\textbf{(HS)} \qquad \sum_{i=1}^{p_0} \pi_i^0 \left|a_i^0\right|^s < 1. \tag{3.34}$$

The hypothesis **(HS)** does not request every component to be stationary, and it allows non-stationary *regimes*, as long as they do not appear too often. This property is quite interesting in practice for applications. More particularly, since multilayer perceptrons are bounded functions, this hypothesis will be naturally fulfilled.

With the model introduced above, the *true* conditional density of Y_t, conditionally to Y_{t-1} and marginally in X_t may be written as:

$$g^0(y_t \mid y_{t-1}) = \sum_{i=1}^{p_0} \pi_i^0 f^0_{\theta_i}\left(y_t - F^0_{\theta_i}(y_{t-1})\right) \tag{3.35}$$

For some large fixed P, representing the maximal number of regimes, the class of possible densities defined in Equation now writes as

$$\mathcal{G}_P = \bigcup_{p=1}^{P} \mathcal{G}_p \ , \ \mathcal{G}_p = \left\{ g(y_1, y_2) = \sum_{i=1}^{p} \pi_i f_{\theta_i}(y_2 - F_{\theta_i}(y_1)), \pi_i \geq \eta > 0, \sum_{i=1}^{p} \pi_i = 1, F_{\theta_i}(y) \in \mathcal{F}, f_{\theta_i} \sim \mathcal{N}(0, \sigma_i^2) \right\}. \tag{3.36}$$

Similarly to Section 3.2.4, and for every $g \in \mathcal{G}_P$, one may define the number of regimes $p(g)$ as in Equation 3.15 and the estimate of the number of regimes $\hat{p}$ as the argument minimizing the criterion $T_n(p)$ in Equation 3.16, where the term $l_n(g)$ in the definition of $T_n(p)$ is the log-likelihood marginal in $(X_k)_k$, $l_n(g) = \sum_{k=2}^{n} \ln g(y_k|y_{k-1})$.

With the previous notations and definitions, one may now state the following theorem, similar to 3.2.3, which is an extension of a result for mixtures of distributions in [56]. One may easily see that, with the exception of assumption (A4), the rest of conditions are easily verifiable by many practical model configurations.

Theorem 3.4.1 *Consider the regression model $(Y_k, X_k)_k$ defined by Equation 3.33 and the penalized-likelihood criterion defined in Equation* **??**, *with $l_n(g) = \sum_{k=2}^{n} \ln g(y_{k-1}, y_k)$ the marginal likelihood in $(X_k)_k$, for a given observed sample $(y_1, ..., y_n)$. Let us introduce the next assumptions :*

(A1) $a_n(\cdot)$ is an increasing function of p, $a_n(p_1) - a_n(p_2) \to \infty$ when $n \to \infty$ for every $p_1 > p_2$ and $\frac{a_n(p)}{n} \to 0$ when $n \to \infty$ for every p;

(A2) the model verifies the weak identifiability assumption

$$\sum_{i=1}^{p} \pi_i f_i(y_2 - F_i(y_1)) = \sum_{i=1}^{p_0} \pi_i^0 f_i^0\left(y_2 - F_i^0(y_1)\right) \Leftrightarrow \sum_{i=1}^{p} \pi_i \delta_{\theta_i} = \sum_{i=1}^{p_0} \pi_i^0 \delta_{\theta_i^0} \tag{3.37}$$

(A3) the parameterization $\theta_i \to f_{\theta_i}(y_2 - F_{\theta_i}(y_1))$ is continuous for every (y_1, y_2) and there exists $m(y_1, y_2)$ an integrable map with respect to the stationary measure of (Y_k, Y_{k-1}) such that $|\ln(g)| < m$.

(A4) Y_k is strictly stationary and geometrically β-mixing, and the family of generalized score functions associated to $\mathcal{G}_P$,

$$\mathcal{S} = \left\{ s_g, \ s_g(y_1, y_2) = \frac{\frac{g(y_1,y_2)}{g^0(y_1,y_2)} - 1}{\left\|\frac{g}{g^0} - 1\right\|_{L^2(\mu)}}, g \in \mathcal{G}_P, \left\|\frac{g}{g^0} - 1\right\|_{L^2(\mu)} \neq 0 \right\} \subset \mathcal{L}_2(\mu) \ , \tag{3.38}$$

where μ is the stationary measure of (Y_k, Y_{k-1}), verifies, for every $\varepsilon > 0$,

$$\mathcal{H}_{[\cdot]}\left(\varepsilon, \mathcal{S}, \|\cdot\|_2\right) = \mathcal{O}\left(|\ln \varepsilon|\right) \ , \tag{3.39}$$

where $\mathcal{H}_{[\cdot]}\left(\varepsilon, \mathcal{S}, \|\cdot\|_2\right)$ is the ε-bracketing entropy of $\mathcal{S}$ with respect to the L_2-norm.

Then, under hypothesis (A1)-(A4) and (HS), $\hat{p} \to p_0$ in probability.

3.4.1 An application to mixtures of multilayer perceptrons

Multilayer perceptrons (MLP) are a very useful class of regression models, particularly for handling nonlinear data. They have been intensively used by the machine learning community in the 90's, and are currently making an impressive comeback, with the re-emergence of artificial intelligence, and more powerful computational facilities. When combined with hidden-Markov models, MLP's are able to capture complex regime switches and fit complex nonlinear structures. We illustrated this in a case-study published during my PhD [MO17], while some theoretical properties had been derived by [32] around the same time.

With this in mind, we checked whether the assumptions of the above theorem applied for mixtures of MLP's, so that one can safely use the BIC criterion for model selection purposes. Since non-indentiafiability issues also arise in multilayer perceptrons, the problem was simplified by considering one hidden layer only, and a fixed number of hidden units, k. Then, the i-th component of true model writes, for any $i = 1, ..., p_0$:

$$F^0_{\theta_i}(y) = \alpha_0^{0,i} + \sum_{j=1}^{k} \alpha_j^{0,i} \phi\left(\beta_{0,j}^{0,i} + \beta_{1,j}^{0,i} y\right) \ , \tag{3.40}$$

where where ϕ is the hyperbolic tangent and $\theta_i^0 = \left(\alpha_0^{0,i}, \alpha_1^{0,i}, ..., \alpha_k^{0,i}, \beta_{0,1}^{0,i}, \beta_{1,1}^{0,i}, ..., \beta_{0,k}^{0,i}, \beta_{1,k}^{0,i}, \sigma^{0,i}\right)$ is the true parameter in the i-th component of the mixture.

(HS) The stationarity and ergodicity assumption (HS) is immediately verified since the output of every perceptron is bounded, by construction. Thus, every regime is stationary, and the global model is also stationary.

(A1) The penalty $a_n(\cdot)$ may be chosen, for example, as in the BIC criterion, $a_n(p) = \frac{1}{2} p \ln(n)$.

(A2)-(A3) If one considers the class of possible densities $\mathcal{G}_P$ as in Equation 3.36, with $F_{\theta_i}(y) = \alpha_0^i + \sum_{j=1}^k \alpha_j^i \phi\left(\beta_{0,j}^i + \beta_{1,j}^i y\right)$, $f_{\theta_i} \sim \mathcal{N}(0, \sigma_i^2)$ and $\theta_i = \left(\alpha_0^i, \alpha_1^i, ..., \alpha_k^i, \beta_{0,1}^i, \beta_{1,1}^i, ..., \beta_{0,k}^i, \beta_{1,k}^i, \sigma^i\right) \in \Theta$, a compact set, the hypothesis (A2) and (A3) are immediately verified. Indeed, Gaussian distributions are both weakly identifiable, and verify the regularity conditions in the theorem.

(A4) This assumption is the only one difficult to prove, since one needs to show that she may control the *dimension* of the class of generalized score functions, $\mathcal{S}$, defined in Equation 3.38. Proving that a parametric family like $\mathcal{S}$ verifies the condition on the bracketing entropy is usually immediate under some good regularity conditions, as given for example in [57]. A sufficient condition is to express the bracketing number as a polynomial function of $\frac{1}{\varepsilon}$. In our case, problems arise when $g \to g^0$ and the limits in $L^2(\mu)$ of s_g have to be computed.

The solution consists in splitting $\mathcal{S}$ into two classes of functions, $\mathcal{S} \setminus \mathcal{S}_0$ and $\mathcal{S}_0$, where $\mathcal{S}_0 = \{s_g,\ g \in \mathcal{F}_0\}$ and $\mathcal{F}_0 = \left\{g \in G_P,\ \left\|\frac{g}{g^0} - 1\right\|_{L^2(\mu)} \leq \varepsilon\right\}$ is a neighborhood of g^0.

On $\mathcal{S} \setminus \mathcal{S}_0$, one may use the setting in [57] and easily prove that the number of ε-brackets necessary to cover $\mathcal{S} \setminus \mathcal{S}_0$ is $\mathcal{N}_{[]}\left(\varepsilon, \mathcal{S} \setminus \mathcal{S}_0, \|\cdot\|_2\right) = \mathcal{O}\left(\frac{1}{\varepsilon}\right)^{6(k+1)P}$.

On $\mathcal{S}_0$, one may use an idea proposed in [55] and re-parameterize the model in a convenient way. Indeed, one may first remark that when $\frac{g}{g^0} - 1 = 0$, and by the weak identifiability assumption in (A2) and the fact that $\pi_i \geq \eta > 0$, for all $i = 1, ..., p$, there exists a vector $t = (t_i)_{0 \leq i \leq p_0}$ such that $0 = t_0 < t_1 < ... < t_{p_0} = p$ and such that, up to a permutation,

$$\theta_{t_{i-1}+1} = ... = \theta_{t_i} = \theta_i^0 \ , \ \sum_{j=t_{i-1}+1}^{t_i} \pi_j = \pi_i^0 \ , \ i \in \{1, ..., p_0\} \, . \tag{3.41}$$

With this remark in mind, one may define $s = (s_i)_{1 \leq i \leq p_0}$ and $q = (q_j)_{1 \leq j \leq p}$ such that, for every $i \in \{1, ..., p_0\}$ and $j \in \{t_{i-1} + 1, ..., t_i\}$,

$$s_i = \sum_{j=t_{i-1}+1}^{t_i} \pi_j - \pi_i^0, \; q_j = \frac{\pi_j}{\sum_{l=t_{i-1}+1}^{t_i} \pi_l} \; , \tag{3.42}$$

and introduce the new parameterization (ϕ_t, ψ_t), such that $\phi_t = ((\theta_j)_{1 \leq j \leq p}, (s_i)_{1 \leq i \leq p_0 - 1})$ and $\psi_t = (q_j)_{1 \leq j \leq p}$. The advantage of this new writing is that ϕ_t now contains all the identifiable part of the model, whereas ψ_t contains the non-identifiable one. For $g = g^0$, one has that

$$\phi_t^0 = \; (\underbrace{\theta_1^0, ..., \theta_1^0}_{t_1} \; , ..., \; \underbrace{\theta_{p_0}^0, ..., \theta_{p_0}^0}_{t_{p_0} - t_{p_0 - 1}}, \; \underbrace{0, ..., 0}_{p_0 - 1} \;)^T \; . \tag{3.43}$$

With this new parameterization and by remarking that when $\phi_t = \phi_t^0$, $\frac{g}{g^0}$ does not vary with ψ_t, one may derive a second-order Taylor expansion of $\frac{g}{g^0} - 1$ at ϕ_t^0, and embed $\mathcal{S}_0$ into a set of function which bracketing number is smaller or equal to $O\left(\frac{1}{\varepsilon}\right)^{3p_0 \times (3k+1)+1}$.

Hence, we are able to prove the consistency of the estimate for the number of regimes, under very mild assumptions, which are mainly Gaussian noise, bounded from below mixing probabilities and a fixed structure for the multilayer perceptrons. This results is very important for practical applications.

An example on a real dataset We illustrated the model selection criterion on the complete laser series excerpted from the *Santa Fe time series prediction and analysis competition*, [60]. The level of noise in this series is very low, the main source being the errors of measurement. The length of the series used for estimation was 12,500, and we displayed the last 1,000 patterns in Figure 3.1.

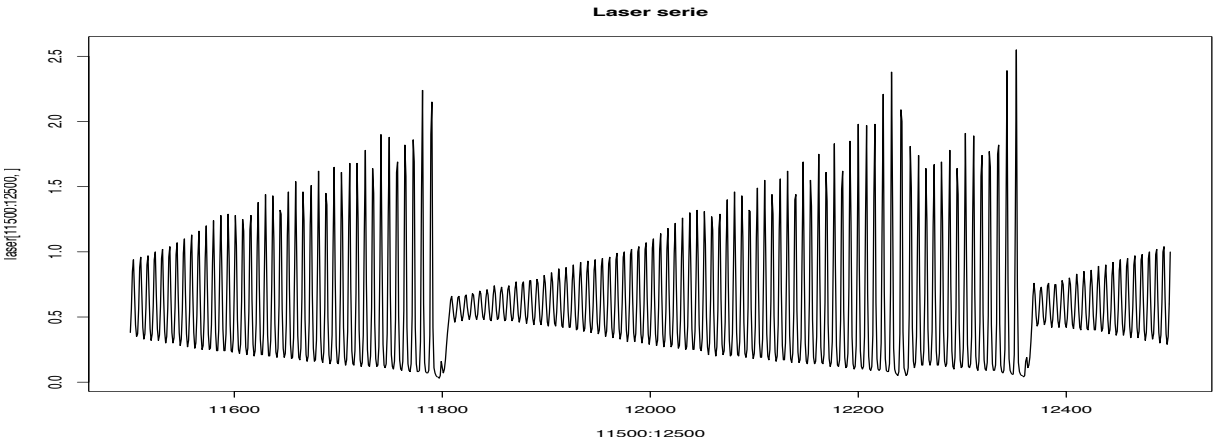


Figure 3.1: Laser series used for training, the last 1,000 time instants.

The architecture of the multilayer perceptrons was based on a previous study [32], and contains 10 entries, hence 10 time lags, and 5 units on the hidden layer. The activation function is a hyperbolic tangent, and the penalty term in the criterion is equal to that of the BIC. As this is a real application, it is impossible to check the main assumption of our theory, and verify whether the true model belongs to the set of possible ones. Nevertheless, the developed theory provides an insight on how to choose the number of experts.

The parameters were again estimated using the standard EM algorithm. In order to avoid local maxima, the algorithm was trained with 100 different initializations. For each estimation, we used 200 iterations of the EM algorithm and for each M-step we optimized the parameters of the multilayer perceptrons until their error of prediction didn't improve any longer. Here also, and mainly for computational reasons, the maximum number of regimes P was fixed equal to 3.

Table 3.2 displays a summary of the results, with the values of the penalized log-likelihood criterion and the mixture probabilities for the best architectures with one, two, and respectively three regimes.

According to these results, the best model contains two regimes. It is however difficult to give an interpretation of the regimes, as we did for the historical data in the previous chapter, since the mixing probabilities remain constant over time, and the switches are independent. We mentioned in the article that according to the

Number of experts	Penalized criterion with BIC penalty	Mixture probabilities
1	-32.17	1
2	-25.92	(0.8 ; 0.2)
3	-38.42	(0.7 ; 0.21 ; 0.09)

Table 3.2: Penalized criterion results for the Laser series

prediction made by each expert, it appeared that one of them was specialized in the general regime of the series, while the other in the collapse regime. Nevertheless, I was unable to find in my archives the estimated values of the parameters, or any other supplementary information that I could add here to help with the interpretation.

In any case, our main goal was to prove that a penalized log-likelihood criterion is consistent and allows to select the number of components with a sound theoretical basis. I believe our work brought a step in this direction. One would naturally wish to improve these models, so that the mixing probabilities would no longer be independent. One option would be to make them depend on the previous values of the time series as in the *gating experts* models introduced by [61], or of time, as in hidden Markov models. In practice, these models prove to be much more competitive, and the theory needed to deal with such complex modeling still needs to be improved.

3.5 Mixtures of experts models

Derived from neural networks literature, *mixtures of experts* (ME) [38] and *hierarchical mixtures of experts* (HME) [39] generalize linear regression models. HME are mixtures of *experts* (for example, linear regression models) organized in a tree-structured network. The network assigns a weight to each expert and then produces an output which combines the outputs produced by all experts according to their weights. Unlike mixtures of regression models, the weights depend on some input z. The ME discussed in this paper is a particular case of HME, where the network has only one layer.

The conditional density of a ME can be generally written as:

$$g\left(y|z,\phi\right)=\sum_{i=1}^{p}\pi_{\nu_i}(z)g_{\theta_i}(y|z), \tag{3.44}$$

where $\phi=\left(\nu_1^T,...,\nu_p^T,\theta_1^T,...,\theta_p^T\right)$ is the parameter of the model. Usually, the weights or *gating functions* are chosen to be logistic type

$$\pi_{\nu_i}(z)=\frac{\exp\left(\nu_i^T z\right)}{\sum_{j=1}^{p}\exp\left(\nu_j^T z\right)}, \tag{3.45}$$

while g_θ may be Poisson, Binomial or Gaussian distributed.

When the model is assumed to be correctly specified, the maximum likelihood estimates converge to the true values of the parameters and are normally distributed [62]. However, the true model is not usually known and the true parameter is unidentifiable.

Theorems 3.2.1 and 3.2.2 may be extended in a quite straightforward way to this class of models, as we proved in [MO11]. One may also prove that the LRTS is divergent if there exists a sequence of parameters $\nu_1,...,\nu_k,...$ such that $\lim_{k\to\infty}\mathbb{E}(\pi_{\nu_k}(Z))=0$. If, on the contrary, one imposes that $\exists\delta>0$ such that $\forall\mu$, $\mathbb{E}(\pi_{\nu_k}(Z))\geq\delta$, the LRTS is tight and the true number of components in the mixture may be selected thanks to classical penalized log-likelihood criteria such as the BIC. Since both the assumptions and the proofs of these results are very similar to those in the previous sections, I will not go into any further pointless technical details here.

3.6 What about autoregressive Markov-switching models?

Several times in this chapter I mentioned that the starting point of the work presented within it was a case study, where we had trained a hidden Markov model combined with multi-layer perceptrons on a financial time series, and where we had to deal with the model selection issue. The latter proved to be a very complicated question to tackle directly, which is the reason for which we first looked at mixtures, and derived the consistency of the BIC criterion.

Eventually, we went back to our starting point and investigated whether one could generalize the theorems for mixtures to the hidden Markov case. The model in Equation 3.1 now states that X_t is a homogeneous Markov

chain, irreducible and aperiodic. Let $\theta_1^0, ..., \theta_{p_0}^0$ denote the true parameters in the regression functions F, and $\Pi = (\pi_{ij}^0)_{i,j=1,...,p_0}$ the true transition matrix, where p_0 is the true number of regimes.

According to Yao and Attali (2000), this model has a unique strictly-stationary and geometrically-ergodic solution under the following hypothesis:

(HS1) The regression functions F_θ are sublinear, that is they are continuous and $\forall \theta \in \Theta$, $\exists\, (a_i, b_i) \in \mathbb{R}_+^2$ such that $|F_{\theta_i}(y)| \leq a_i |y| + b_i$, $(\forall)\ y \in \mathbb{R}$;

(HS2) For every $i \in \{1, ..., p_0\}$, the noise $(\varepsilon_{\theta_i}(t))$ has a strictly positive density f_{θ_i}, with respect to the Lebesgue measure, and there exists $s \geq 1$ such that $\mathbb{E}|\varepsilon_{\theta_i}(1)|^s < \infty$.

(HS3) The spectral radius $\rho(Q_s) < 1$, where

$$Q_s = \begin{pmatrix} \left(a_1^0\right)^s \pi_{11}^0 & \cdots & \left(a_{p_0}^0\right)^s \pi_{1p_0}^0 \\ \vdots & \ddots & \vdots \\ \left(a_1^0\right)^s \pi_{p_0 1}^0 & \cdots & \left(a_{p_0}^0\right)^s \pi_{p_0 p_0}^0 \end{pmatrix} \tag{3.46}$$

The hypothesis **(HS3)** is clearly verified whenever $a_i^0 < 1$, for all $i \in \{1, ..., p_0\}$.

3.6.1 On the difficulty of defining a contrast function

Let $\{Y_1, ..., Y_n\}$ be a n-sample stemming from the true model introduced in Equation 3.1, with X_t is a homogeneous Markov chain, irreducible and aperiodic. One may attempt to extend the criteria in the previous sections. As we proved in [MO12], when trying to do so, several issues arise: on the one hand, the non-identifiability, one but we have already seen that this could be dealt with using a good reparameterization of the model, and, on the other hand, the dependency structure of $(X_t)_t$. Indeed, the dependency will not allow for an explicit form for the conditional density, marginally in X_t:

$$f^0(Y_k \mid Y_{k-1}, ..., Y_0) = \sum_{i=1}^{p_0} \mathbb{P}_{\theta^0}(X_k = i \mid Y_{k-1}, ..., Y_0)\, f_{\theta_i^0}(Y_k - F_{\theta_i^0}(Y_{k-1}))\ , \tag{3.47}$$

since $\mathbb{P}_{\theta^0}(X_k = i \mid Y_{k-1}, ..., Y_0)$ has to be computed recursively. Nevertheless, since X_t is stationary and following the same idea as [56], a cost function which involves the invariant probability measure of the hidden Markov chain can still be defined.

Hence, instead of the true log-likelihood, one may consider an approximation where the conditional probabilities $\mathbb{P}_{\theta^0}(X_k = i \mid Y_{k-1}, ..., Y_0)$ are replaced by the invariant measure of the Markov chain. With this in mind, a cost function may be defined as

$$\tilde{l}_n(g) = \sum_{t=2}^{n} \ln g(Y_{t-1}, Y_t) = \sum_{t=2}^{n} \ln(\sum_{i=1}^{p} \pi_i f_{\theta_i}(Y_t - F_{\theta_i}(Y_{t-1})))\ , \tag{3.48}$$

where $g \in \mathcal{G}_{\mathcal{P}}$, and

$$\mathcal{G}_{\mathcal{P}} = \left\{ g \mid g(y_1, y_2) = \sum_{i=1}^{p} \pi_i f_{\theta_i}(y_2 - F_{\theta_i}(y_1)), \theta_i \in \Theta, p \leq P \right\} \tag{3.49}$$

is the class of possible densities, with Θ a compact set and P the maximum number of regimes. One would intuitively expect that $\tilde{l}_n(g)$ is maximized by the *true density*,

$$g^0(y_1, y_2) = \sum_{i=1}^{p_0} \pi_i^0 f_{\theta_i^0}(y_2 - F_{\theta_i^0}(y_1)), \tag{3.50}$$

where π_i^0 is the invariant probability under the true model. When trying to prove this, one will eventually find that:

- $\tilde{l}_n(g)$ is maximized by g^0 if the regime switches are independent, which corresponds to the case of mixtures, already studied in the previous sections.
- $\tilde{l}_n(g)$ is maximized by g^0 if the regression functions F_θ are constant, which corresponds to "simple" hidden Markov models, already treated in [56].
- in the general case, there is no reason for g^0 to maximize $\tilde{l}_n(g)$, as proven by some simulations here below.

3.6.2 Simulation results

We performed various simulations where the true model was a two-regimes process. I will summarize some of them here. We considered three possibilities for the transition matrix, $\Pi_1^0 = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$, $\Pi_2^0 = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$ and $\Pi_3^0 = \begin{pmatrix} 0.9 & 0.5 \\ 0.1 & 0.5 \end{pmatrix}$. The first transition matrix corresponds to independent regime switches. The regression functions F_θ were considered either linear, or constant. The latter correspond to hidden Markov chains. The noise was considered normally distributed $N\left(0, 0.5^2\right)$ and the log-likelihood was penalized with the BIC penalty term. For every model, several sample sizes were considered (from 200 up to 2000 input values) and for each model and sample size, twenty different samples were simulated. For each case, Table 3.3 contains the estimated number of regimes (the maximum number P was fixed equal to three).

		Π_1^0			Π_2^0			Π_3^0		
	n	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$
$F_1^0(y) = 0.8y - 1$	200	0	20	0	**0**	**15**	**5**	**0**	**17**	**3**
$F_2^0(y) = 0.3y + 1$	500	0	20	0	**0**	**17**	**3**	**0**	**8**	**12**
	1000	0	20	0	**0**	**6**	**14**	**0**	**4**	**16**
	1500	0	20	0	**0**	**1**	**19**	**0**	**5**	**15**
	2000	0	20	0	**0**	**1**	**19**	**0**	**5**	**15**
$F_1^0(y) = -1$	200	0	20	0	0	20	0	0	20	0
$F_2^0(y) = 1$	500	0	20	0	0	20	0	0	20	0
	1000	0	20	0	0	20	0	0	20	0
	1500	0	20	0	0	20	0	0	20	0
	2000	0	20	0	0	20	0	0	20	0

Table 3.3: Selected number of regimes for the BIC-penalized criterion with the approximated log-likelihood defined in Equation 3.48.

Simulation results prove that the penalized estimate $\hat{p}$ converges when the true model has independent switches (transition matrix Π_1^0), or when the regression functions are constant. The criterion diverges when the true model is a two-regime autoregressive Markov-switching model. This means that the cost function that was considered as a generalization of the *marginal likelihood* does not have the right properties to be a contrast function and the problem of estimating p_0 remains open in the general case of autoregressive Markov switching models.

We have tested empirically the behavior of a penalized criterion using the exact likelihood (see Table 3.4), and in this case the convergence was achieved. Hence, for practical applications, BIC criterion with exact likelihood provides reliable results. However, we do not have the theoretical foundations for this result yet. Some first results were published in [63], who showed that the LRTS was divergent for hidden Markov models, but that a penalized likelihood estimate was consistent under some good hypothesis.

		Π_1^0			Π_1^0		
	n	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$
$F_1^0(y) = 0.8y - 1$	200	0	16	4	0	15	5
$F_2^0(y) = 0.3y + 1$	500	0	16	4	0	19	1
	1000	0	17	3	0	19	1
	1500	0	18	2	0	19	1
	2000	0	19	1	0	20	0

Table 3.4: Selected number of regimes for the BIC-penalized criterion with the exact log-likelihood

3.7 Conclusion and perspectives

In this chapter, I aimed at looking with a more theoretical eye at how autoregressive regime-switching models – either with independent regime switches, or with Markovian ones – may be used in applications, and under which conditions they prove to be not only useful, but also reliable in practice.

Under some very general conditions, mainly related to the complexity of the class of generalized score functions, we proved the consistency of a penalized log-likelihood criterion for selecting the number of regimes, in the case

of independent regime switches. We also checked that these conditions were verified by some classes of common models, such as mixtures of linear autoregressive models, mixtures of multilayer perceptrons and mixtures of experts.

Let me recall here that, due to identifiability issues and to the complexity of the model, we supposed in all the results above that the architecture of the multilayer perceptrons, for example, was priorly known (number of hidden layers, number of hidden units in the layers), and the model selection issue focused on the number of "experts" only. This framework is obviously not a completely realistic one, and one may want to drop this constraint and select both the number of regimes and the architecture of the model in each regime. Let me mention here also, that [64] had already investigated the consistency of a penalized likelihood criterion for selecting the number of hidden units in a one hidden-layer perceptron. It would be of course of interest to study if and how the two results may be mixed, and if the consistency of the BIC and of related versions would still hold with a double source of non-identifiability. Although I left aside these topics for a while, at some point I would like to go back to studying them.

As for Markovian regime switches, establishing consistency results in the case of autoregressive components proved to be a difficult task. The criterion based on the approximation of the marginal likelihood failed to select the correct number of regimes if the autoregressive functions were not constants. To my knowledge (although not exactly up to date!), the question remains open. Using the exact likelihood provided consistent results on several simulations, but the theoretical proofs were not established. I do acknowledge that I would very much like to take some time in the years to come and go back to these issues, while updating on the latest results and emerging questions.

Chapter 4

Self-organizing maps for complex data

"*Friendship is everything. Friendship is more than talent. It is more than the government. It is almost the equal of family.*"
Don Corleone

4.1 Introduction

When starting to write this chapter, the first thing coming to my mind was that this part of my work had a great deal to do with a family affair. And, of course, this goes beyond the Italian background of one of the colleagues involved in this collaboration. More seriously, I had the privilege of being a SAMM member at the same time and together with three recognized experts in self-organizing maps (SOM) and related methods. I'm talking here about Marie Cottrell, Nathalie Vialaneix and Fabrice Rossi. They are my colleagues, but also my friends, and in some sense my family also. The fruitful interactions and discussions we had together over the years led to a series of developments and extensions of SOM for complex data that I will present next.

Indeed, right after my PhD I became involved in various applied research projects which were aimed at exploring complex data, such as categorical sequences or longitudinal data, social networks, corpuses of historical documents,... Most of this data stemmed from humanities and social sciences, had a temporal component, and was not easily described by numerical vectors. However, in most of these cases, the data could be known through pairwise similarities (kernels) or dissimilarities, but also through multiple dissimilarities or similarities.

In this context, self-organizing maps appeared as a very convenient tool, on the one hand, due to their nice properties of both clustering and visualization, and, on the other hand, since extensions of the original algorithm, designed for numerical vectors, had already been proposed in the literature, and the question of further developing them, while summarizing the existing state of the art into a unified view, was a topical one.

Inspired from neurobiological learning paradigms, self-organizing maps (SOM) were introduced by T. Kohonen [65, 66], as both a clustering and a visualization technique, based on a vector quantization principle. They may be seen as a generalized version of the k-means algorithm with a neighborhood constraint on the cluster structure, which provides very nice properties for visualization, making them a popular technique in practical applications (over 10,000 papers published on SOM in 2016). More precisely, the original data is mapped onto a lower dimensional space, usually a finite two-dimensional rectangular grid. The grid or the map is equipped with a topology which allows defining a distance between its units. Each unit of the grid is also equipped with a prototype, a mathematical object living in the same space as the original data. The algorithm aims at partitioning the original space, the number of elements in the partition being equal to the size of the map, with the constraint of topology preservation: if two input data are close in the original space, they should be mapped onto the same unit on the grid, or onto neighbor units. The prototype of a cell is then a generalized barycenter, computed as a weighted mean of the input data in the cell and also the neighboring ones. The weights are supposed to be decreasing with the neighborhood radius.

Initially designed for numerical vector data, SOM was extended starting with the late 90's to non-vector data, giving rise to an abundant literature. I will not review it here (some elements are available in our review papers [MO60] and [MO4]), and I will focus instead on a particular class of extensions, relational and kernel SOM. This chapter summarizes the contributions we proposed for handling complex data described by similarities or dissimilarities, and more particularly the extensions for large datasets, which are based either on bagging, or on sparsified versions.

In Section 4.2, I will start by presenting the original SOM algorithm, designed for numerical vectors, both in *online* and *batch* frameworks. This will serve as a basis for presenting the kernel and relational versions in Section 4.3, where I will try to give a unified view of the different approaches, while stressing our contribution. I will also discuss here the complexity issues arising in relational frameworks and present our contribution for an accelerated implementation which improves the computational burden from a cubic to a quadratic one. I will illustrate how the relational algorithm works in practice on a graph summarizing the characters in the novel *Les Misérables*. This example was published in [MO7], and is part of the R-package SOMbrero, that we developed during this collaboration, and which is available on CRAN. SOMbrero provides, among others, online implementations for the numerical and relational algorithms, and comes with a user-friendly Shiny interface which makes it appealing for researchers in humanities and social sciences.

Our contribution to exploring data described by multiple kernels or multiple dissimilarities is presented in Section 4.4. I illustrate this approach with an example on longitudinal data, encoding for career paths of young high-school graduates, excerpted from [MO34]. The efficient extensions of relational SOM for large datasets are introduced in Section 4.5. I will describe three different approaches, based on sparse approximations, and compare and discuss them on several examples. Eventually, a conclusion providing some work in progress and perspectives for future work will end this chapter.

The work I will present here was published as three original journal papers [MO6], [MO7], [MO8], three review articles, among which one as invited authors [MO4], [MO9], [MO60], and ten peer-reviewed proceedings [MO28], [MO29], [MO30], [MO31], [MO32], [MO33], [MO34], [MO35], [MO38], [MO39].

As I have already mentioned above, the contributions I will present here were developed jointly with Nathalie Vialaneix, Fabrice Rossi and Marie Cottrell, on a time span running broadly from 2012 to 2017. Laura Bendhaiba, now a data scientist in a private company, and Julien Boelaert, now an Assistant Professor at Université Lille 2, contributed to the writing of the SOMbrero package during respectively a master's internship (Summer 2013) and a postdoc (2014). They were both co-supervised by Nathalie Vialaneix and myself. The extensions of SOM to large datasets were mainly developed during Jérôme Mariette's PhD. Jérôme, who is now a research engineer at INRA, was supervised by Nathalie, and parts of his work were done in collaboration with Fabrice Rossi and myself. The data on career paths and school-to-job transitions was made available through a collaboration with Patrick Rousset (CEREQ), and was analyzed first during the internship and later on during the subsequent collaboration with Sébastien Massoni, whom I supervised and who is now an Assistant Professor in Economics in University of Nancy.

I mentioned a family affair in the very beginning of this chapter, and I could not end this Introduction without speaking of the broader SOM family. With the emergence of SOM and related methods (neural gas, soft topographic mapping, learning vector quantization, ...), WSOM, a workshop on self-organizing maps, was created in 1997 by T. Kohonen. This meeting gathers a small community of about sixty researchers in a friendly atmosphere every two years. I joined them in 2005 while I was doing my PhD and participated in the organization of the event in Paris. Since then, I contributed both as an author and a reviewer to most of the events. In 2017, Jean-Charles Lamirel (LORIA) and I co-organized the event in Nancy. As a member of the Steering Committee, I am involved in the sustainable organization of the conference in the years to come.

4.2 Self-organizing maps (SOM) for numerical data

The self-organizing map algorithm was initially designed, in Kohonen's seminal papers [65] and [66], for vector data $\{x_1, ..., x_n\}$ belonging to some subset $\mathcal{G}$ of $\mathbb{R}^m$. One has to specify a low-dimensional regular grid composed of U units (generally in a one or two-dimensional array) and to define a neighborhood function H on $\mathcal{U} \times \mathcal{U}$, where $\mathcal{U} = \{1, ..., U\}$ is the set indexing the units of the grid. The neighborhood function may depend on time (time will be understood here in terms of iterations in the training procedure), in which case we will use the notation H^t. Usually, H^t only depends on the grid-distance between units u and u', denoted $d(u, u')$ in the following. It is common to set $H^t : \mathbb{R}^+ \to \mathbb{R}^+$, $H^t(0) = 1$ and $\lim_{x \to +\infty} H^t(x) = 0$, which means that the neighborhood relationship is decreasing with the distance between the units. Among the most common choices for the neighborhood function, one will either use the *step function*,

$$H^t(d(u, u')) = \mathbf{1}_{d(u,u') \leq r^t} \ , \tag{4.1}$$

where r^t is the neighborhood radius, a constant that may decrease with time, or the *Gaussian kernel*,

$$H^t(d(u, u')) = \exp\left(-\frac{d(u, u')^2}{2\sigma(t)^2}\right) \ , \tag{4.2}$$

where the parameter $\sigma(t)$ may be decreasing with time, so as to reduce the intensity of the neighborhood.

To each unit of the grid u, one associates a prototype $p_u \in \mathbb{R}^m$ (in the abundant literature, prototypes are also called weight vectors, code-vectors, centroids, codebook vectors, ...). SOM algorithm aims at updating these prototypes according to the input data fed at each iteration, so that eventually they represent the input space as accurately as possible (from a quantization point of view), while preserving the topology of the data by matching the regular low-dimensional grid with the data original structure. For each prototype p_u, the set of inputs closer to p_u than to any other prototype will define the associated cluster (also called a Voronoï cell) in the input space, and the neighborhood structure on the grid will induce a neighborhood structure on the clusters. In other words, after having trained a SOM map, close inputs should belong to the same cluster, or to neighbor clusters.

In the original *online* setting (*online* meaning here that only one input data is fed at each iteration), the algorithm alternates an *assignment* step and a *representation step*, similarly to a k-means training, except for the update step which takes into account the neighborhood structure, as illustrated in Algorithm 1.

Algorithm 1 Online numerical SOM

1: For all $u = 1, \dots, U$, randomly initialize $p_u^0 \in \mathbb{R}^m$, $u = 1, ..., U$.
2: **for** $t = 1, \dots, T$ **do**
3: Randomly choose an input $x_i \in \mathcal{G}$.
4: *Assignment step*: find the unit of the closest prototype or the *best matching unit(BMU)*

$$f^t(x_i) \leftarrow \arg \min_{u=1,\dots,U} \|x_i - p_u^{t-1}\|^2 .$$

5: *Representation step*: $\forall\, u = 1, \dots, U$,

$$p_u^t \leftarrow p_u^{t-1} + \mu(t) H^t(d(f^t(x_i), u)) \left(x_i - p_u^{t-1}\right)$$

where $\mu(t)$ is a learning rate (positive, less than 1, constant or decreasing with t).
6: **end for**

In addition to the *online* setting, SOM was extended to a *batch* setting [67], where all the input data is fed at each iteration. Initially, this framework was meant for applications where the stochasticity of the online results was seen as a drawback. Conditionally to the initialization of the prototypes, batch SOM is completely deterministic, and this leads to reproducible training results. Without going into computational details, the procedure may be summarized as in Algorithm 2.

Algorithm 2 Batch numerical SOM

1: For all $u = 1, \dots, U$, randomly initialize $p_u^0 \in \mathbb{R}^m$, $u = 1, ..., U$.
2: **for** $t = 1, \dots, T$ **do**
3: *Assignment step*: For all $x_1, ..., x_n$, find the best matching units

$$f^t(x_i) \leftarrow \arg \min_{u=1,\dots,U} \|x_i - p_u^{t-1}\|^2 .$$

4: *Representation step*: $\forall\, u = 1, \dots, U$,

$$p_u^t \leftarrow \sum_{i=1}^{n} \frac{H^t(d(f^t(x_i), u))}{\sum_{i=1}^{n} H^t(d(f^t(x_i), u))} x_i$$

5: **end for**

One may easily see that if the neighborhood function is the step function with a null radius, then prototypes are computed as barycenters of the clusters, so one gets exactly the batch version of the k-means algorithm.

Batch and *online* numerical SOM have been extensively used in various applications. If *batch* SOM has the advantage of being deterministic conditionally to the initial values of the prototypes, *online* SOM was proved to be less sensitive to initial conditions, to converge (empirically) faster, and to lead to better results in terms of topology preservation [68], although PCA-based initializations for *batch* were shown to drastically improve the results as compared to random ones [69].

In terms of complexity, the assignment step in the online algorithm is $\mathcal{O}(Um)$, while the representation step is also $\mathcal{O}(Um)$. For the batch version, the assignment step is of order $\mathcal{O}(nUm)$, while the representation step is $\mathcal{O}(nUm)$. One may thus conclude that the online version is preferable, since it will generate less computational

burden. Nevertheless, empirical observations showed that in general one input data should be fed several times to the algorithm in the online setting in order to achieve a good convergence, which means that potentially the two versions could be equivalent.

Let me also mention that the success of Kohonen's algorithm in practical applications incited many researchers to study its theoretical properties. In the *batch* framework, one may show ([68], [70]), for a fixed neighborhood function, that the procedure is a quasi-Newtonian algorithm which minimizes the extended within-class variance,

$$\mathcal{E}(p_1, ..., p_U) = \sum_{u,u'=1}^{U} H(d(u,u')) \int_{x \in C_u} \|x - p_{u'}\|^2 \mu(dx) \ , \tag{4.3}$$

where C_u is the Voronoï cell or the cluster associated to prototype p_u and μ is the probability distribution of the data. Whereas in the *online* framework, despite a large amount of work and empirical evidences, the main theoretical properties of the algorithm do not have complete proofs yet, only some partial answers being currently available, as discussed, among others, in our review papers [MO60] and [MO4].

Eventually, let me remark that neither *batch* SOM, nor *online* SOM can be applied as such on data other then numerical vectors. If one wishes to extend the algorithm to another context (categorical variables, texts, time-series, graphs, ...), she should either provide some numerical summaries of the data, or modify the algorithm accordingly. Here again, there are many variants and extensions in the literature, such as the *korresp* algorithm for contingency tables [71], *median* SOM for dissimilarity data [72], ... Among all these, I will focus on kernel and relational versions only, which have the advantage of embedding the data into suitable spaces equipped with a dot product, where operations such as linear combinations become feasible. This property is particularly interesting for defining prototypes and for updating them in the representation step. In the following, I will describe these two approaches, in *batch* and *online* versions, and stress our contribution to the matter.

4.3 Kernel and relational SOM

Kernel and relational SOM are clustering procedures designed for being trained on data living in some abstract space. I will try to illustrate throughout this chapter the potential of these approaches, and notably in the context of analyzing data coming from humanities and social sciences. First, I will start by focusing on the data itself, on how it may be presented and subsequently embedded into suitable mathematical spaces so that its analysis becomes feasible. Then, I will describe the extensions of the SOM algorithm for dissimilarity data, both in *online* and *batch* versions. I will discuss the links between numerical, kernel and relational SOM and give the conditions under which they become equivalent. Also, I will focus on complexity issues which are not negligible in this context and may generate heavy computations, and describe the accelerated version we have introduced. Eventually, I will briefly describe SOMbrero, the R-package that we developed, and end with an illustrative example.

4.3.1 Kernel and dissimilarity data, how does one deal with it?

Analyzing non-standard vector data is obviously not a trivial task, especially for a quantitative social scientist who has to find her away in the prolific labyrinth of the state-of-the-art, moreover not unified, spanning from computational statistics, machine learning, signal processing, statistical physics etc. Very broadly speaking, since establishing any panorama of the existing literature is not the aim of this manuscript, she essentially has two choices: either extract some meaningful numerical features from the data (by performing a factorial analysis for categorical data, a Fourier or wavelet decomposition for time series, ...), or establish a measure of pairwise similarities or dissimilarities between the elements of her dataset. Here, I will only focus on the latter case.

Consider that the data at hand comprises n inputs $x_1, ..., x_n$ sampled in an abstract space $\mathcal{G}$. Two scenarii will be of interest.

Kernel data. In this setting, the data is supposed to be known through a similarity function known as *kernel*, $K : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$, such that K is symmetric

$$K(x, x') = K(x', x), \forall x, x' \in \mathcal{G} \ , \tag{4.4}$$

and positive definite

$$\forall m > 0, \forall (x_1, ..., x_m) \in \mathcal{G}^m, (\alpha_1, ..., \alpha_m) \in \mathbb{R}^m, \sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \geq 0 \ . \tag{4.5}$$

In some cases, instead of K, one only has the kernel matrix computed for the n inputs, $\boldsymbol{K} = (K_{ii'})_{i,i'=\overline{1,n}} = (K(x_i, x_{i'}))_{i,i'=\overline{1,n}}$. This is more restrictive, but has no impact on the developments and discussions hereafter.

Dissimilarity data. This situation is more general than the previous one, since here one assumes that the data is known through a positive similarity function $\delta : \mathcal{G} \times \mathcal{G} \to \mathbb{R}_+$, such that δ is symmetric

$$\delta(x, x') = \delta(x', x), \forall x, x' \in \mathcal{G} \ , \tag{4.6}$$

and reflexive

$$\delta(x, x) = 0, \forall x \in \mathcal{G} \ . \tag{4.7}$$

Here also, instead of δ, one often disposes of the dissimilarity matrix $\boldsymbol{\Delta} = (\delta_{ii'})_{i,i'=\overline{1,n}} = (\delta(x_i, x_{i'}))_{i,i'=\overline{1,n}}$ only.

4.3.2 Embeddings for kernel and dissimilarity data

Very few techniques allow to analyze complex data described by similarities or dissimilarities as such. This is due to the fact that some operations on the data are generally necessary for training the algorithms. Usually, this is achieved by embedding the data into some convenient metric space.

Kernel data. For data described by kernels, the Moore-Aronszajn theorem [73] states that there exists a unique Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ and a mapping function $\phi : \mathcal{G} \to \mathcal{H}$, called *feature map*, such that $K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$, $\forall x, x' \in \mathcal{G}$ and with $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ being the inner product in $\mathcal{H}$.
This embedding in a Hilbert space allows to define a natural distance in $\mathcal{G}$:

$$\delta_K^2(x, x') = \langle \phi(x) - \phi(x'), \phi(x) - \phi(x') \rangle_{\mathcal{H}} \ , \ \forall x, x' \in \mathcal{G} \ , \tag{4.8}$$

and it is immediate to see that this is equivalent to

$$\delta_K^2(x, x') = K(x, x) + K(x', x') - 2K(x, x') \ , \ \forall x, x' \in \mathcal{G} \ , \tag{4.9}$$

which means that knowing the kernel function only is sufficient for performing all the necessary operations, without the explicit knowledge of $\mathcal{H}$ or ϕ. This is also known as the *kernel trick*.

Dissimilarity data. For dissimilarity data available as a square matrix Δ, two situations may occur, according to whether Δ is Euclidean or not.

Definition 4.3.1 *A square dissimilarity matrix Δ, symmetric, positive and null on the diagonal, is Euclidean if it can be embedded in a Euclidean space $(\mathbb{R}^p, \| \cdot \|_{\mathbb{R}^p})$, where $p < n$, which means that there exists a set of vectors $\{v_1, ..., v_n\} \in \mathbb{R}^p$ such that the dissimilarities in Δ can be expressed as Euclidean distances in $\mathbb{R}^p$:*

$$\delta_{ij} = \delta(x_i, x_j) = \|v_i - v_j\|_{\mathbb{R}^p} \ , \ \forall i, j = 1, ..., n.$$

There are several ways to check whether a dissimilarity matrix is Euclidean. If Δ fulfills the conditions in [74, 75, 76] which require the matrix with elements

$$s_{ij} = \frac{1}{2} \left(\delta(x_i, x_n)^2 + \delta(x_j, x_n)^2 - \delta(x_i, x_j)^2 \right) \ , \ \forall i, j = 1, ..., n \tag{4.10}$$

to be positive definite, or, similarly, if the matrix with elements

$$s_{ij} = -\frac{1}{2} \left(\delta^2(x_i, x_j) - \frac{1}{n} \sum_{k=1}^{n} \delta^2(x_i, x_k) - \frac{1}{n} \sum_{k=1}^{n} \delta^2(x_k, x_j) + \frac{1}{n^2} \sum_{k,k'=1}^{n} \delta^2(x_k, x_{k'}) \right) \ , \ i, j = 1, ..., n \tag{4.11}$$

as proposed in [77], is positive definite, then Δ is Euclidean.

If the dissimilarity matrix Δ is not Euclidean, the Euclidean space is not "large enough" to reconstruct the dissimilarities, due to the negative eigenvalues of the corresponding similarity matrix. In this case, one may use a so called *pseudo-Euclidean* space, as described in [78], in which any premetric finite dissimilarities are embeddable.

Definition 4.3.2 *A pseudo-Euclidean space $\mathcal{E} = \mathbb{R}^{(p_+,p_-)}$ is a real vector-space equipped with a non-degenerate, indefinite, inner product $\langle\cdot,\cdot\rangle_\mathcal{E}$. $\mathcal{E}$ admits a direct orthogonal decomposition $\mathcal{E} = \mathcal{E}_+ \oplus \mathcal{E}_-$, where $\mathcal{E}_+ = \mathbb{R}^{p+}$ and $\mathcal{E}_- = \mathbb{R}^{p-}$, and the inner product is positive definite on $\mathcal{E}_+$ and negative definite on $\mathcal{E}_-$. The space $\mathcal{E}$ is therefore characterized by the signature (p_+, p_-).*

Pseudo-Euclidean spaces are more general versions of Euclidean spaces, with a correction of the non-Euclideaness: the first p_+ components can be considered as a standard Euclidean contribution, whereas the next p_- components serve as a correction. The inner product in a pseudo-Euclidean space writes naturally as $\langle\cdot,\cdot\rangle_\mathcal{E} = \langle\cdot,\cdot\rangle_{\mathcal{E}_+} - \langle\cdot,\cdot\rangle_{\mathcal{E}_-}$.

Similarly to kernels, but starting from a dissimilarity matrix Δ, one may thus state that there exists a pseudo-Euclidean space $\mathcal{E}$ and a unique *feature map* $\phi : x \in \mathcal{G} \to \phi(x) = (\phi^+(x), \phi^-(x)) \in \mathcal{E}$, such that one has the embedding

$$\delta^2(x_i, x_{i'}) = \langle \phi(x_i) - \phi(x_{i'}), \phi(x_i) - \phi(x_{i'}) \rangle_\mathcal{E} \ , \ \forall x_i, x_{i'} \in \mathcal{G} \ . \tag{4.12}$$

Let me remark here that the distance naturally induced by the pseudo-inner product is not necessarily positive, which will in practice raise numerical and convergence issues.

Link between kernel and relational approaches The two frameworks introduced above are very similar, and hence so are the algorithms based on one or the other. The relational framework is more general, but also, because of its weaker hypothesis, more difficult to assess theoretically and numerically.

If one considers a dissimilarity measure defined by the squared distance in Equation 4.9, then a kernel-based and a dissimilarity-based algorithm will be equivalent. This is why, throughout the rest of this chapter, I will only present the relational algorithms, the kernel ones being immediately tractable.

Reciprocally, suppose that the dissimilarity matrix Δ may be embedded in a Euclidean space. If this is the case, training a kernel algorithm on $\boldsymbol{K} = (K(x_i, x_j))_{i,j=1,...,n}$, where $K(x_i, x_j) = s_{ij}$ in Equations 4.10 or 4.11 is equivalent to training a relational algorithm on Δ.

Let me remark here also that if the dissimilarity matrix is Euclidean, relational SOM (both in *online* and *batch* versions) is exactly identical to the standard numerical SOM, as long as the prototypes of the numerical SOM are initialized in the convex hull of the input data.

Nevertheless, as explained in [79], some useful dissimilarities (e.g., shortest path lengths in graphs or optimal matching dissimilarities for sequences of events, [80, 81]) do not fulfill the required conditions allowing them to be embedded in a Euclidean space. In these cases, kernel and relational approaches are no longer equivalent. In particular, the similarity matrix $\boldsymbol{S} = (s_{ij})_{i,j=1,...,n}$ defined as in Equation 4.10 or 4.11 may have negative eigenvalues. If these are negligible with respect to the positive ones, one may simply consider them as noise and ignore them. If their values are high enough to generate numerical issues, then one may preprocess $\boldsymbol{S}$ and make it positive definite using techniques such as *clipping*, *flipping*, *shifting*, ..., as described in [82].

4.3.3 The algorithm

Suppose, in the following, that the data belongs to a general abstract space $x_1, ..., x_n \in \mathcal{G}$, and is known through a dissimilarity matrix, $\boldsymbol{\Delta} = (\delta_{ij})_{i,j=1,...,n} = (\delta(x_i, x_j))_{i,j=1,...,n}$, as introduced in the previous section. Additionally, define a low dimensional map of U units, each of them represented by a prototype p_u, $u = 1, ..., U$. Each prototype is expressed as a convex combination in the embedding space $\mathcal{E}$ (which may be Euclidean or pseudo-Euclidean), $\phi(p_u) = \sum_{i=1}^n \beta_{u,i}\phi(x_i)$, where $\beta_{u,i} \geq 0$, $\sum_{i=1}^n \beta_{u,i} = 1$, $\forall u = 1, ..., U$, and ϕ is the *feature map* defined in the previous subsection.

Furthermore, suppose that the map is equipped with a distance d between the units, and consider a neighborhood function H verifying the assumptions $H : \mathbb{R}^+ \to \mathbb{R}^+$, $H(0) = 1$ and $\lim_{x\to+\infty} H(x) = 0$, and μ a training parameter. H and μ are supposed to be decreasing with the number of iterations during the training procedure.

Let us now take a look at the two steps of the numerical algorithm and see how they modify in the relational framework. In the assignment step, the dissimilarity between an input data and a prototype may be then computed as:

$$\delta(x_i, p_u) = \langle \phi(x_i) - \phi(p_u), \phi(x_i) - \phi(p_u) \rangle_\mathcal{E} \equiv \boldsymbol{\Delta}_i\beta_u - \frac{1}{2}\beta_u^T\boldsymbol{\Delta}\beta_u \ , \tag{4.13}$$

where $\boldsymbol{\Delta}_i$ is the *i-th* row of $\boldsymbol{\Delta}$ (the formula is justified and proved for example in our paper [MO8]). In the representation step, it is enough to remark that the updates of the prototypes will concern the coefficients β_u only.

Batch SOM In the batch framework, the kernel version of the algorithm was first described in [83], while the generalization to the relational one was proposed in [84]. We summarize the procedure in Algorithm 3 below.

Algorithm 3 Batch relational SOM

1: For all $u = 1, \dots, U$ and $i = 1, \dots, n$, initialize $\beta_{u,i}^0$ such that $\beta_{u,i}^0 \geq 0$ and $\sum_i^n \beta_{u,i}^0 = 1$.
2: **for** $t = 1, \dots, T$ **do**
3: *Assignment step*: for each $i = 1, ..., n$, find the unit of the closest prototype

$$f^t(x_i) \leftarrow \arg \min_{u=1,\dots,U} \left(\beta_u^{t-1}\mathbf{\Delta}\right)_i - \frac{1}{2}(\beta_u^{t-1})^T \mathbf{\Delta} \beta_u^{t-1}$$

4: *Representation step*: for each $u = 1, \dots, U$,

$$\beta_{u,i}^t = \frac{H^t(d(f^t(x_i), u))}{\sum_{i'} H^t(d(f^t(x_{i'}), u))} \ , \ \forall i = 1, ..., n.$$

5: **end for**

Online SOM In the online framework, the kernel version of the algorithm was first described in [85], while we generalized it to the relational version in [MO35]. We summarize the procedure in Algorithm 4 below.

Algorithm 4 On-line relational SOM

1: For all $u = 1, \dots, U$ and $i = 1, \dots, n$, initialize $\beta_{u,i}^0$ such that $\beta_{u,i}^0 \geq 0$ and $\sum_i^n \beta_{u,i}^0 = 1$.
2: **for** $t = 1, \dots, T$ **do**
3: Randomly choose an input x_i
4: *Assignment step*: find the unit of the closest prototype

$$f^t(x_i) \leftarrow \arg \min_{u=1,\dots,U} \left(\beta_u^{t-1}\mathbf{\Delta}\right)_i - \frac{1}{2}(\beta_u^{t-1})^T \mathbf{\Delta} \beta_u^{t-1}$$

5: *Representation step*: $\forall\, u = 1, \dots, U$,

$$\beta_u^t \leftarrow \beta_u^{t-1} + \mu(t) H^t(d(f^t(x_i), u)) \left(\mathbf{1}_i - \beta_u^{t-1}\right)$$

where $\mathbf{1}_i$ is a vector with a single non null coefficient at the ith position, equal to one.
6: **end for**

Complexity and accelerating techniques When considering the relational algorithms, both in online and batch versions, one may see that the complexity of the assignment step is $\mathcal{O}(n^2U)$ and that of the representation step is $\mathcal{O}(nU)$. The main complexity burden comes mainly from the evaluation of the matrix product $(\beta_u^{t-1})^T \mathbf{\Delta} \beta_u^{t-1}$. Hence, both algorithms have a complexity of $\mathcal{O}(n^2U)$ for one iteration. If, as we have observed in [MO35], one trains the online algorithm during $\mathcal{O}(\beta n)$ iterations in order to have good convergence properties, then the global complexity of the online setting becomes $\mathcal{O}(\beta n^3 U)$. Hence, this version of the algorithm, although appealing from many points of view, becomes rapidly not suited for large datasets.

In order to overcome this numerical difficulty, we proposed in [MO28] a reformulation of the relational online algorithm, which leads to a drop in the complexity from cubic to quadratic, making it faster and comparable in terms of computational burden with the numerical version. Briefly, the assignment and representation steps in Algorithm 4 are modified as follows:

1. The assignment step is rewritten as

$$f^t(x_i) \leftarrow \arg \min_{u=1,\dots,U} \left(B_{u,i}^{t-1} - \frac{1}{2} A_u^{t-1} \right) \ , \tag{4.14}$$

where $B^{t-1} = \left(\sum_{j=1}^n \beta_{u,j}^{t-1} \delta_{i'j}\right)_{u=1,\dots,U;i'=1,\dots,n}$ is a $U \times n$ matrix, and $A^{t-1} = \left(\sum_{j,j'=1}^n \beta_{u,j}^{t-1} \beta_{u,j'}^{t-1} \delta_{jj'}\right)_{u=1,\dots,U}$ is a U vector.

2. The updates of A^t and B^t are performed recursively in the representation step. First, we rewrite the update of the prototypes as

$$\beta_u^t \leftarrow (1 - \lambda_u(t))\, \beta_u^{t-1} + \lambda_u(t) \mathbf{1}_i \ , \ \forall u = 1, ..., U, \tag{4.15}$$

where $\lambda_u(t) = \mu(t)H^t(d(f^t(x_i), u))$. This alternative writing leads to simplified updates of A^t and B^t, which do not longer need the use of the entire matrix Δ:

$$B^t_{u,i'} = (1 - \lambda_u(t))\, B^{t-1}_{u,i'} + \lambda_u(t)\delta_{ii'} \ , \ \forall u = 1, ..., U; i' = 1, ...n; \tag{4.16}$$

$$A^t_u = (1 - \lambda_u(t))^2\, A^{t-1}_u + 2\lambda_u(t)\,(1 - \lambda_u(t))\, B^{t-1}_{u,i} \ , \ \forall u = 1, ..., U. \tag{4.17}$$

We proved in [MO28] that, with this alternative writing, the complexity of one iteration of the relational online SOM is $\mathcal{O}(nU)$, and if one performs $\mathcal{O}(\beta n)$ iterations in order to guarantee good empirical convergence properties, then the total complexity drops to $\mathcal{O}(\beta n^2 U)$. In the same paper, we illustrated the significant improvement in terms of computational time on several examples, where the accelerated version is 30 to 40 times faster then the "naïve" one, while being also comparable from this point of view with numerical SOM.

4.3.4 Using SOM(brero) for clustering and visualizing graphs

I chose to illustrate relational SOM in the particular case of graph clustering, which is a topic intensively studied these last years in humanities and social sciences. The example below was designed for static and undirected graphs, but I will give some hints in the conclusion on how this could be extended to more general situations. This section is greatly inspired from our paper [MO7], where we also describe with details the R package that we implemented, SOMbrero, and which is available on CRAN.

With a relational SOM trained on a kernel or dissimilarity associated to a graph, one may draw a simplified representation of it. In this approach, the vertices of the simplified graph represent the clusters, each cluster being represented by a disk which area is proportional to the cluster size, *i.e.* to the number of vertices of the original graph that are associated to the cluster. The edges between the clusters have widths which are usually proportional to the number of edges (or the sum of the weights of the edges) between the pairs of vertices clustered in the two corresponding clusters. In general, when the clustering is a prior step to the visualization, the clustered graph is drawn using modified force-directed placement algorithm [86], which can cope with vertices having non-uniform sizes [87, 88, 89]. In our case, since a map is directly associated to the clustering procedure, this second step is no longer necessary since the prior structure of the grid provides natural positions for the clusters: as shown in [90], standard grids position units in $\mathbb{R}^2$ at coordinates $(1, 1)$, $(1, 2)$, ..., (p, q) where p and q are the length and width of the grid (and thus $pq = U$): the induced graph is thus displayed with vertices positioned at these coordinates. [91] uses a similar approach to represent graphs, by defining a topographic map algorithm. This method does not rely on a dissimilarity but on a criterion specific to graphs and derived from the modularity [92], which is a very standard quality measure for clustering the vertices of a graph.

Dissimilarities for graphs Before presenting the results, I will briefly describe some standard dissimilarities and kernels used for measuring the dissemblance between vertices in a graph. The graph is supposed here as static, undirected, and connected.

One of the most standard dissimilarities is the length of the shortest path between two vertices, expressed as the number of hops. This measure however does not take into account the number of paths that link two vertices.

Another common dissimilarity comes from the eigenvalue decomposition of the Laplacian, and the so called *spectral clustering*, [93]. The Laplacian of the graph, $L = (l_{ij})_{i,j=1,...,n}$, is a matrix encoding the graph structure

$$l_{ij} = \begin{cases} -w_{ij} & \text{if } i \neq j, \\ d_i = \sum_{k \neq i} w_{ij} & \text{otherwise,} \end{cases}$$

where $W = (w_{ij})_{i,j=1,...,n}$ is the (possibly weighted) adjacency matrix. This method first performs an eigenvalue decomposition of the Laplacian, $((\lambda_i)_{i=1,...,n}, (v_i)_{i=1,...,n})$ with $\lambda_1 = 0 \leq \lambda_2 \leq ... \leq \lambda_n$ the eigenvalues of the Laplacian (in increasing order: the first eigenvalue is always equal to zero) and $v_i \in \mathbb{R}^n$. Second, every vertex x_i is represented by the entries of the eigenvectors associated to the U smallest positive eigenvalues of the Laplacian, where U is the number of clusters that are searched for: thus, each vertex is transformed into a vector in $\mathbb{R}^U$, $\mathbf{v}_i = (v_{1i}, \ldots, v_{Ui})$, where v_{ji} is the *i-th* coefficient of the *j-th* eigenvector. This method can be seen as a relaxed version of the normalized cut problem in which the number of edges of the original graph between vertices that belong to two different clusters is minimized while ensuring that the cluster size is large enough. Applied to relational SOM, this gives the idea to use the (squared)-Euclidean distance between $\mathbf{v}_i$ and $\mathbf{v}_{i'}$ to measure the dissimilarity between vertices x_i and $x_{i'}$.

This method can be further refined, using standard kernels for graphs, such as the ones proposed and discussed in [94]. These are based on regularized versions of the Laplacian and include, among others, the *commute time*

kernel, $K_{\mathrm{CT}} = L^+$ [95], that can be interpreted as the average time a random walk takes to connect two vertices in the graph or the *heat kernel*, $K_{\mathrm{H}} = e^{-\gamma L}$, $\gamma > 0$ [96], can be interpreted in term of a diffusion process on the graph.

Eventually, dissimilarities between vertices could also be computed using the *modularity matrix*. For a partition $\mathcal{C}_1, \ldots, \mathcal{C}_U$ of the vertices in a graph, the modularity is equal to

$$\mathcal{Q}(\mathcal{C}_1, ..., \mathcal{C}_U) = \frac{1}{2m} \sum_{k=1}^{U} \sum_{x_i,\, x_j \in \mathcal{C}_k} \left(W_{ij} - \frac{d_i d_j}{2m} \right), \tag{4.18}$$

with m the total number of edges of the graph (so that $\sum_{ij} D_{ij} = \sum_{ij} W_{ij} = 2m$). Finding a partition that maximizes the modularity is thus equivalent to find a partition in which the observed intra-cluster weights, W_{ij} for x_i and x_j in the same cluster, are larger than those expected in a null model where the weights of the edges would depend only on the degrees of the afferent vertices. In [97], the modularity maximization problem is addressed using the eigen-decomposition of a matrix called the *modularity matrix* which can be expressed as

$$B = W - D$$

where D is the matrix $D_{ij} = \frac{d_i d_j}{2m}$. Similarly to the spectral clustering, this eigen-decomposition can be used as a mean to compute a dissimilarity measure between vertices in the graph: as suggested in [97], the eigenvectors of B associated with the positive eigenvalues, $b_1, \ldots, b_p$ (with $p \leq n$), make it possible to provide the following p-dimensional representation for the vertex x_i: $\mathbf{b}_i = (b_{1i}, \ldots, b_{pn})$ and the squared Euclidean distance between those vectors defines a dissimilarity matrix for the graph.

These different dissimilarities are illustrated and compared in the following.

Results for *Les Misérables* The data has been described in [98] and comes from the novel "Les Misérables". The connected graph extracted from this novel is a graph of co-appearance of the characters of the novel: the vertices of the graph are the 77 characters of the novel and the 254 edges stand for a co-appearance of the corresponding two characters in a same chapter of the novel. The edges are weighted by the number of co-appearances. The graph can be downloaded at `http://people.sc.fsu.edu/~jburkardt/datasets/sgb/jean.dat` and is also available as a `igraph` object in the R package `SOMbrero`.

A 5×5 map was trained on the dissimilarity matrix computed as the length of the shortest paths on the graph, with $T = 500$ iterations and prototypes initialized at random among the input data. The resulting clustering has a topographic error equal to 0 (which means the quality of the mapping is very good), and a quantization error equal to 0.61. The map is displayed in Figures 4.1 and 4.2. Among the possible graphs implemented in `SOMbrero`, we show the hitmap, which is the default plot and displays a rectangle which area is proportional to the number of vertices clustered in every unit. We also show the projected graph, as described above, and the clusters composition in terms of vertices (character names).

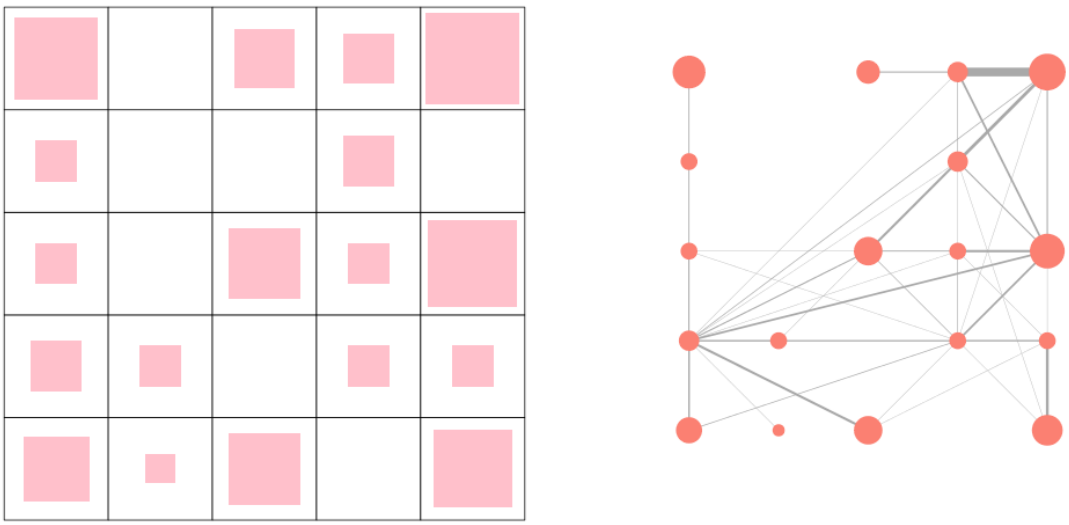


Figure 4.1: Hitmap and projected graph

Figure 4.2: Cluster composition

At this step of the analysis, due to the large number of clusters, the representation of the projected graph is still a bit messy and the clustering quality is not very good (the modularity is equal to 0.389). This is due to the fact that SOM is being used mainly as a nonlinear mapping technique for visualizing data and also as a dimensionality reduction (in terms of number of inputs) rather than a clustering one with a small number of clusters, as one would do in a model-based framework, for instance. In order to overcome this, an additional hierarchical clustering (HAC) step could be added. This HAC step is directly applied on the prototypes, and more specifically one applies the Ward criterion on the dissimilarity matrix computed for the prototypes only. The results are illustrated in Figure 4.3. In particular, figure (c) shows that the clusters obtained make sense (each one is associated to an important character of the novel : Valjean, Myriel, Fantine, Gavroche, Cosette, Javert and the Thenardier family) and the modularity is improved (0.529). In figure (d), we used the super-clustering to obtain a final simplified representation of the graph in which all super-clusters were positioned at their gravity center on the grid.

Influence of the dissimilarity measure Let us now investigate the influence of the dissimilarity on the SOM performances. To do so, we have used different dissimilarities with the graph of co-appearance from "Les Misérables". The following dissimilarities were computed on this graph:

(a) the shortest path lengths, computed with the unweighted graph,

(b) the dissimilarity based on the weighted Laplacian eigen-decomposition with $U = 25$ (to be used with a 5×5 map),

(c) the squared distance induced from the commute time kernel which uses also the weighted Laplacian,

(d) the dissimilarity based on the modularity matrix eigen-decomposition, which also uses the weighted graphs.

The different dissimilarities are illustrated on the heatmaps of Figure 4.4. This figure shows that, as expected,

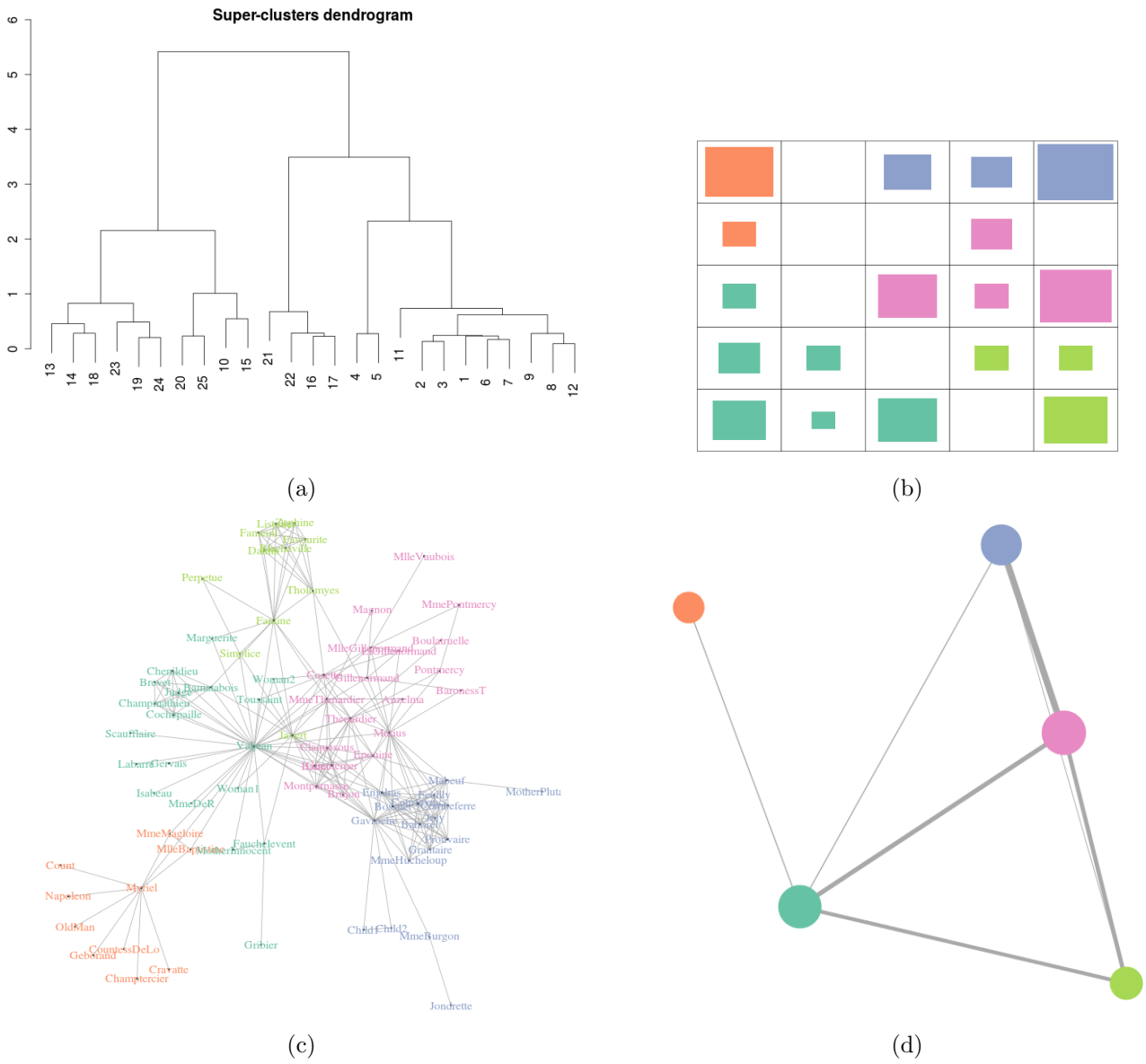


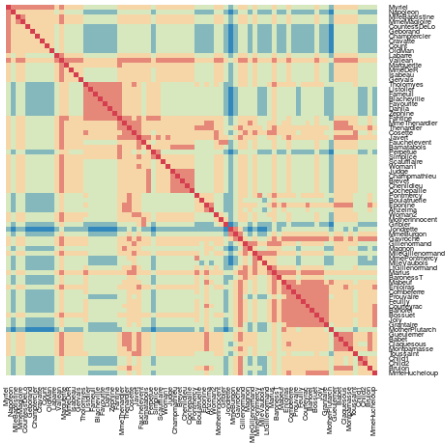
Figure 4.3: (a) Dendrogram of the clustering of the prototypes: from this figure, we chose to select 6 super-clusters, (b) hitmap colored with the result of the super-clustering, (c) original graph with the result of the super-clustering, (d) projected graph using the super-clustering

the dissimilarity based on the eigen-decomposition of the Laplacian and the one computed from the commute time kernel present very similar patterns, the latter being slightly less contrasted than the former which uses only a part of the eigen-decomposition. The shortest path length (which is based on the unweighted graph, contrary to the other three) exibits rather different patterns and seems to have a larger number of distant vertices than the two previous. Finally, the dissimilarity based on the modularity matrix has exactly the opposite behaviour: very few pairs of vertices are considered as distant.

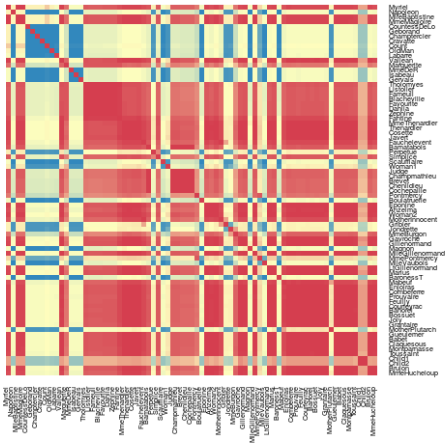
For every dissimilarity matrix, three different initializations of the SOM algorithm were tested: random, random among the input data, on a regular grid in the first two PCA axes. Then, for every dissimilarity and every type of initialization, 1,000 maps were trained with a 5×5-grid and 500 iterations. To test if a combination of dissimilaritiescorresponding to different features can improve the results, two combined dissimilarities were also computed as suggested in [MO7]:

$$\boldsymbol{\Delta}^{\text{sp+modularity}} = \frac{\boldsymbol{\Delta}^{\text{sp}}}{\|\boldsymbol{\Delta}^{\text{sp}}\|_F} + \frac{\boldsymbol{\Delta}^{\text{modularity}}}{\|\boldsymbol{\Delta}^{\text{modularity}}\|_F} \quad \text{and} \quad \boldsymbol{\Delta}^{\text{sp+spec. clust}} = \frac{\boldsymbol{\Delta}^{\text{sp}}}{\|\boldsymbol{\Delta}^{\text{sp}}\|_F} + \frac{\boldsymbol{\Delta}^{\text{spec. clust}}}{\|\boldsymbol{\Delta}^{\text{spec. clust}}\|_F} \tag{4.19}$$

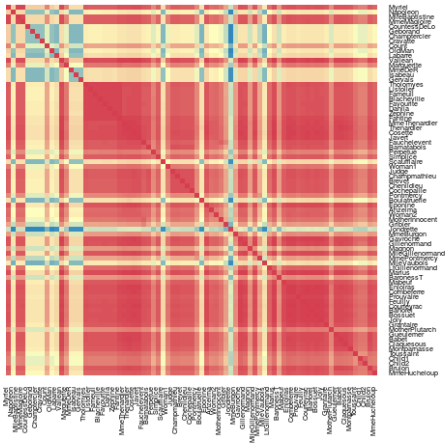
with $\boldsymbol{\Delta}^{\text{sp}}$ the dissimilarity matrix based on shortest path length, $\boldsymbol{\Delta}^{\text{modularity}}$ the dissimilarity matrix based on the eigen-decomposition of the modularity matrix, $\boldsymbol{\Delta}^{\text{spec. clust}}$ the dissimilarity matrix based on the eigen-decomposition of the Laplacian and $\|.\|_F$ the Frobenius norm to make the two combined dissimilarities have



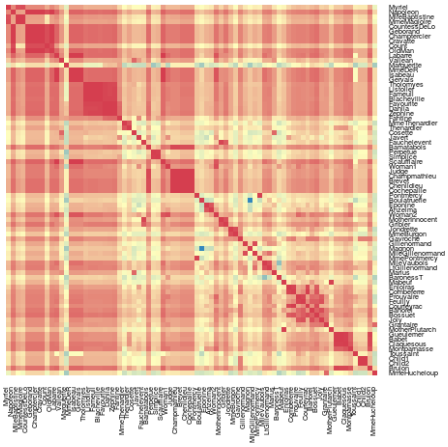
(a) shortest path lengths



(b) Laplacian



(c) commute time kernel distance



(d) modularity

Figure 4.4: Heatmap of the different dissimilarities obtained from the graph "Les Misérables". Red corresponds to small (or null) dissimilarities, blue to large ones.

comparable scales.

The different results were compared using four quality measures:

1. the topographic error, which provides an insight of the good organization of the mapping;
2. the quantization error, which helps quantifying the quality of the clustering. This quantity was divided by age Frobenius norm of the corresponding dissimilarity matrix to allow for similar and comparable scales between the different dissimilarities;
3. the modularity as defined in Equation 4.18. Two versions of the criterion were computed: one using the weighted graphs and one the unweighted graph.

The distribution of the different quality measures over the $1,000$ maps are given by dissimilarity and type of initialization in Figure 4.5. Additionally, Table 4.1 gives the average performance for the four quality measures whatever the initialization.

Several conclusions can be drawn from these results: first, the type of initialization does not seem to have a strong impact on the quality of the results, which is expected for a stochastic algorithm. Second, the maps obtained with the shortest path length are the best for the standard quality criteria for SOMs but are not good

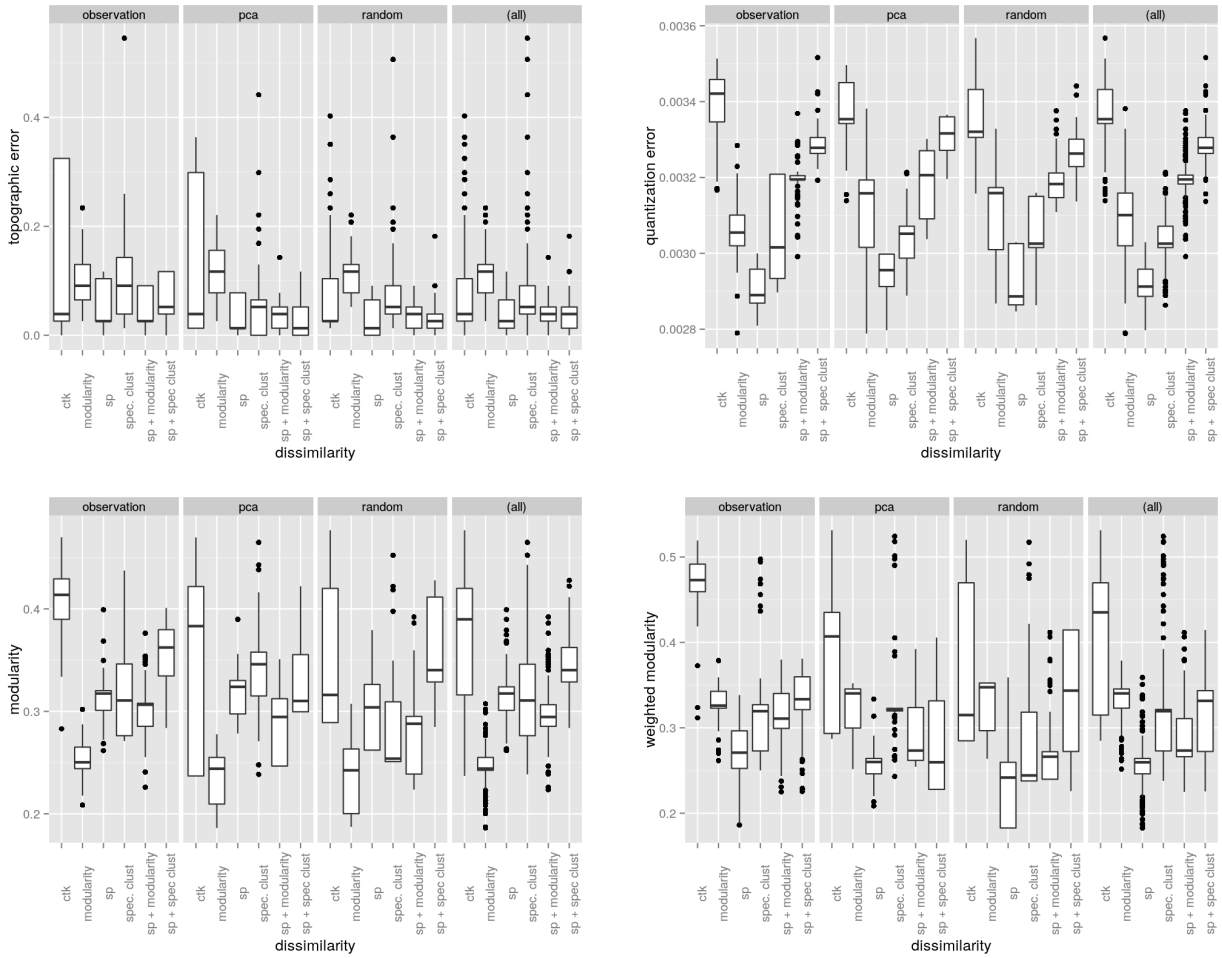


Figure 4.5: Distribution of the topographic and quantization errors and of the modularity (based on the weighted or unweighted) graph, given by dissimilarity and type of initialization. "ctk" means "commute time kernel", "sp" means "shortest path length" and "spec. clust" stands for the dissimilarity based on the eigen-decomposition of the Laplacian. "observations" means random initialization so that, for every u, $\beta_{u,i(u)} = 1$ for a unique $i(u)$, "pca" is the initialization based on a PCA-like approach and "random" is the initialization in which $\beta_{u,i}$ are sampled in $[0, 1]$ and then scaled such that $\sum_i \beta_{u,i} = 1$. The last panel of each figure provides the distribution whatever the type of initialization (distribution over $3,000$ maps).

dissimilarity	topographic error	quantization error	modularity	weighted modularity
ctk	0.100	0.00338	**0.369**	**0.406**
modularity	0.106	0.00310	0.242	0.330
sp	**0.037**	**0.00293**	0.309	0.253
spec. clust	0.065	0.00305	0.308	0.301
sp + modularity	0.038	0.00319	0.287	0.288
sp + spec. clust	0.039	0.00329	0.347	0.318

Table 4.1: Average performance for four quality criterion over all $3,000$ maps produced for each dissimilarity. "ctk" means "commute time kernel", "modularity" is the dissimilarity based on the modularity matrix eigen-decomposition, "sp" means "shortest path length" and "spec. clust" stands for the dissimilarity based on the eigen-decomposition of the Laplacian. The "+" indicates that the two dissimilarities have been combined as in Equation (4.19).

from the modularity perspective (especially the weighted version, which is expected since this dissimilarity does not use the information on weights). On the contrary, the maps based on the dissimilarity computed from the commute time kernel exhibit rather poor topological preservation and quantization error as compared to the other dissimilarities but they have the best performances for the modularity (at the cost of a very large variability). This means that, since they depend on the dissimilarity itself which represents more or less faithfully the graph topology, the standard quality criteria for SOMs are probably not to be trusted too much to compare

maps obtained from different dissimilarities. Surprisingly, the dissimilarity which is based on the modularity matrix did not outperform the others for the modularity criteria. This might be due to the fact that the matrix only uses the positive part of the modularity matrix spectrum, which may well be insufficient to grab most of the information needed to optimize the modularity itself[1]. Also, the dissimilarity based on the modularity matrix had a very poor topographic preservation (which could be the consequence of uninterpretable values, as described above). Finally, the combined dissimilarities provide almost everywhere average performance, except for the quantization error which is among the worst (but this error is probably the most dependent from the values of the dissimilarity themselves and maybe not as reliable as the others).

4.4 Multiple relational SOM

One of the challenges stemming from the analysis of complex data, and particularly complex data from humanities and social sciences, is the fact that more than one kernel or dissimilarity functions may be useful for describing the data. Indeed, in some cases, data may come from various sources: a graph structure together with additional information on the vertices, which may be numerical variables associated with categorical ones, texts, or time series. This situation, called *multiple view*, is quite common in a variety of applications. We may cite, for instance, node clustering in a social network taking into account a set of attributes describing the nodes, [99, 100]. Another issue which is quite similar in terms of methodological approach at least, is data which may be described by several similarity or dissimilarity measures, each encoding for specific features of the data, but none of them being acknowledged as more informative than the others. In social sciences for instance, the choice of a good dissimilarity to describe the resemblance between two categorical time series is still an open issue [101, 102].

This section describes our contribution to the unsupervised analysis of complex data using various sources of information and is summarizing the results published in [MO33], [MO34] and [MO8]. Again, I will only describe the methodology for relational data, since the kernel version may immediately inferred and is available in the cited references.

Combining several sources of information Combining all sources of information or several dissimilarities aims at increasing the relevance of the clustering. In the literature, this issue has already been tackled by different approaches: some rely on clustering ensembles, combining together the clusterings obtained from each view or from each dissimilarity into a consensus clustering [103]. A more complex strategy, described in [104], iteratively updates the different clusterings using a global log-likelihood approach until they converge to a consensus. Other authors propose to concatenate all data/views prior to the clustering. If kernels are available, this method is known as multiple kernel clustering: the different kernels are combined by using a convex combination and the coefficients of the convex combination are optimized together with the clustering [105, 106]. In a similar way, if the data are described by numeric variables belonging to different feature groups, [107] proposes to weight each group and to optimize simultaneously the clustering and the group weights. For the SOM algorithm as well, a few articles tackle related issues: in particular, [108] combines numeric and binary variables to produce a single map by optimizing two quantization energies in parallel. Our approach is using a similar approach by combining different dissimilarities into a convex combination, and is learning an optimal combination in an online fashion, by minimizing the energy function.

Computing a multiple dissimilarity Suppose now that the observations $x_1, \ldots, x_n$ are no longer described by a single dissimilarity matrix $\mathbf{\Delta}$, but by D dissimilarity matrices $\mathbf{\Delta}^1, \ldots, \mathbf{\Delta}^D$, where $\mathbf{\Delta}^d = \left(\delta^d(x_i, x_j)\right)_{i,j=1,\ldots,n}$. The dissimilarities can be either different dissimilarities on the same data or dissimilarities computed from different variables measured on the same individuals.

In relational SOM, as described in the previous section, the entries of each of these dissimilarity matrices, $(\delta^d_{ij})_{i,j=1,\ldots,n}$, are assimilated to a squared-norm $\|x_i - x_j\|^2_d$ in the Euclidean framework. Similarly to the multiple kernel approach described in [109] or in [100] (for multiple kernel SOM), we propose to combine all the similarities into a single one, defined as a convex combination:

$$\delta^\alpha_{ij} = \sum_{d=1}^{D} \alpha_d \delta^d_{ij} \tag{4.20}$$

[1]Note that [91] has already observed that using the eigen-decomposition of the modularity matrix is not the best method for defining a kernel which should optimize the modularity.

where $\alpha_d \geq 0$ and $\sum_{d=1}^{D} \alpha_d = 1$. In the Euclidean framework, this approach is strictly equivalent to the kernel SOM approach because $\|x_i - x_j\|_d^2 = \langle x_i - x_j, x_i - x_j \rangle_d$ (multiple kernel is a convex combination of dot products whereas Equation 4.20 is based on a convex combination of squared distances). In practice, to ensure similar scales for all dissimilarity matrices, a normalization step is performed on the similarity matrices computed according to Equation 4.11 for each $\mathbf{\Delta}^d$, $d = 1, ..., D$: they are all scaled to have a unit Frobenius norm (after having removed the negative eigenvalues, if any).

Online multiple relational SOM If the (α_d) are given, relational SOM based on the dissimilarity introduced in Equation 4.20 aims at minimizing (over $(\beta_u)_u$) the following energy function

$$\mathcal{E}((\beta_u)_u, (\alpha_d)_d) = \sum_{u=1}^{U} \sum_{i=1}^{n} H\left(d\left(f(x_i), u\right)\right) \delta^{\alpha}\left(x_i, p_u(\beta_u)\right),$$

where $\delta^{\alpha}\left(x_i, p_u(\beta_u)\right)$ is defined as in Equation 4.13 by

$$\delta^{\alpha}\left(x_i, p_u(\beta_u)\right) \equiv \mathbf{\Delta}_i^{\alpha} \beta_u - \frac{1}{2} \beta_u^T \mathbf{\Delta}^{\alpha} \beta_u \tag{4.21}$$

with $\mathbf{\Delta}^{\alpha} = \sum_d \alpha_d \mathbf{\Delta}^d$

When there is no a-priori on the $(\alpha_d)_d$, we propose to include the optimization of the convex combination into the online algorithm that trains the map. This idea is similar to the one proposed in [110] for optimizing a kernel parameter in vector quantization algorithms. More precisely, a stochastic gradient descent step is added to the original online relational SOM algorithm to optimize the energy $\mathcal{E}((\beta_{u,i})_{u,i}, (\alpha_d)_d)$ both over $(\beta_{u,i})_{u,i}$ and $(\alpha_d)_d$. To perform the stochastic gradient descent step on the (α_d), the computation of the derivative of

$$\mathcal{E}|_{x_i} = \sum_{u=1}^{U} H\left(d\left(f(x_i), u\right)\right) \delta^{\alpha}\left(x_i, p_u(\beta_u)\right)$$

(the contribution of the randomly chosen observation $(x_i)_i$ to the energy) with respect to α is needed. Since

$$\frac{\partial}{\partial \alpha_d} \left[\delta^{\alpha}(x_i, p_u)\right] = \delta^d(x_i, p_u),$$

we have

$$\mathcal{D}_{id} = \frac{\partial \mathcal{E}|_{x_i}}{\partial \alpha_d} = \sum_{u=1}^{U} H\left(d\left(f(x_i), u\right)\right) \left(\mathbf{\Delta}_i^d \beta_u - \frac{1}{2} \beta_u^T \mathbf{\Delta}^d \beta_u\right).$$

Following an idea similar to that of [109], the SOM is trained by performing, alternatively, the standard steps of the SOM algorithm (i.e., assignment and representation steps) and a gradient descent step for the $(\alpha_i)_i$. The methodology is described in Algorithm 5.

To ensure that the gradient step respects the constraints on α ($\alpha_d \geq 0$ and $\sum_d \alpha_d = 1$), the following strategy is used: first, similarly to [111, 112, 109], the gradient $\left(\frac{\partial \mathcal{E}^{t-1}|_{x_i}}{\partial \alpha_d}\right)_d$ is reduced and is projected such that the non-negativity of α is ensured. The following modified descent step is thus used:

$$\widetilde{\mathcal{D}}_d = \begin{cases} 0 & \text{if } \alpha_d = 0 \text{ and } \mathcal{D}_d - \mathcal{D}_{d_0} > 0 \\ -\mathcal{D}_d + \mathcal{D}_{d_0} & \text{if } \alpha_d > 0 \text{ and } d \neq d_0 \\ \sum_{d \neq d_0,\ \alpha_d > 0} \left(\mathcal{D}_d - \mathcal{D}_{d_0}\right) & \text{else} \end{cases}$$

The descent step $\nu(t)$ is decreased with the standard rate of ν_0/t with an initial ν_0 small enough to ensure the positivity constraint on $(\alpha_d)_d$.

An application to social sciences — which dissimilarity is to be used when extracting typologies in sequence analysis? We illustrate this algorithm on data related to school-to-work transitions. We used the data in the survey "Generation 98"[2]. According to the French National Institute of Statistics (INSEE), 22.7% of young people under 25 were unemployed at the end of the first semester 2012.[3] Hence, it is crucial to understand how the transition from school to employment or unemployment is achieved, in the current economic context. The data set contains information on 16 040 young people having graduated in 1998 and monitored

[2]Available thanks to Génération 1998 á 7 ans - 2005, [producer] CEREQ, [diffusion] Centre Maurice Halbwachs (CMH)

[3]The graphical illustrations were carried out using the `TraMineR` package [113].

Algorithm 5 On-line multiple dissimilarity SOM

For all $u = 1, \ldots, U$ and $i = 1, \ldots, n$, initialize $\beta_{u,i}^0$ such that $\beta_{u,i}^0 \geq 0$ and $\sum_{i=1}^n \beta_{u,i}^0 = 1$.

For all $d = 1, \ldots, D$, initialize $\alpha_d^0 \in [0,1]$ st $\sum_d \alpha_d^0 = 1$. **return** $\delta^{\alpha,0} \leftarrow \sum_d \alpha_d^0 \delta^d$.

for t=1,...,T **do**

Randomly choose an input x_i

Assignment step: find the unit of the closest prototype

$$f^t(x_i) \leftarrow \arg \min_{u=1,\ldots,M} \delta^{\alpha,t-1}\left(x_i, p_u(\beta_u)\right)$$

where $\delta^{\alpha,t-1}\left(x_i, p_u(\beta_u)\right)$ is defined as in Equation (4.21).

Representation step: update all the prototypes according to the new classification: $\forall\, u = 1, \ldots, U$,

$$\beta_u^t \leftarrow \beta_u^{t-1} + \mu(t) H^t\left(d\left(f(x_i), u\right)\right)\left(\mathbf{1}_i - \beta_u^{t-1}\right)$$

Gradient descent step: update the convex combination parameters: $\forall\, d = 1, \ldots, D$,

$$\alpha_d^t \leftarrow \alpha_d^{t-1} + \nu(t)\mathcal{D}_d^t$$

where $\mathcal{D}_d^t$ is the descent direction and update $\delta^{\alpha,t}$

$$\delta^{\alpha,t} \leftarrow \sum_d \alpha_d^t \delta^d.$$

end for

during 94 months after having left school. The labor-market statuses have nine categories, labeled as follows: permanent-labor contract, fixed-term contract, apprenticeship contract, public temporary-labor contract, on-call contract, unemployed, inactive, military service, education. The following stylized facts are highlighted by a first descriptive analysis of the data as shown in Figure 4.6:

- permanent-labor contracts represent more than 20% of all statuses after one year and their ratio continues to increase until 50% after three years and almost 75% after seven years;
- the ratio of fixed-terms contracts is more than 20% after one year on the labor market, but it is decreasing to 15% after three years and then seems to converge to 8%;
- almost 30% of the young graduates are unemployed after one year. This ratio is decreasing and becomes constant, 10%, after the fourth year.

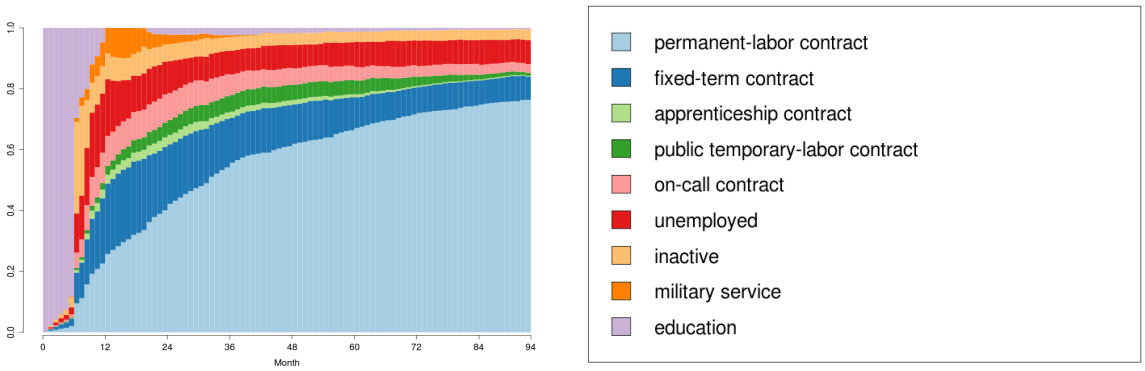


Figure 4.6: Labor market structure

The dissimilarities between sequences were computed using optimal matching (OM). Also known as "edit distance" or "Levenshtein distance", optimal matching was first introduced in biology by [80] and used for aligning and comparing sequences. In social sciences, the first applications are due to [81]. The underlying idea of optimal matching is to transform one sequence into another using three possible operations: insertion, deletion and substitution. A cost is associated to each of the three operations. The dissimilarity between sequences is then computed as the cost associated to the smallest number of operations which allows to transform the sequences into each other. The method seems simple and relatively intuitive, but the choice of the costs is a delicate operation in social sciences. This topic is subject to lively debates in the literature [101, 102] mostly

Metric	OM	HAM	DHD
α-Mean	0.43111	0.28459	0.28429
α-Std	0.02912	0.01464	0.01523

Metric	OM	HAM	DHD	Optimally-tuned α
Quantization error	92.93672	121.67305	121.05520	114.84431
Topographic error	0.07390	0.08806	0.08124	0.05268

Table 4.2: Preliminary results for three OM metrics (average over 50 random subsamples): Optimally-tuned α (top table) and Quality criteria for the SOM clustering (bottom table).

because of the difficulties to establish an explicit and sound theoretical frame.

In our application, all career paths have the same length, the status of the graduate students being observed during 94 months. Hence, we suppose that there are no insertions or deletions and that only the substitution costs have to be defined for OM metrics. Among optimal-matching dissimilarities, we selected three dissimilarities: the OM with substitution costs computed from the transition matrix between statuses as proposed in [114], the Hamming dissimilarity (HAM, no insertion or deletion costs and a substitution cost equal to 1) and the Dynamic Hamming dissimilarity (DHD as described in [115]).

In order to identify the role of the different dissimilarities in extracting typologies, we considered several samples drawn at random from the data. For each of the experiments below, 50 samples containing 1 000 input sequences each were considered. In order to assess the quality of the maps, two indexes were computed: the quantization error for quantifying the quality of the clustering and the topographic error for quantifying the quality of the mapping, [116]. These quality criteria all depend on the dissimilarities used to train the map but the results are made comparable by using normalized dissimilarities.

The results are listed in Table 4.2. According to the mean values of the α's, the three dissimilarities contributed to extracting typologies. The Hamming and the dynamical Hamming dissimilarities have similar weights, while the OM with cost-matrix defined from the transition matrix has the largest weight. The mean quantization error computed on the maps trained with the three dissimilarities optimally combined is larger than the quantization error computed on the map trained with the OM metric only. On the other hand, the topographic error is improved in the mixed case. In this case, the joint use of the three dissimilarities provides a trade-off between the quality of the clustering and the quality of the mapping. The averaged results over the 50 samples are given in Table 4.2. They confirm the difficulty to define adequate costs in optimal matching and the fact that the metric has to be chosen according to the aim of the study: building typologies (clustering) or visualizing data (mapping).

Finally, a multiple relational SOM was trained on the entire data set with the optimal convex combination of dissimilarities. The final map is illustrated in Figure 4.7. Several typologies emerge from the map: a fast access to permanent contracts (clear blue), a transition through fixed-term contracts before obtaining stable ones (dark and then clear blue), a holding on precarious jobs (dark blue), a public temporary contract (dark green) or an on-call (pink) contract ending at the end by a stable one, a long period of inactivity (yellow) or unemployment (red) with a gradual return to employment. The mapping also shows a progressive transition between trajectories of exclusion on the west and quick integration on the east. A more detailed study of this data set is available in [MO34].

4.5 Efficient versions for large data sets

Despite their ability of handling complex data, kernel and relational SOM are not well suited for large datasets. Indeed, they suffer of two important drawbacks stemming from the large dimensionality of the embedding space (which is equal to the number of observations, n). First, as pointed out in [MO34], the complexity (in n) is at least quadratic, and the algorithms will be very slow, with prohibitive computational times. Second, as emphasized in [117], since the prototypes are written as convex combinations in the original dataset, they are no longer explicit representative points, and the interpretability is lost. The prototypes in this case are not much more informative than the clustering itself.

Over the years, different strategies have been developed to handle large datasets (in terms of number of inputs), and are available in the literature. The standard approaches include (i) *divide and conquer* approaches ([118], [119], [120]) in which data are split into several bits of data which are processed separately. The results are aggregated afterwards to obtain a final solution which is supposed to well approximate the solution that would

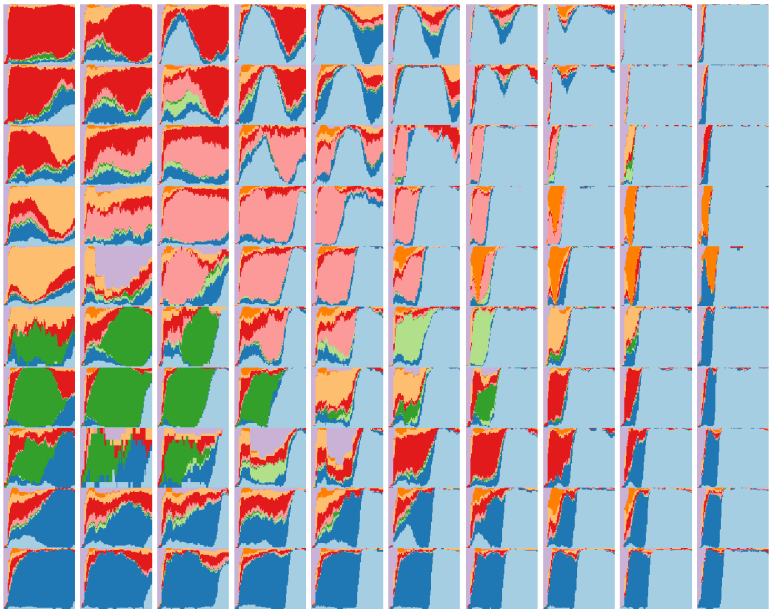


Figure 4.7: Final map obtained with the OM dissimilarities

have been obtained if the entire dataset had been processed at once; (ii) *subsampling methods*([121], [122], [123], [124], [125]), which consist in using a restricted subset (usually carefully designed) of the original data, in order to approximate the solution that could have been obtained with the entire dataset and (iii) *online updates* ([126], [127]), in which the results are updated with sequential steps, each having a low computational cost.

A particular case of the subsampling strategy is the Nyström approximation [128], which consists in sampling a small number of rows/columns in square matrices and in obtaining an approximation of its eigendecomposition at a very reduced computational cost. The eigendecomposition is even exact when the matrix is of low rank (when the size of the subsample is larger than the rank of the matrix). This method is frequently used for kernel and dissimilarity-based algorithms.

Regarding kernel and relational data, several extensions of the SOM algorithm or of related methods (k-means, topographic mapping, LVQ, ...) have been proposed in the literature, and build on the strategies cited above. Most of them seek a simplified or a sparse representation of the prototypes, and/or a reduced computational time.

In the relational k-means framework, [129] proposed a sparse extension of the batch algorithm: every prototype is represented by at most K (K fixed) observations per cluster, selected at each step of the algorithm. In the supervised framework, [130] used a similar strategy for batch LVQ, by selecting the most representative observations (with different methods to obtain them, including approximation heuristics and L1 penalty) in every cluster and at each step of the algorithm. A similar method was used in [131], combined with the Nyström approximation of the LVQ algorithm, in order to obtain sparse prototypes at very low computational cost. The Nyström approximation was also used for obtaining faster versions of topographic mapping methods [132], [133] and for reducing the computational cost of the clustering.

Nevertheless, these approaches do not lead to a simplified (and thus interpretable) representation of the prototypes. Furthermore, all of them are restricted to the batch framework and most of them are performed after each iteration of a batch algorithm, i.e., after all observations have been processed at least once. An alternative to these methods consists in splitting the data into several subsets on which independent algorithms are trained. [134] use *patch clustering* as particularly suited for streaming data, but also for large dimensional data. The underlying principle is to randomly split the initial data into a partition of patches, $(\mathcal{P}_b)_b$, that the algorithm processes iteratively. At step 1, the first patch is trained until convergence. The resulting prototypes are approximated by the closest P input data points. During each of the next steps indexed by b, the index set of the P-approximations of all prototypes and the index set of the next patch $\mathcal{P}_{b+1}$ are put together into the extended patch $\mathcal{P}^\star_{b+1}$, and the clustering process is performed on all the observations indexed by $\mathcal{P}^\star_{b+1}$. This is iterated until all patches have been clustered. This approach leads to good clustering results, but it is not parallelizable and the algorithm may be sensitive to the order in which patches are processed.

Our contributions in this context consisted in three different algorithms, all three aimed at tackling the dimensionality issue. We payed a significant amount of attention to the prototype interpretability question. Our first

contribution was introduced in [MO32] and consists of a bagging approach which only uses a small subset of the original data. Roughly, the method we propose works as a sophisticated sampling technique, by combining the results of several bags to select the most representative observations, which are then utilized to define the prototypes in the final map. This approach is both sparse (since the resulting map is based on a small subset of observations only), fast and parallelizable. The second approach was introduced in [MO6] and takes advantage of a preprocessing step of the data, consisting into embedding it in a low dimensional Euclidean space and then performing standard numerical SOM on the resulting vectors. Because of the computational complexity of the preprocessing, an approximation using the Nyström technique is added. Eventually, an alternative approach to obtain sparse prototypes is to take advantage of the online updates of the stochastic version of the algorithm. Here, unlike sparse relational SOM with Nyström approximation, prototypes are directly written as convex combinations of the images of the input data, but, in this case, they are restricted to the input data already fed to the algorithm, and, more particularly, to the most important of them. We introduced this version for relational data in [MO29], and described its equivalent for kernels in [MO6].

In the following, I will briefly describe these three algorithms and compare their performances with the original one. Since all three are approximation techniques, their clustering performances will be usually lower than those of the relational algorithm trained on the entire data, but we will seek for the best trade-off between clustering abilities on the one hand, and computational complexity and prototype interpretability on the other. Eventually, the three algorithms will be illustrated on several case studies.

4.5.1 Bagged relational SOM

The algorithm More formally, the algorithm we introduced consists in sampling at random among the n inputs, B *small* subsets $(\mathcal{S}_b)_b$ of size $n_B \ll n$. On each sample $\mathcal{S}_b$, one trains an online relational SOM with U units, and with the resulting embedded prototypes $\phi(p_u^b) = \sum_{x_i \in \mathcal{S}_b} \beta_{u,i}^b \phi(x_i)$, where ϕ is the *feature map* associated with the dissimilarity δ. Next, for each trained map b and for each prototype u, the inputs with the first P largest weights, where $P \in \mathbb{N}$ is a fixed hyper-parameter, are chosen as the most representative observations,

$$\mathcal{I}_u^b := \{x_i : \beta_{u,i}^b \text{ is among the first } P \text{ largest weights } (\beta_{u,j}^b)_{x_j \in \mathcal{S}_b}\} , \tag{4.22}$$

and one may define the set of the most relevant observations for the bth sample and the bth map as $\mathcal{I}^b = \cup_u \mathcal{I}_u^b$. The sets $(\mathcal{I}^b)_b$ of the most representative inputs are eventually merged into

$$\mathcal{S} := \{x_i : \mathcal{N}(x_i) \text{ is among the } PU \text{ largest numbers } (\mathcal{N}(x_j))_{j=1,\cdots,N}\} , \tag{4.23}$$

where $\mathcal{N}(x_i) := \#\{b : x_i \in \mathcal{I}^b\}$ is the frequency with which x_i is selected as relevant for each sample and for each map. Eventually, a final map with U clusters is trained on the inputs selected in $\mathcal{S}$. The final clustering for all inputs is then derived by applying the assignment step of the relational SOM algorithm to each x_i, $i = 1, \cdots, n$. One shall remark here that in the assignment step, the distance between one input x_i and one prototype p_u writes as:

$$\delta(x_i, p_u) = \|\phi(x_i) - p_u\|_{\mathcal{H}}^2 = \beta_u^T \tilde{\mathbf{\Delta}}_j - \frac{1}{2} \beta_u^T \tilde{\mathbf{\Delta}} \beta_u , \tag{4.24}$$

where $\tilde{\mathbf{\Delta}}_j^T = (\delta(x_i, x_j))_{x_j \in \mathcal{S}}$ and $\tilde{\mathbf{\Delta}} = (\delta(x_j, x_{j'}))_{(x_j, x_{j'}) \in \mathcal{S}}$. The proposed methodology is summarized in Algorithm 6.

With the remarks on the complexity in the previous sections, the bagged algorithm trained with the accelerated version of the relational online SOM has a complexity $\mathcal{O}(n_b^2 U)$ for each bag (with each observation being fed at least once in the training procedure), and $\mathcal{O}(P^2 U^3)$ for the final map. When compared with the algorithm trained on the whole data, the computational approach is more interesting as long as $B n_B^2 + U^2 P^2 < n^2$, although the performances may be further improved by training the B initial SOM's in parallel. Usually, B is chosen to be large, n_B is small compared to n, and P is small also, in order to have a sparse representation of the prototypes. We investigated the computational performances and the sensitivity of the results with respect to the hyper-parameters B, n_B and P in [MO32]. According to these, the performances of bagged relational SOM, even for small values of P, are very similar to those of the full relational SOM, and in some cases even better. This suggests that training a clustering procedure, and in particular a relational SOM algorithm, on the most representative inputs chosen from the data, tends to remove the noise and provide more robust clustering structures when compared with what one gets by clustering the entire data set with no preprocessing. Obtained prototypes are also easier to interpret, as based on a smaller number of observations.

An illustration on network data We illustrated the bagged version of the algorithm on one of the ego-Facebook networks described in [135]. We used the network 107, for which we extracted the largest connected

Algorithm 6 Bagged relational SOM

Initialize for all $i = 1, \ldots, n$, $\mathcal{N}(x_i) \leftarrow 0$
for $b = 1 \rightarrow B$ **do**
  Sample randomly n_B observations in $(x_i)_{i=1,\ldots,n}$ **return** $\mathcal{S}_b$
  Perform **relational SOM** with $\mathcal{S}_b$ **return** prototypes $(p_u^b)_u \sim (\beta_{u,i}^b)_{u,i}$
  for $u = 1 \rightarrow U$ **do**
    Select the P largest $(\beta_{u,i}^b)_{x_i \in \mathcal{S}_b}$ and **return** $\mathcal{I}_u^b$ (set of the observations corresponding to the selected $\beta_{u,i}^b$)
  end for
  for $i = 1 \rightarrow n$ **do**
    if $x_i \in \cup_u \mathcal{I}_u^b$ **then**
      $\mathcal{N}(x_i) \leftarrow \mathcal{N}(x_i) + 1$
    end if
  end for
end for
Select the PU observations corresponding to the largest $\mathcal{N}(x_i)$ **return** $\mathcal{S}$
Perform **relational SOM** with $\mathcal{S}$ **return** prototypes $(p_u)_u \sim (\beta_{u,i})_{u=1,\ldots,U,x_i \in \mathcal{S}}$ and classification $(f(x_i))_{x_i \in \mathcal{S}}$
Affect $(x_i)_{x_i \notin \mathcal{S}}$ with

$$f(x_i) := \arg\min_u \|\phi(x_i) - p_u\|_{\mathcal{H}}^2$$

return final classification $(f(x_i))_{i=1,\ldots,n}$ and sparse prototypes $(p_u)_u \sim (\beta_{u,i})_{u=1,\ldots,U,x_i \in \mathcal{S}}$

	Bagged K-SOM	Full K-SOM	Random K-SOM
Quantization error	**7.66**	9.06	8.08
Topographic error	**4.35**	5.22	6.09
Node purity	**89.65**	86.53	87.26
Normalized mutual information	**70.10**	53.79	60.79
Modularity	**0.47**	0.34	0.40

Table 4.3: Quality measures for different versions of kernel SOM (standard using all data, bagged, standard using randomly selected data) on Facebook ego-network.

component, which contained 1,034 nodes. Standard kernel SOM and bagged kernel SOM were trained on 10×10 two-dimensional grids and then compared in terms of performances.

As explained in [90, 91], using such mapping provides a simplified representation of the graph, which may be useful for the user for understanding the macro-structures, before focusing on some selected clusters. The kernel used for computing the similarities between the vertices was the *commute time kernel* described in Section 4.3.4. As shown in [136], the commute time kernel yields fo a simple similarity interpretation because it computes the average time needed for a random walk on the graph to reach a node starting from another one.

We compared three different approaches: (i) the *standard kernel SOM* (on-line version), using all available data; (ii) the *bagged kernel SOM*, with $B = 1000$ bootstrap samples, $n_B = 200$ in each sample and $P = 3$ observations selected per prototype and (iii) a *standard kernel SOM* trained with an equivalent number of randomly chosen observations. The relevance of the results was assessed using different quality measures. Some quality measures were related to the quality of the map (quantification error and topographic error) and some were related to a ground truth: some of the nodes have been indeed labeled by users to belong to one *list* (as named by Facebook). We confronted these groups to the clusters obtained on the map calculating (i) the average node purity and (ii) the normalized mutual information [137] and also to the graph structure using the modularity [138], which is a standard quality measure for node clustering.

The results are summarized in Table 4.5.1. Surprisingly, the maps trained with a reduced number of input data (bagged K-SOM and random K-SOM) obtain better quality measures than the map trained with all the available data. Using a bootstrapping approach to select the relevant observations also significantly improves all quality measures as compared to a random choice with the same number of observations. The results obtain with the bagged SOM are displayed in Figure 4.8. They show that the nodes are mainly dispatched into four big clusters, which correspond each to approximately only one "list", as defined by the user. The results provided with the K-SOM using all the data tend to provide smaller communities and to scatter the biggest lists on the map. Using this approach, it is however hard to conclude if the interpretability has been increased (i.e., if the selected observations used for training are representative of their cluster) as they do not seem to have a

particularly high degree or centrality.

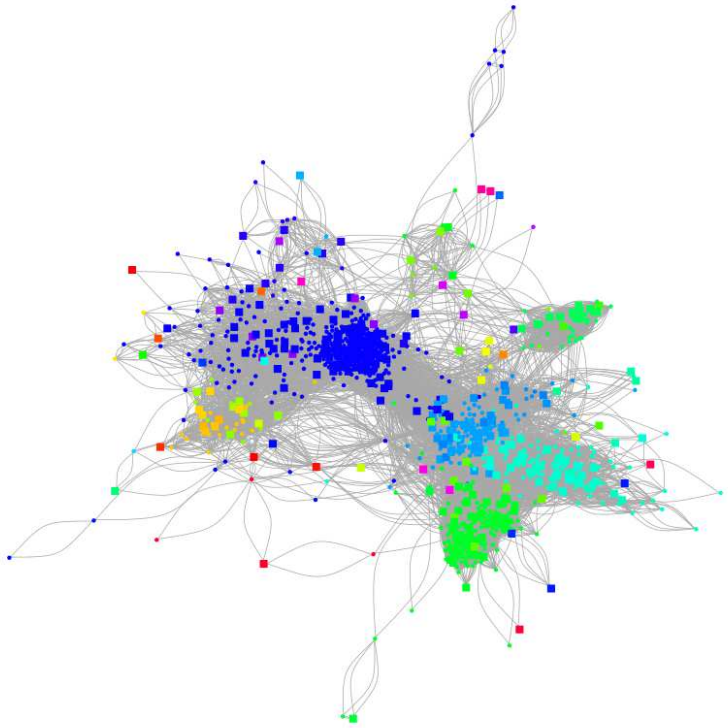


Figure 4.8: The Facebook network represented with a force-directed placement algorithm. Colors represent the clusters on the map and selected nodes used to train the map are represented by squares (instead of circles).

4.5.2 Sparse relational SOM with Nyström approximation

The second algorithm we introduced was inspired by a dimensionality reduction principle. We proposed to rely on a preprocessing of the data based on a PCA (principal component analysis) in the feature space $\mathcal{E}$, and then on training the numerical SOM algorithm on the subspace of $\mathcal{E}$ spanned by the first eigenvectors of the PCA. Eventually, the computational burden was reduced by computing the PCA eigenvectors thanks to a Nyström approximation.

Linear embedding of the dissimilarity matrix The first step consists in a partial Euclidean embedding of the dissimilarity matrix $\mathbf{\Delta}$. One starts by computing a similarity matrix $\boldsymbol{S} = (s_{ij})_{i,j=1,...,n}$ using the double centering formula in Equation 4.11. Next, she computes the eigenvalues $(\lambda_k)_{k=1,...,n}$ (ordered decreasingly) and the corresponding eigenvectors $(\boldsymbol{\alpha}_k)_{k=1,...,n}$ of $\boldsymbol{S}$. This diagonalization is equivalent to finding the eigenvectors in the feature space $\mathcal{E}$ of the covariance matrix of the (centered) images of the original data by ϕ, the feature map associated to $\mathcal{E}$. These vectors, noted $(a_k)_{k=1,...,n} \in \mathcal{E}$, lie in the span of $\{\phi(x_i)\}_{i=1,...,n}$ and may be expressed as

$$a_k = \sum_{i=1}^{n} \alpha_{k,i} \phi(x_i) \ , \ \forall k = 1, ..., n \ . \tag{4.25}$$

a_k are orthonormal in $\mathcal{E}$ and the principal components are the coordinates of the projections of the images of the original data $\{\phi(x_i)\}_{i=1,...,n}$ onto the eigenvectors $(a_k)_{k=1,...,n}$. Eventually, the principal components write as follows:

$$\mathcal{P}_{a_k}(\phi(x_i)) = \langle a_k, \phi(x_i) \rangle_{\mathcal{E}} \cdot a_k = (\lambda_k \alpha_{k,i}) a_k \ , \ \forall k = 1, ..., n; i = 1, ..., n. \tag{4.26}$$

This PCA/MDS embedding may be used to approximate the data in a reduced space by selecting p axes $(a_k)_{k=1,...p}$ (associated with the p largest eigenvalues) in the feature space $\mathcal{E}$, $p \ll n$, on which the data will be projected. Let me remark here that if $\mathbf{\Delta}$ is not Euclidean and $\boldsymbol{S}$ is not positive definite, one will have $n_+ \leq n$ positive eigenvalues and $n_- \leq n$ negative eigenvalues. Then, the dimension p will be usually selected such that $p \ll n_+$. This restriction is equivalent to performing K-PCA on a kernel pre-processed with the standard *clip* approach as suggested in [82].

Relational SOM on linear embeddings The next step of the algorithm is to define the U prototypes of the map in $A = \text{span}\{a_1, ..., a_p\}$ instead of the entire feature space, $\phi(p_u) = \sum_{k=1}^{p} \beta_{u,k} a_k$, where $\beta_{u,k} \geq 0$ and $\sum_{k=1}^{p} \beta_{u,k} = 1$.

The *assignment step* then writes as

$$f^t(x_i) := \arg \min_{u=1,...,U} \|\phi(p_u) - \phi(x_i)\|_A^2 \ ,$$

in which

$$\|\phi(p_u) - \phi(x_i)\|_A^2 = \boldsymbol{\beta}_u^\top \boldsymbol{\beta}_u - 2\boldsymbol{\beta}_u^\top \Lambda \boldsymbol{\alpha}_{.i} + \|\phi(x_i)\|_A^2 \ , \tag{4.27}$$

where $\boldsymbol{\beta}_u = (\beta_{u,k})_{k=1,...,p}$, $\Lambda = \text{Diag}\,(\lambda_1, \ldots, \lambda_p)$ and $\boldsymbol{\alpha}_{.i}$ is the i-th column of $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p]^\top$. This step is thus equivalent to minimizing $\boldsymbol{\beta}_u^\top \boldsymbol{\beta}_u - 2\boldsymbol{\beta}_u^\top \Lambda \boldsymbol{\alpha}_{.i}$ over $u \in \{1, \ldots, U\}$ and is also equivalent to minimizing $\|\phi(p_u) - \mathcal{P}_A(\phi(x_i))\|^2$ over u.

The updates of the prototypes in the *representation step* can be explicitly written as

$$\begin{aligned} \phi(p_u^t) &= \phi(p_u^{t-1}) + \mu(t) H^t(d(f^t(x_i), u)) \left(\mathcal{P}_A(\phi(x_i)) - \phi(p_u^{t-1})\right) \\ &= \sum_{k=1}^{p} \beta_{u,k}^{t-1} a_k + \mu(t) H^t(d(f^t(x_i), u)) \left(\sum_{k=1}^{p} (\lambda_k \alpha_{ki}) a_k - \sum_{k=1}^{p} \beta_{u,k}^{t-1} a_k \right), \end{aligned} \tag{4.28}$$

which is equivalent to update the coefficients as:

$$\boldsymbol{\beta}_u^t = \boldsymbol{\beta}_u^{t-1} + \mu(t) H^t(d(f^t(x_i), u)) \left(\Lambda \boldsymbol{\alpha}_{.i} - \boldsymbol{\beta}_u^{t-1}\right). \tag{4.29}$$

Eventually, this approach is simply the standard (numerical) SOM with entries the $n \times p$ matrix $\boldsymbol{\alpha}^\top \Lambda$. Using this approximation, the complexity is reduced from $\mathcal{O}(nUT)$ (accelerated online relational SOM, T iterations) to $\mathcal{O}(pUT) + \mathcal{O}(npU)$ (online numeric SOM in $\mathbb{R}^p$, cost of the T iterations and cost of the final clustering computation), once the embedding is given. Hence, since T is generally of the order of n, this gives a linear complexity for the algorithm, which makes it very interesting for large dataset. However, the embedding itself can have a large (cubic) complexity. This issue is addressed in the next paragraph.

Linear embedding of relational data with Nyström approximation Although the linear embedding with a dimensionality reduction in the MDS step reduces the relational algorithm to a numerical one, and improves the complexity to linear in the size of the input data, the computational burden is actually shifter in the pre-processing step, since the eigenvalue decomposition of the Gram matrix $\boldsymbol{S}$ has a cubic complexity in n.

A very effective approach for improving the scalability of this step is to use a Nyström approximation [128]. A similar technique had been already used in [133] for improving the computational cost of topographic maps for dissimilarity data. They reduced the computational complexity of $\|\phi(p_u) - \phi(x_i)\|^2$ from $\mathcal{O}(n^2)$ to $\mathcal{O}(m^2 n)$, where $m \ll n$ is a number of inputs randomly sampled within the original data, and m is supposed to be close to the rank of $\boldsymbol{S}$. In our approach, it is the pre-processing step only that is addressed by the Nyström approximation, while the SOM procedure is reduced to a numerical version with a linear complexity.

More precisely, we approximate the eigendecomposition of $\boldsymbol{S}$ by selecting m observations, $\mathcal{T}_m$ among $(x_i)_{i=1,...,n}$, and by using the eigendecomposition of the reduced matrix $\boldsymbol{S}^{(m)} = (s_{ij})_{x_i, x_j \in \mathcal{T}_m}$. In practice, the selected observations $\mathcal{T}_m$ are chosen at random, although more efficient sampling techniques such as the ones described and evaluated in [139] could also be used.

If the eigenvalues and the (orthonormal) vectors of $\boldsymbol{S}^{(m)}$ are denoted by $(\lambda_k^{(m)})_{k=1,...,m}$ and $(v_k^{(m)})_{k=1,...,m}$ respectively, then the eigenvalues and (orthonormal) eigenvectors of $\boldsymbol{S}$ are given by

$$\lambda_k \simeq \frac{n}{m} \lambda_k^{(m)} \qquad \text{and} \qquad v_{k,i} \simeq \sqrt{\frac{m}{n}} \frac{1}{\lambda_k^{(m)}} \boldsymbol{S}_{i.}^{(n,m)} v_k^{(m)} \ , \ \forall k = 1, ..., m \ , \ \forall i = 1, ..., n, \tag{4.30}$$

with $\boldsymbol{S}_{i.}^{(n,m)}$ the i-th row of the matrix $\boldsymbol{S}^{(n,m)} = (s_{ij})_{i=1,...,n,\, x_j \in \mathcal{T}_m}$. If the rank of $\boldsymbol{S}^{(m)}$ is equal to the rank of the original matrix $\boldsymbol{S}$, then the approximation even becomes an equality. Then, assuming that $\boldsymbol{S}$ (which is supposed to be centered) is known or at least that any of the pairs $(s_{ij})_{i,j=1,...,n}$ can be computed at low cost, the linear embedding requires to obtain the entries $(\lambda_k \alpha_{ki})_{k=1,...,p} \in \mathbb{R}^p$ for all $i = 1, \ldots, n$, where $(\boldsymbol{\alpha}_k)_k$ are the eigenvectors of $\boldsymbol{S}$ which are orthogonal with respect to the norm induced by $\boldsymbol{S}$. We can easily show that

$$\boldsymbol{\alpha}_k = \frac{v_k}{\sqrt{\lambda_k}} \ , \ \forall k = 1, ..., p, \tag{4.31}$$

and therefore, the linear embedding may be computed with entries the rows of the $n \times p$ matrix $\boldsymbol{\alpha}^\top \Lambda$ with $\forall\, i = 1, \ldots, n$ and $\forall\, k = 1, \ldots, p$,

$$\lambda_k \alpha_{ki} = \sqrt{\lambda_k} v_{ki} = \frac{1}{\sqrt{\lambda_k^{(m)}}} K_{i.}^{(n,m)} v_k^{(m)} = K_{i.}^{(n,m)} \boldsymbol{\alpha}_k^{(m)} \tag{4.32}$$

with $\boldsymbol{\alpha}_k^{(m)} = \frac{v_k^{(m)}}{\sqrt{\lambda_k^{(m)}}}$. This simplified representation is a good approximation of the linear embedding with $\boldsymbol{S}$. The complexity of the thus preprocessing is reduced from $\mathcal{O}(n^3)$ to $\mathcal{O}(m^3) + \mathcal{O}(nm^2)$. The complete algorithm is provided in Algorithm 7.

Algorithm 7 Online K-PCA SOM

Nyström approximation of K-PCA
Select at random m observations $\mathcal{T}_m = \{x_{i(1)}, \ldots, x_{i(m)}\}$ from the original dataset
Compute the first p eigenvalues and orthonormal eigenvectors of $K^{(m)}$, $(\lambda_k^{(m)})_{k=1,\ldots,p}$, $(v_k^{(m)})_{k=1,\ldots,p}$ and obtain, for $k = 1, \ldots, p$, $\boldsymbol{\alpha}_k^{(m)} = \frac{v_k}{\sqrt{\lambda_k}}$
Compute $\left(K^{(n,m)} \boldsymbol{\alpha}_k^{(m)}\right)_{k=1,\ldots,p}$, which is an $n \times p$ matrix $B = (b_{ik})_{i=1,\ldots,n,\, k=1,\ldots,p}$
K-PCA SOM
Initialize prototypes randomly: $\forall\, u = 1, \ldots, U$, $\boldsymbol{\beta}_u^0 \in [0,1]^p$ and $\sum_{k=1}^p \beta_{uk}^0 = 1$
for $t = 1, \ldots, T$ **do**
    Randomly choose an input $i \in \{1, \ldots, n\}$
    Assignment step: find the unit of the prototype closest to i-th row of B, $\mathbf{b}_i$:

$$f^t(x_i) \leftarrow \arg \min_{u=1,\ldots,U} \|\boldsymbol{\beta}_u^{t-1} - \mathbf{b}_i\|_{\mathbb{R}^p}^2$$

    Representation step: $\forall\, u = 1, \ldots, U$,

$$\boldsymbol{\beta}_u^t \leftarrow \boldsymbol{\beta}_u^{t-1} + \mu(t) H^t(d(f^t(x_i), u)) \left(\mathbf{b}_i - \boldsymbol{\beta}_u^{t-1}\right)$$

end for

4.5.3 Direct sparse SOM

Finally, the third approach we proposed takes advantage of the online updates of the stochastic version of the relational algorithm in order to obtain sparse prototypes. The sparse representation is achieved both through the initialization and a thresholded update.

Prototypes are initialized at random among the input data; thus, the method first selects U inputs at random (or considers the first U observations if the data is acquired "on the fly") and uses them as initial values for the U prototypes. At a given step t of the algorithm, each prototype is written as a convex combination of the images of the most important past observations: if $I_u(t-1)$ denotes the set of the most important observations selected for prototype p_u before step t, p_u writes $\phi(p_u) = \sum_{j \in I_u(t)} \beta_{u,j} \phi(x_i)$. Eventually, the distance in the feature space $\mathcal{E}$ between a new input, x_i selected at random, and p_u is given by:

$$\|\phi(x_i) - \phi(p_u)\|^2 = \sum_{j \in I_u(t-1)} \beta_{u,j} \delta(x_j, x_i) - \frac{1}{2} \sum_{j,j' \in I_u(t-1)} \beta_{u,j} \beta_{u,j'} \delta(x_j, x_{j'}) \;,\; \forall u = 1, ..., U; i = 1, ..., n. \tag{4.33}$$

In order to maintain prototypes as sparse combinations of the input data, they are periodically updated and the most important coefficients only are kept. The update instants may be performed throughout the iterations using various strategies: for instance, they can be uniformly distributed during the learning process or distributed according to some geometric distribution. The parameter of the geometric distribution may be fixed, during the whole training, or varying (in ascending or descending fashion) with the iterations. We showed in [MO6] that results on simulations are globally similar, with a slight advantage for the "ascending random" strategy.

As suggested by [130] for a batch sparse LVQ method, sparsity could also be achieved by selecting the first P most important coefficients, where P is a fixed integer. However, in order to allow for more flexibility in the expression of the prototypes, we choose to select the most important coefficients according to their value, by fixing a threshold: let $0 < \nu \leq 1$ be the selected threshold. At time step t at which an update occurs and for every

$u = 1, ..., U$, the coefficients of the prototype p_u are first ordered in descending order, $\beta_{u,(1)} \geq ... \geq \beta_{u,(\sharp I_u(t))}$. Then, the integer N_u such that

$$N_u = \arg \min_{k=1,...,\sharp I_u(t)} \left\{ \sum_{i=1}^{k} \beta_{u,(i)} \geq \nu \right\} \tag{4.34}$$

is introduced. The most important coefficients are finally updated as follows

$$\beta_{u,(i)} = \begin{cases} \frac{\beta_{u,(i)}}{\sum_{j=1}^{N_u} \beta_{u,(j)}} & \text{if } (i) \leq N_u \\ 0 & \text{if } (i) > N_u \end{cases}, \tag{4.35}$$

and $I_u(t)$ is updated accordingly afterwards by keeping the observations that correspond to non zero coefficients only.

The sparse relational SOM algorithm is entirely described in Algorithm 8.

Algorithm 8 Sparse online K-SOM

For all $u = 1, \ldots, U$, initialize p_u^0 among U randomly selected observations in $(x_i)_i$. Initialize $I_u(0) = \{i(u)\}$, with $i(u) \in \{1, \ldots, n\}$ for all u and $\boldsymbol{\beta}_u^0 = 1$.
for $t = 1, \ldots, T$ **do**
  Randomly choose an input x_i, $i \in \{1, ..., n\}$.
  Assignment step: find the unit with the prototype closest to $\phi(x_i)$:

$$f^t(x_i) \leftarrow \arg \min_{u=1,...,U} \left[(\boldsymbol{\beta}_u^{t-1})^\top \mathbf{K}_{I_u(t-1)} \boldsymbol{\beta}_u^{t-1} - 2 \sum_{j \in I_u(t-1)} \beta_{uj}^{t-1} K(x_j, x_i) \right]$$

  where $\mathbf{K}_{I_u(t-1)} = (K(x_j, x_{j'}))_{j,j' \in I_u(t-1)}$.
  Representation step: $\forall\, u = 1, \ldots, U$
  if $i \in I_u(t-1)$, **then**
    $\boldsymbol{\beta}_u^t \leftarrow \boldsymbol{\beta}_u^{t-1} + \mu(t) H^t(d(f^t(x_i), u)) \left(\mathbf{1}_i - \boldsymbol{\beta}_u^{t-1} \right)$
    $I_u(t) = I_u(t-1)$
  else if $i \notin I_u(t-1)$, **then**

$$\boldsymbol{\beta}_u^t \leftarrow \left[1 - \mu(t) H^t(d(f^t(x_i), u)) \right] \left(\boldsymbol{\beta}_u^{t-1}, 0 \right) + \mu(t) H^t(d(f^t(x_i), u)) (\underbrace{0, ..., 0}_{\sharp I_u(t-1)}, 1)$$

    $I_u(t) = I_u(t-1) \cup \{i\}$.
  end if
  Sparse representation
  if t is an update instant **then**
    Sparsely update the prototypes: $\forall\, u = 1, \ldots, U$ and with $\beta_{u,(1)}^t \geq ... \geq \beta_{u,(\sharp I_u(t))}^t$, set

$$N_{t,u} = \arg \min_{k=1,...,\sharp I_u(t)} \left\{ \sum_{i=1}^{k} \beta_{u,(i)}^t \geq \nu \right\}$$

  and $\forall\, (i)$, st $i \in I_u(t)$,

$$\beta_{u,(i)}^t \leftarrow \begin{cases} \frac{\beta_{u,(i)}^t}{\sum_{j=1}^{N_{t,u}} \beta_{u,(j)}^t} & \text{if } (i) \leq N_{t,u} \\ 0 & \text{if } (i) > N_{t,u} \end{cases}$$

  and $I_u(t) \leftarrow \{i : (i) \leq N_{t,u}\}$.
  end if
end for

Contrary to Nyström approximated SOM, in this version the prototypes may directly be interpreted through the observations that are used in their representation. Also, the sparsity is updated during the training and the induced dimension reduction is thus not constrained to the efficiency of a given dimension reduction technique such as PCA. However, due to the sparse representation step, the algorithm can be computationally more expensive than Nyström approximated SOM and the amount of information preserved in the sparse representation is not as well controlled as in the PCA mapping. Finally, the complexity of the method, for T iterations, is not easily obtained: if a total mass equal to ν results in no more than $P(\nu)$ observations for every prototype at each

step, then the global complexity of the method has an upper bound of order $\mathcal{O}\left((P(\nu))^2UT\right) + \mathcal{O}\left(n(P(\nu)^2\right)$ (respectively for the iterations and the final clustering computation). However, the relation between ν and $Q(\nu)$ is hard to know in advance and can depend on the dataset distribution and on the training itself.

4.5.4 Mining job trajectories with sparse relational SOM algorithms

The data set already presented in Section 4.4 will be used here also for illustrating and comparing the two latter sparse approaches. The dissimilarities between career trajectories were computed this time using the optimal matching [80, 81] on the 12,560 unique career paths. This resulted in a non positive dissimilarity (6,651 eigenvalues out of 12,500 were found positive).

To assess the accuracy and the computational cost of both PCA SOM and sparse relational SOM, 100 maps were trained, using an R implementation of the methods, on a 40-nodes computer without concurrent access. All maps were trained for a 10×10 grid, equipped with a piecewise linear neighborhood, with 60,000 iterations. The entropy ratio preserved by PCA SOM was varied in {20%, 40%, 60%, 80%}. For sparse relational SOM, the mass parameter ν was varied in {95%, 99%} and the update parameter κ was varied in {1, 50}. Only 10 maps were trained using the standard relational SOM due to its very high computational cost.

Table 4.4 presents the results obtained in terms of normalized quantization error (QE), average intra-cluster inertia (ICI), topographic error (TE) and CPU time (only the clustering time is reported).The last column provides the final dimension or number of coefficients of the prototypes.

Methods	QE (×100)	ICI	TE (%)	CPU time	Stability (%)	Dimension
Relational SOM	20.99 (0.12)	**23.94** (0.24)	7.91 (0.66)	949582 (1 373)	77.65 (3.66)	12 500
PCA (80%)	23.49 (0.10)	24.07 (0.31)	8.27 (0.92)	251 (78)	75.81 (1.82)	392
PCA (60%)	15.36 (0.12)	24.32 (0.32)	8.57 (0.77)	136 (44)	75.72 (1.95)	44
PCA (40%)	5.61 (0.09)	26.26 (0.37)	6.98 (0.75)	114 (40)	77.13 (3.24)	8
PCA (20%)	**0.37** (0.00)	31.92 (0.95)	**0.82** (0.86)	**112** (33)	**86.31** (5.69)	2
sparse (95%, 1)	32.40 (0.74)	28.66 (1.36)	30.86 (6.88)	378 (3)	55.79 (1.37)	14
sparse (95%, 50)	25.36 (0.35)	26.57 (0.59)	11.15 (1.86)	655 (32)	63.64 (0.88)	14
sparse (99%, 1)	25.11 (0.17)	27.09 (0.46)	5.26 (0.72)	1025 (194)	68.08 (1.25)	50
sparse (99%, 50)	26.76 (0.36)	27.36 (0.71)	31.59 (4.53)	381 (28)	59.81 (0.90)	8

Table 4.4: Performance results of PCA SOM and sparse SOM (average over 100 maps and standard deviation between parenthesis) for the "trajectories" dataset. Parameters for the methods are given between parenthesis after the method name (% of entropy preserved in the projection for PCA SOM and maximum mass, ν, and update parameter, κ, for random ascending updates in sparse SOM).

As one may easily see, results demonstrate a high efficiency, in term of computational cost, of both approaches, while still preserving accurate results. The results also show that PCA SOM provides a good trade-off between the quality of the map and the dimensionality of the prototypes, outperforming the direct sparse approach. The best results for PCA SOM are obtained with 20% entropy-rate preserved. Nevertheless, this strategy selects only two dimensions, which increases the redundancy in the data and tends to produce clusters with few observations. Thus, the PCA SOM preserving 40% entropy should be preferred. The best results for the sparse K-SOM are obtained with a mass equal to 95%.

Both PCA SOM and sparse relational SOM provide accurate results in a reasonable computational time. For sparse relational SOM, prototypes can be interpreted by inspecting the properties of the few observations used to represent them. For PCA SOM, the projection of the data on a subspace requires to interpret the axes of the PCA as an extra step in order to understand the meaning of the prototypes.

Interpretability of PCA relational SOM

For illustrating how the results of PCA relational SOM may be interpreted despite the PCA pre-processing, we selected the map with the lowest ICI among the 100 with 40% preserved entropy rate. The mapping of the data on a lower-dimensional subspace requires to interpret the PCA axes first. Figure 4.9 (left) presents the entropy supported by the first 15 axes and shows that the first two axes are enough to provide relevant information on the data. Figure 4.9 (right) displays the projections of the observations on the first two principal axes. The first axis represents 16.90% of the total entropy and opposes permanent-labor and fixed term contracts. Stable job trajectories have the smallest coordinates on the first axis while fixed-term contract or unemployed

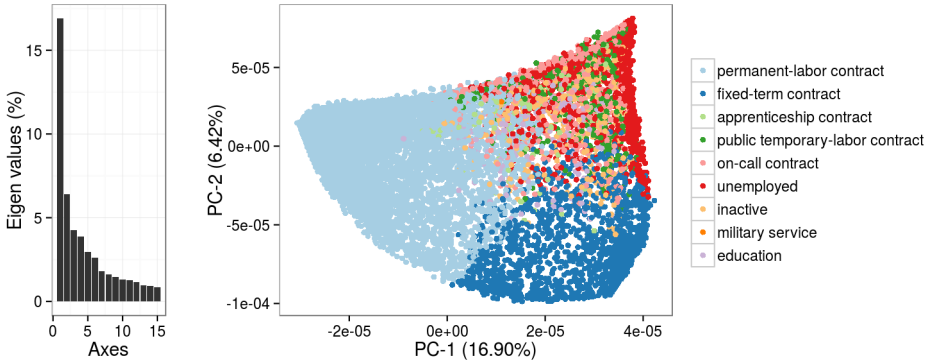


Figure 4.9: Entropy preserved by the 15 first axes on the left and projection of the observations on the first two principal components on the right. Colors represent the contract that appears the most often (mode) in the trajectory.

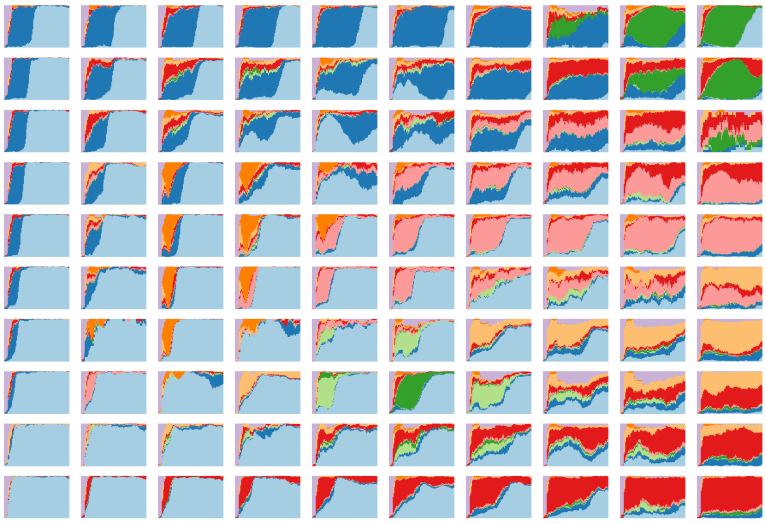


Figure 4.10: **PCA relational SOM**. For each neuron of the map, job trajectories distribution is represented using the observations classified in the corresponding unit. Colors represent the type of contract.

have the highest. Figure 4.9 (right) also demonstrates that the second axis separates two kinds of precarious situations. Fixed-term contracts are opposed to highly precarious contracts such as unemployment, inactivity, on-call contract and public temporary-labor contract.

The distribution of the job trajectories within each neuron of the map is represented in Figure 4.10. First note that the presented map is comparable in term of topology to the one described in [MO8]. Different typologies can be highlighted: a fast access to permanent contracts (clear blue) on the bottom-left corner of the map, a transition through fixed-term contracts before obtaining stable ones (dark and then clear blue) on the map top-left corner, temporary jobs (dark blue) on the top-middle neurons, a long period of inactivity (yellow) or unemployment (red) on the map bottom-right corner.

The map organization is in accordance with the axis interpretation. Figure 4.11 (top) displays the average values on the first and second principal components in every cluster of the map. Results show a gradient of the observation coordinates on the first PCA axis between the bottom-left and the right side of the map. This confirms that the first principal component (and corresponding diagonal on the map) separates permanent contracts from instable career paths. In Figure 4.11 (top), a gradient can also be observed for the second PCA axis between the top-left, where trajectories correspond to a fast access to permanent-labor contracts and the bottom-left corner of the map, where trajectories pertaining to precarious jobs are gathered.

Interpretability of sparse relational SOM

Similarly to above, let us discuss one of the final results obtained from sparse relational SOM. The selected map is again the one with the smallest ICI among all maps obtained with $\nu = 95\%$ and $\kappa = 50$.

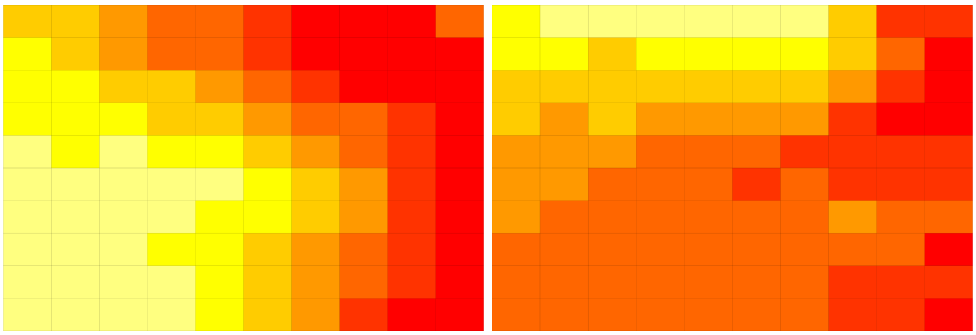


Figure 4.11: **PCA relational SOM**. Representation of the SOM map with neurons filled using colors according to the average coordinate of the observations for the first (on the top-left) and the second (on the top-right) principal component.

The resulting distribution of the job trajectories within the clusters of the map is provided in Figure 4.12. This distribution is fairly similar to the one obtained above: the left hand side of the map corresponds to a fast access to permanent contracts whereas the right hand side corresponds to different types of precarious situations. Two main differences can be highlighted: first, the class are more homogeneous in sparse relational SOM, especially at the border of the map. This is a direct effect of the dimension reduction in PCA relational SOM: since the dimension reduction increases redundancy in the dataset, some clusters (mostly located at the borders of the map) contain more observations and are thus less homogeneous. Second, the precarious situations (on the right hand side of the map) are organized a bit differently (with on-call contracts in the middle or the bottom of the map). However, both representations are realistic, with most of the clusters in the map being homogeneous.

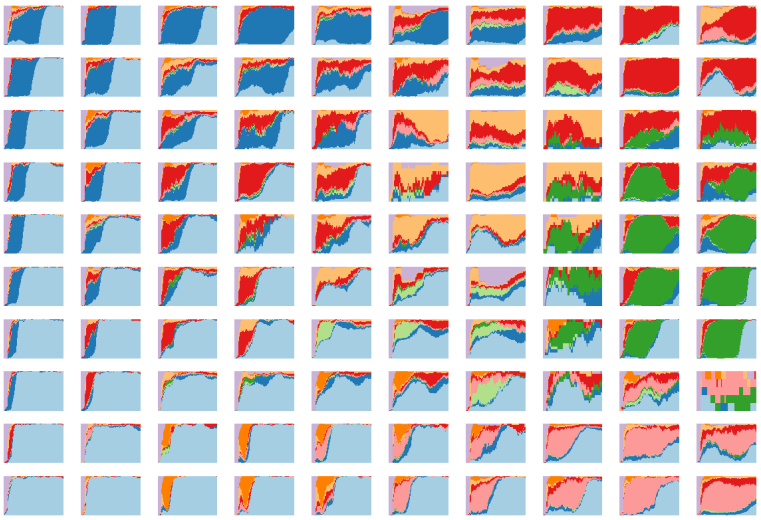


Figure 4.12: **Sparse relational SOM**. For each neuron of the map, job trajectories distribution is represented using the observations classified in the corresponding unit. Colors represent the type of contract.

Nyström approximation

To evaluate the relevance of using a Nyström approximation, the PCA relational SOM with 40% preserved entropy rate was used as a reference. Table 4.5 presents PCA relational SOM results using a Nyström approximation with different rates of observations sampled to perform the approximation. These were varied in {100%, 25%, 10%, 5%, 1%}, in which the 100%-results are reported from Table 4.4. The coefficients given to the PCA relational SOM are restricted to the first eight PCA axes everywhere, to avoid any bias related to data dimensionality. The computational time is reported in Table 4.5 and gives the time needed to perform the PCA only, excluding the training and clustering times.

Results demonstrate a high efficiency in terms of computational time of the Nyström approximation while producing accurate results. In fact, none of the tested values lead to deteriorate the map quality in term of QE, ICI and TE, while the PCA is $\sim$ 1000 times faster when using 10% of observations. The best ICI is even

Methods	QE (×100)	ICI	TE (%)	CPU time	Stability (%)
PCA (100%)	**5.61** (0.09)	26.26 (0.37)	**6.98** (0.75)	8 153 (205)	**77.13** (3.24)
PCA (25%)	5.62 (0.09)	26.11 (0.40)	7.12 (0.77)	101.38 (18.96)	75.84 (2.38)
PCA (10%)	5.62 (0.13)	26.13 (0.40)	7.00 (0.76)	7.39 (1.23)	74.68 (2.25)
PCA (5%)	5.64 (0.15)	26.05 (0.45)	7.11 (0.92)	0.86 (0.38)	73.10 (1.72)
PCA (1%)	5.65 (0.18)	**25.99** (0.47)	7.02 (1.02)	**0.02** (0.01)	69.32 (1.26)

Table 4.5: Performance results of the PCA relational SOM with PCA performed through a Nyström approximation (average over 100 maps and standard deviations between parenthesis) for the "trajectories" dataset. After the method name and between parenthesis, the percentage of observations used to perform the approximation is given.

obtained using only 1% of the observations. The clustering stability decreases with the number of observations used by the Nyström approximation, even if the stability is still high when using at least 10% of the observations.

The maps with the smallest ICI among the 100 maps generated from a Nyström approximation using 1% and 5% of the observations are displayed in Figure 4.13. Results show the ability of the Nyström approximation to preserve a realistic representation of the dataset while reducing the computational time.

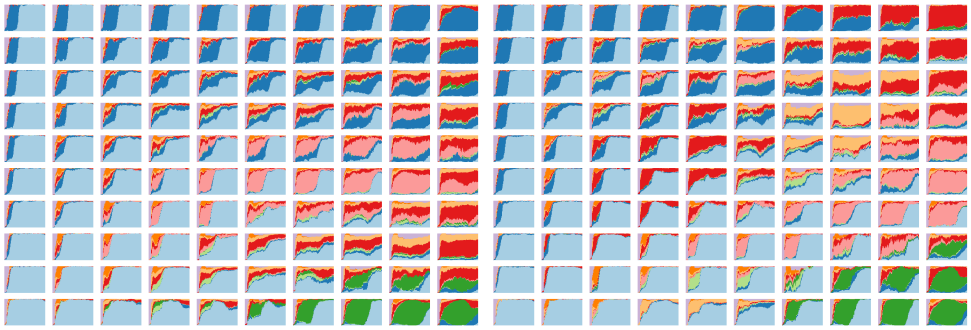


Figure 4.13: PCA relational SOM performed through a Nyström approximation using 1% (left) and 5% (right) of the observations: For each neuron of the map, job trajectories distribution is represented using the observations classified in the corresponding unit. Colors represent the type of contract.

The map obtained with 5% of the observations shows an organization similar to the one presented in Figure 4.10.With a Nyström approximation using 1% of the observations, trajectories mostly containing fixed-term contracts are located on the right hand side of the map. This output is different from the other maps illustrated here. As expected, clusters obtained using 1% of the observations are less homogeneous than those obtained with 5%. The differences between these two maps might have different causes. Firstly, the instability of the SOM algorithm can explain the differences in terms of map organization: different runs of the algorithm give different results. This is particularly critical when the dataset to be analyzed is high dimensional as can be the "Generation 98" survey (even with a subsampling rate of 1%, the dimensionality of the problem is still larger than 100). This issue could be addressed by aggregating strategies, as described in [140]. Secondly, the differences between the two maps in Figure 4.13 might be explained by the high redundancy of trajectories with fixed-term contracts in the dataset: a very small subsampling might enforce the over-representation of these trajectories and affect the result. Such a problem could be addressed by using more efficient sampling techniques such as the ones described in [139].

The choice of the ratio m/n of observations to select in order to obtain accurate results highly depends on the quality of the kernel approximation provided by the Nyström technique. This quality is strongly influenced by the rank of the kernel, which can not easily be obtained when n is very large. Adaptative sampling technique for the Nyström algorithm, such as the one described in [141, 142] are based on an unequal probability sampling, which is performed iteratively and depends on the reconstruction error. [139] proposes an improved version in which the full kernel is not even needed to estimate the reconstruction error. Such methods could be relevant to assess the evolution of the quality reconstruction in a growing sample and to stop the Nyström sampling when this quality is considered good enough.

4.6 Conclusion and perspectives

In this chapter, I described our contributions related to the extensions of self-organizing maps for complex data described by kernels and dissimilarities. More particularly I focused first on a version of the algorithm designed to handle multiple kernels or dissimilarities, and second on several frameworks for large datasets which need sparse approaches in order to have reasonable computational times and interpretable prototypes in the outputs. I also mentioned the R-package SOMbrero that we implemented, and which is used by several of my colleagues in statistics and quantitative social sciences.

Currently, there are several aspects that I would like to study or further develop.

First, from a very practical point of view, SOMbrero needs additional facilities to be implemented. One of them is handling missing data in the numerical vectors context. This is one recurrent demand coming from various users of the package, and it is an important issue in practice. Thus, we aim at adding new functions inspired by the methodology described in [143] for dealing with this question. Also, the current version of the package does not contain neither the extensions to multiple dissimilarities, nor the sparse ones. These should be progressively added to the package, together with new vignettes illustrating these algorithms on real data.

From a more methodological point of view, I am particularly interested in connecting relational self-organizing maps and related algorithms with the exploratory analysis of temporal directed graphs. I am currently studying such a network (French cattle trading system), which is additionally a large one, as an investigator in the ANR-funded Cadence project, hosted by MaIAGE, INRA. For the moment, I chose to compute dissimilarities between vertices based on temporally reachable paths, and train a bagged kernel SOM on the resulting matrices of distances. However, computing temporal distances has a large complexity, while the directed edges lead to non-symmetric dissimilarities. The latter was managed in practice by considering a bipartite graph, where each vertex appears twice as outbound or inbound, and by adapting the algorithm for taking this into account. Although the first results are encouraging in terms of interpretability, the computational burden is quite heavy. Mixing this approach with a spectral one, comparing it with other clustering techniques for networks both in terms of interpretability and complexity are some of the extensions and developments to be further considered.

Another aspect that I would look at into more details, and also from a methodological point of view, is the possible input of self-organizing maps and related methods in the study of spatial data, and more particularly in the study of residential segregation. This is a topic that I will further develop in the next chapter, but I would like to mention here a very nice property of self-organizing maps which is the stochasticity of the online algorithm and the absence of a unique solution for the clustering. This property of obtaining (slightly) different clusterings when a large number of maps is being trained, leads to the emergence of *strong patterns*, and also to that of *fickle* inputs. When combined with the spatial distribution, this could be particularly informative on the underlying structure of the data.

Chapter 5

When space steps in

"La forme d'une ville
Change plus vite, hélas! que le coeur d'un mortel."
Charles Baudelaire

5.1 Introduction

This summer, as I was finishing O. Pamuk's "A strangeness in my mind" and with my mind full of the recent history of Istanbul and of its demographic changes, I was brought to reflect upon the surprising domino effects that hazard produces, making all pieces of a puzzle falling into place. I was just about to start this last chapter, devoted to the use of exploratory techniques for assessing residential segregation. This chapter is the outcome of some of my most recent work, and some of my most recent collaborations. Although it all started two years ago as some "armchair philosophy" about spatial patterns and individual perceptions of the city, my feeling is that we hopefully managed to nail some core issues in the study of urban segregation. I am convinced that this happened thanks to a genuine interdisciplinary exchange within a small group of geographers, statisticians, statistical physicists, and machine learning scientists.

According to [144], cities are the perfect illustration of complex systems: individual and institutional agents interact on multiple levels of multiple networks (both physical and virtual), leading to nontrivial collective behaviors and nontrivial patterns at many scales. Among a number of intricate questions that have emerged in the study of urban systems, that of socio-spatial dissimilarities stands out as one that has aroused interest for decades, and across a wide range of fields. Assessing the complexity and the multiple facets of urban segregation remains a hot topic despite the abundant literature addressing it, and even more so these last years, with more and more fine-grained data becoming available and giving rise to new computational and modeling challenges. In this context, an interdisciplinary approach is not only desirable or trendy, it becomes vital for driving things forward.

Various contexts can be characterized as reflecting one form or another of segregation. The results I will present in this chapter are related to residential segregation, defined, for example, as a phenomenon where the distribution of a given variable locally (at the neighborhood scale) differs substantially from its distribution at the scale of the whole city. I will not discuss here more complex forms of segregation, taking into account geo-mobility, social networks or other forms of interactions in the city. Some of the methods we developed may be potentially extended to some of these cases, but I will come back to these issues in the conclusion of the chapter.

From where I stand and as far as I see things, I would say the question of residential segregation has been tackled using two very different approaches.

On the one hand, there is a broad literature (briefly reviewed, for instance, in our paper [MO1]), notably supplied by the statistical physics community, and focused on multi agent and interacting particle systems modeling. These are mostly mechanistic models, inspired by the pioneer work of Schelling's checker-board simulation of a two-community dynamics [145], [146]. Although these theoretical studies come with very elegant settings and sophisticated mechanisms explaining the emergence of segregation, rare are the instances where real-world data have been compared with theoretical results other than stylized facts [147], [148], [149]. One reason for this is the notorious difficulty to elaborate, based on available data, segregation indices [150] that could correspond to some of the theoretical ones - the evolution of which is described in agent-based modeling.

On the other hand, an impressively large number of exploratory analyses on real data led to many proposals of segregation indices, aimed at capturing the details of any regularity emerging in the spatial distribution [151]. The numerous existing indices may be classified in at least two categories [152, 153]. First, there are the zone-based indices such as the dissimilarity index [154, 155, 156], the proximity index [157] or the concentration profile [158] which work at fixed scales. They are all liable to the Modifiable Areal Unit Problem (MAUP) [159]. Second, surface-based measures [160, 161] use a continuous population density surface to circumvent the MAUP. But data most often comes as already aggregated units, so these indices usually require refined statistical interpolation techniques. Furthermore, they are not scale-free, since one has to select values for the radius within which the population density is estimated and the dissimilarity indices computed. Another class of indices largely used in spatial statistics is that based on spatial autocorrelations [162], sometimes coupled to a subsequent clustering. Although easy to compute and helpful in practice, these require introducing an a priori dependence structure on the grid of spatial units and on the shape of the influence of one spatial unit upon another.

However, as pointed out in [163] or [164], segregation is essentially a multiscalar phenomenon, while multiscalar approaches have been introduced only recently [165], [166].

The starting point of the work leading to the contributions I will present in this chapter was the questioning of whether starting from real data and an exploratory setting, one may drive the analysis, step by step, towards a more theoretical, possibly mechanistic, framework. In particular, we aimed at including a multidimensional perspective (use several variables describing the data instead of a unique one as is commonly done in studies on segregation), and a multiscalar one. Our contributions are thus two-folded, and may be seen as two separate, or at least not unified yet, research projects.

The first project was focused on the multidimensional aspects of the real data, and mainly involved Aurélien Hazan, Marie Cottrell, Julien Randon-Furling and myself. Since we started from real data and with an exploratory goal, at least in the beginning, we turned towards the self-organizing maps algorithm that I presented in the previous section and which was particularly appealing in this context. Indeed, SOM produces homogeneous clusters starting from multi-dimensional data, while mapping the data in a low-dimensional space preserves its topology. The clustering, together with the stochasticity of the online version of the algorithm, allows to find structure in the data, and to assess how discriminant this structure is. The mapping, when combined with the geographical distribution of the data, provides a good indicator of whether the patterns identified with the clustering have specific spatial distributions. Our empirical findings on a Parisian dataset were published in [MO27] and [MO1], and I will describe them in Section 5.2. We are currently investigating a new multidimensional index of segregation, based on SOM, mixing a robustness index on the underlying structure of the data with the correlation between the geographical distances and the distances on the SOM mapping. I will further discuss this in the conclusion of the chapter.

The second study was carried out mainly with William Clark and Julien Randon-Furling, as I will present it in Section 5.3. We tackled the multiscalar issue starting from a common-sense remark, that an individual perceives segregation all the more acutely as she has to go *a longer way* from her home to discover what the city in its entirety might look like. We built upon the idea of *egocentric profiles* or *egocentric signatures* already proposed in [166] among others, and computed individual trajectories in the city, using all possible scales in the data. Using these trajectories, we introduced new concepts such as *focal distances* and *distortion coefficients*, which are new measures for measuring the perceived segregation at individual level. Other colleagues contributed to this work: Antoine Luquiaud, a PhD student supervised by Julien Randon Furling who took a look at the ballot theorem in our paper [MO3], Cécile de Bezenac who has been implementing some of the algorithms in Python during her summer internship, Jean-Charles Lamirel who greatly contributed to one of the last papers published on this topic [MO24].

All the results to be presented next are mostly related to exploratory statistics, and are inspired by machine learning, spatial statistics, and statistical physics. In some sense, this might be frustrating, since the mathematical formalisms are not completely established yet. But, at the same time, I believe that these very new tools that we introduced, and especially those in Section 5.3, open new paths of research, both theoretical and applied, that I will try to summarize in the perspectives section. Our contributions up to present have been published in three journal papers [MO1], [MO2], [MO3] and three peer-reviewed conference proceedings [MO24], [MO25], [MO27].

I should mention here that this work benefitted from a significant amount of institutional interest, which materialized both as financial support, and as additional collaborations. Julien Randon-Furling and myself got a research grant from the board of the University, which allowed us to invite William Clark several times in Paris, and also to make a short visiting stay at UCLA in 2018 (while I'm about it, let me also say that Julien Randon-Furling was my main collaborator for all the results presented here, and I shall never be grateful enough towards him for his networking skills!). Together with some Cuban colleagues from the University of Habana,

equally interested in the emergence of spatial patterns in cities and particularly in the spatial distribution of the aging population and her access to public facilities, we obtained a two-years PHC funding (2019-2020). This grant was combined with a joint PhD project: Dafne Garcia de Armas started to work with us early 2019, under the supervision of Sira Allende (Habana) and myself.

I've already said that measuring residential segregation and measuring spatial inequalities are by essence an interdisciplinary topic, but I should also say here that the impact of the research findings on this issue echoes beyond the scientific community, and more particularly in the eyes of public policy makers and local authorities. In 2018, I joined the Convergence Institute "Migrations", and, around the same dates, submitted a proposal for a data challenge organized by the European Commission and aimed at studying the integration of migrants in cities, `https://bluehub.jrc.ec.europa.eu/datachallenge/`. Our contribution was published in a technical report [167] after a showy workshop in Brussels.

In the next two sections, I will briefly present our findings. Section 5.2 describes how we used self-organizing maps for highlighting multidimensional spatial patterns in the cities, while Section 5.3 summarizes the new concepts we introduced, such as *focal distances* and *distortion coefficients*. A conclusion with perspectives for current and future work will follow.

5.2 Assessing residential segregation with self-organizing maps (SOM)

5.2.1 The data

We illustrate the proposed methodology on a dataset comprising several variables for the city of Paris, recorded in 2014 and provided by INSEE (*Institut National de la Statistique et des Études Économiques*), France's Census Bureau, as well as some additional data provided by the IGN (*Institut Géographique National)* and RATP (*Régie Autonome des Transports Parisiens)*, Paris public transport agency.

Census data is provided as aggregated values on fixed spatial units, called IRIS (*Ilôts Regroupés pour l'Information Statistique*). They do not correspond to a fixed surface area, nor to a fixed number of inhabitants, although they are supposed to correspond to around 2,000 inhabitants, in average. For the city of Paris, having about 2 million inhabitants, the number of spatial units is just under 1,000, although some are purely geographical, with no or very few inhabitants. INSEE provides data such as the number and types of shops, public service offers, health facilities. They also provide deciles of the income distribution within each census unit, summary statistics on the age, education, social status, ... of the inhabitants. From the metropolitan authority for public transportation we obtained the geographical coordinates of all access points to underground, tramway and bus stations, and were able to compute for each spatial unit the number of underground and tramway lines available within an 800 meter radius (this was computed with respect to the centroid of the unit, and should be further refined in our future work).

Our goal was to first compare different sets of variables in terms of spatial distributions, and assess whether they produced different spatial patterns, and, eventually, more or less marked signs of segregation. From the available dataset, we built three sets of characteristics:

1. Revenue and income. We used the first and the ninth deciles, as well as the median of the income distribution within a spatial unit, the fraction of revenue coming from assets and other patrimonial sources, and the fraction of revenue coming from minimal social benefits.
2. Population characteristics. We focused on age (average and standard deviation), number of people under 18 in the household, education level for the head of the household.
3. Urban facilities and services. We included the rate of social housing, the access to public transportation, the access to medical and health services, the number of shops, of sports facilities, of primary and secondary schools, including primary schools in special urban and education development projects, called *éducation prioritaire* (EP).

5.2.2 Three facets of the city

We used the numerical SOM algorithm described in Section 4.2 for clustering the available data, using as input data each set of variables measured for all spatial units summarizing the city of Paris. For each set of variables, we trained a 8×8-map, which was further clustered by applying hierarchical clustering (HAC) on the resulting prototypes. Eventually, four super-clusters were obtained for the first two sets of variables, and six super-clusters for the third set. I briefly describe the features of each clustering.

Set 1 The variables in the first set, income and revenue, are the most commonly used in socioeconomic studies on segregation (along with ethnic groups, but let me recall here that ethnic statistics are not available in France). As in can be seen in Figure 5.1, SOM followed by HAC yields four easily identifiable types of spatial units (see also Table 5.1). The population in supercluster 1 is richer than average, and more particularly the top 10%, and has also a very substantial part of revenues coming from financial and other patrimonial assets. At the other end of the spectrum, the population in supercluster 4 is poorer than the Parisian average. In between, one has the upper (supercluster 2) and the lower (supercluster 3) middle classes, with again a difference in the level of patrimonial income. The clustering obtained with the first set of variables appears as significantly correlated with spatial segregation. Indeed, Figure 5.1 shows a high level of spatial homogeneity for the superclusters computed with the variables related to income and revenue.

Figure 5.1: Left: the Kohonen map for the variables of Set 1, with values of the variables for the prototypes in each cluster. Colors indicate the final super-clusters obtained with HAC. Right: Spatial distribution of the four super-clusters obtained with the variables of Set 1. Areas in white correspond to parks, train stations, hospitals, and blocks for which data is not available.

Super-cluster	1st decile	Median income	9th decile	Revenue from assets	Revenue from social benefits
1	13,405	44,367	129,538	40	0.2
2	12,504	33,281	73,929	22	0.5
3	9,471	23,963	50,338	13	1.2
4	7,390	15,081	30,840	7	3.6
All	10,672	28,411	65,309	19	1.1

Table 5.1: Per super-cluster averages of some of the variables in Set 1 (in euros for the deciles of income distribution, in percentage for the share of revenue drawn from financial and other assets).

Set 2 The clustering obtained on the second set of variables is much less structured than the first one (see Figure 5.2 and Table 5.2). However, one may eventually identify four superclusters and some underlying trends: the heads of the household are younger than the average in superclusters 1 and 2, whereas the level of education is lower than the average in supercluster 4. Spatially, superclusters are much less grouped in contiguous patterns and their distribution is more heterogeneous than in the previous case .

Super-cluster	Age	Age SD	Children per household	Education
1	39.2	17.7	0.3	2.8
2	39.3	14.4	0.6	2.7
3	44.5	18.6	0.4	2.7
4	46.4	16.7	0.5	2.1
All	42.6	17.6	0.4	2.6

Table 5.2: Per super-cluster averages of some of the variables in Set 2 (education level is from pre secondary (1) to postgraduate (5)).

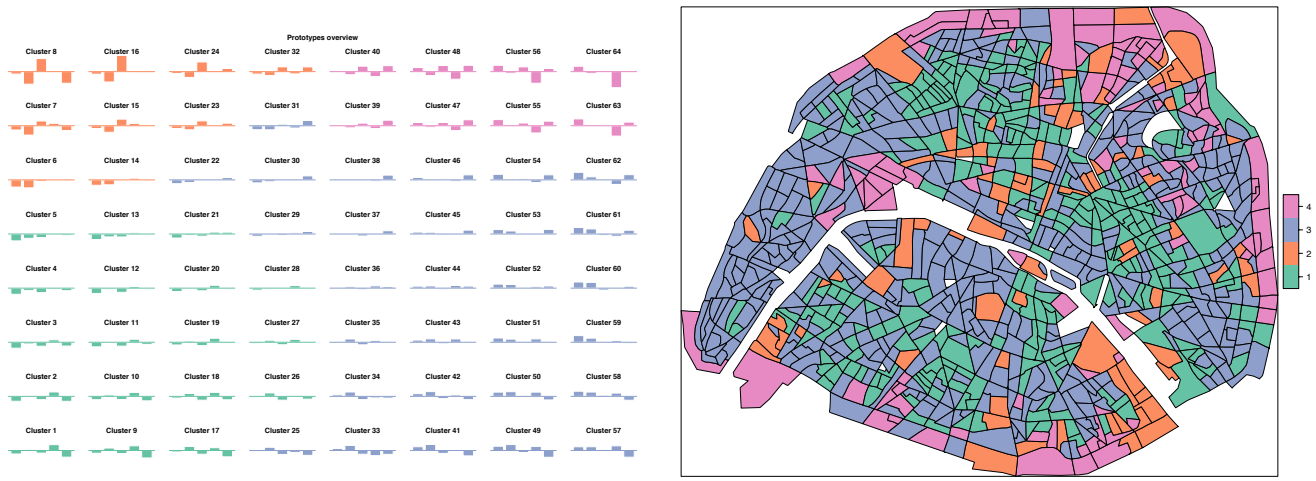


Figure 5.2: Left: the Kohonen map for the variables of Set 2, with values of the variables for the prototypes in each cluster. Colors indicate the final super-clusters obtained with HAC. Right: Spatial distribution of the four super-clusters obtained with the variables of Set 2. Areas in white correspond to parks, train stations, hospitals, and blocks for which data is not available.

Set 3 Processing the third set of variables with SOM allows one to distinguish six well-identified types of spatial units (Figure 5.3 and Table 5.3):

1. areas with more medical services, more private schools, fewer shops and less access to public transports;
2. areas with many shops, facilities, and the highest access to public transportation;
3. areas with a slight concentration of social housing, 9%, and below average for all other variables;
4. areas with a high level of access to public transportation, many shops and facilities, and also a certain number of EP primary schools;
5. areas with a significant proportion of social housing, 34%, very few EP primary schools, and the lowest access to public transportation;
6. areas with the highest proportion of social housing, 42%, the largest number of EP primary schools, and low access to public transportation and other facilities.

On this set of variables, a multidimensional approach sheds light on residential patterns by integrating various sources of information and subsequently of potential inequalities. If one looks at the spatial distribution of the superclusters, there are again some distinct areas that emerge as particularly representative of a given supercluster. There appears to be an important correlation between the third set of variables and the spatial distribution of the units.

Super-cluster	Social Housing	Medical doctors	EP schools	Shops	Public transportation
1	5	298	0.1	362	7.2
2	5	181	1.4	626	16.6
3	9	172	0.5	236	7.2
4	10	129	5.4	406	12.4
5	34	129	2.8	177	5.9
6	42	100	10.5	182	6.3
All	18	173	2.8	288	8

Table 5.3: Per super-cluster averages of some of the variables in Set 3 (social housing rate is a percentage per spatial unit, other variables are raw numbers for each spatial unit and its ten nearest neighbors; access to public transportation is the number of lines within 800m of a unit's centroid).

Which facet is the most segregated? Thanks to its multidimensional nature, SOM yields typologies of neighborhoods according to multiple variables taken into account simultaneously. Now, for a given set of variables, are all types of neighborhoods well mixed across the city, or are there any spatial patterns? In other terms, is there any form of spatial segregation along some of the variables considered here? Is there a set of

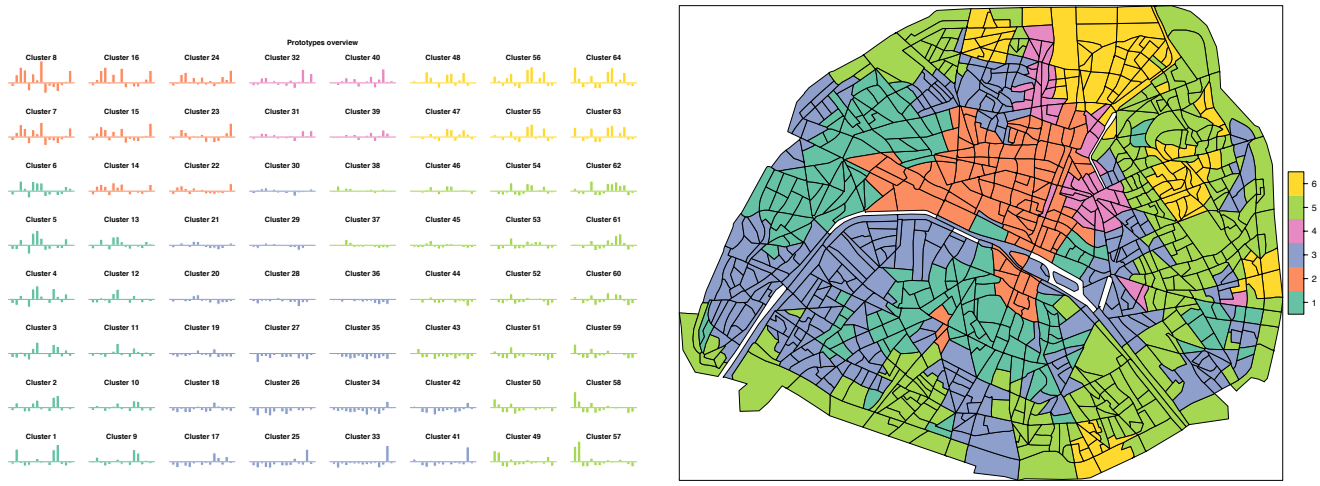


Figure 5.3: Left: the Kohonen map for the variables of Set 3, with values of the variables for the prototypes in each cluster. Colors indicate the final super-clusters obtained with HAC. Right: Spatial distribution of the six super-clusters obtained with the variables of Set 3.

variables for which segregation patterns are stronger than for the other two? Looking at the maps in Figures 5.1, 5.2 and 5.3, one sees spatial patterns, but how significant are they?

Since the data for the city of Paris is available as aggregated spatial units (this is the case for most public census data in most countries), we measure segregation by quantifying two different things:

1. first, we quantify how different spatial units are from one another, or, in other terms, whether there is structure in the multidimensional distribution of the units.

2. second, we are interested in the spatial concentration of the units of a given type, or in how far one stands from a uniform distribution over the whole city for units belonging to each group.

The first point may be addressed using the stochasticity of the SOM algorithm. Indeed, if there were no structure in the data, the clustering would not be very robust to the training of the algorithm with various initializations.

Segregation as an underlying structure of the distribution I will address here the first point above, which consists in assessing the robustness of the clustering, or, on the contrary, its volatility. I imply here that a volatile clustering means there is little underlying structure in the data. For doing this, we used the stochasticity of the map, and took advantage of an old idea of "strong patterns" introduced in [168], and later developed for self-organizing maps in [169], [170] and [171]. Essentially, one trains a large number of SOM's with different, independent initializations, and then looks at the pairs of inputs which are always clustered together or in neighbor clusters, and the pairs of inputs for which the associations appear to be random. A discriminant underlying structure in the data would lead to robust, recurrent associations, and very few random ones.

With some more formalism, let $(x_i)_{i=1,...,n}$ be the input data to be clustered, and L the total number of trainings of the SOM, with different initializations. For each $i, j = 1, ..., n$ and $l = 1, ..., L$, one defines

$$neigh^l_{i,j} = \begin{cases} 1 & , \quad \text{if } x_i \text{ and } x_j \text{ are neighbors in the } l\text{-th training of the SOM;} \\ 0 & , \quad \text{otherwise.} \end{cases} \tag{5.1}$$

In the following, we use a fixed crisp function for deciding whether two inputs are in neighboring units of the map: two SOM units are said to be neighbors if the shortest path distance on the SOM grid is smaller or equal to 1 (this a priori choice should be further investigated in the future by using smoother functions, adapting the threshold to the size of the grid and the number of clusters U, ...).

Next, one defines $Y_{i,j} = \sum_{l=1}^{L} neigh^l_{i,j}$, $\forall i, j = 1, ..., n$, as the number of times x_i and x_j are neighbors after L different, independent trainings, and the stability index $M_{i,j} = Y_{i,j}/L$ as the average over all trainings. Via a statistical hypothesis testing, one may check whether $M_{i,j}$ is completely different from its expected value were x_i and x_j neighbors in a complete random way.

If edge effects are not taken into account and with the neighborhood structure defined above, the number of units involved in a neighborhood for a regular two-dimensional grid is equal to 9. Hence, for any pair x_i and x_j,

the probability of being neighbors in a random way is $\frac{9}{U}$. Using a Gaussian approximation for $Y_{i,j} \sim \mathcal{B}(L, \frac{9}{U})$, one may build a critical region for a test of level 5%:

$$\mathcal{C}_{L,U,0.05} =]-\infty, A - B[\cup]A + B, +\infty[\ , \tag{5.2}$$

where

$$A = \frac{9}{U} \text{ and } B = 1.96\sqrt{\frac{9}{UL}\left(1 - \frac{9}{U}\right)} \ . \tag{5.3}$$

In practice, for each pair of inputs x_i and x_j, one computes the stability index $M_{i,j}$ and applies the following rule:

- if $M_{i,j} > A + B$, then x_i and x_j are almost always neighbors in a significant way, they *attract* each other;
- if $M_{i,j} < A - B$, then x_i and x_j are almost never neighbors in a significant way, they *repulse* each other;
- if $A - B \leq M_{i,j} \leq A + B$, then x_i and x_j are neighbors due to randomness, they are a *fickle* pair.

Eventually, for each input x_i, $i = 1, ..., n$, one may compute its index of *volatility* or *fickleness* as the percentage of fickle pairs to which it belongs:

$$\mathcal{V}_i = \frac{\#\{j \neq i \ , \ |M_{i,j} - A| \leq B\}}{n - 1} \ . \tag{5.4}$$

With this index, one may represent each data point with its associated *volatility* index, and decide upon the randomness of the associations and the lack of structure in the data.

As an illustration, we trained $L = 100$ maps with different initializations on the three sets of variables above. All maps had the same size, $U = 8 \times 8$. I will remark here that, actually, the three maps presented above were each selected as being the one minimizing the quantization error among the 100 clusterings obtained on each set of variables. The percentage of *fickle* pairs is 4% with the first set of variables, 7.5% with the second, and 9.6% with the third. SOM clusterings appear as being quite robust overall, although the one computed with the first set of variables is the less volatile, indicating a greater level of differentiation among the input spatial data, and most probably a more discriminant underlying structure.

In Figure 5.4, we show the *volatility* indices per input data, for the three sets of variables, ordered decreasingly. The larger the number of fickle inputs, the less definite will any clustering be. Note that this is not only a measure of the robustness of the clustering, it is actually measuring the level of heterogeneity between areal units, independently of their spatial distribution. In this respect, *volatility* indices provide a first level of information concerning segregation, in terms of local composition. They will be complemented by the indices taking into account the spatial aspects, that I will present next.

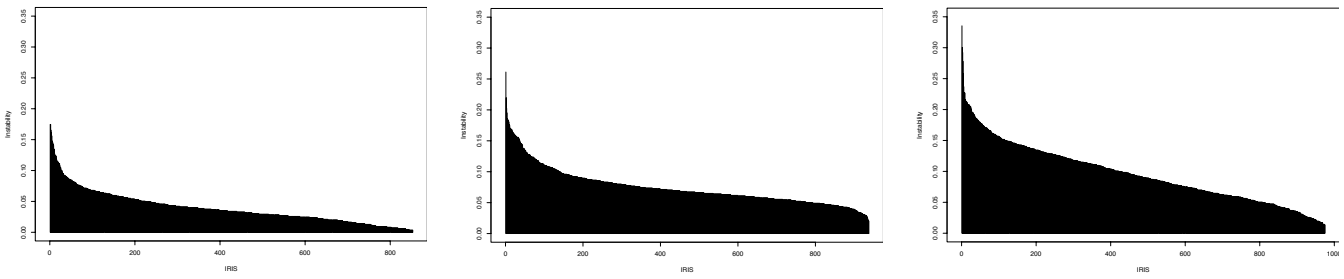


Figure 5.4: Volatility of SOM-based clusterings on each of the three sets of variables. Plots show levels of *fickleness* for each spatial unit, in decreasing order of magnitude for Sets 1 (left), 2 (middle), and 3 (right).

Segregation as a clustering spatial distribution issue The second point above, assessing segregation as a question related to the spatial concentration of some characteristics of the population, may be dealt with using various approaches inspired from physical statistics, spatial statistics or machine learning. One very common approach is to use entropy and information-theory based criteria. We compare these historical indices with a new one, that we introduced in [MO1], and that makes use of the specific properties of SOM.

For illustration purposes, we compute several common indices based on the entropy, such as the spatial dissimilarity index ($\tilde{D}$-index), the spatial relative diversity index ($\tilde{R}$-index), and the spatial information theory index ($\tilde{H}$-index). These indices have been defined for the multi group context and with a spatial constraint in [151] and [150]. They measure the difference between the entropy of the global distribution at the city scale and local distributions. We used the normalized versions of the indices implemented in the R-package seg [153],

which allow for a comparison: the values of the indices range from 0 to 1, where a value of zero represents no segregation and a value of 1 indicates complete segregation. A summary of the values obtained for these three standard segregation indices is given in Table 5.4. All three indices indicate stronger segregation on the third set of variables, and weaker on the second one. The first set appears closer to the third one in terms of segregation.

Set of variables	$\tilde{D}$-index	$\tilde{R}$-index	$\tilde{H}$-index	SOM-based Index
1	0.64	0.43	0.50	0.26
2	0.44	0.22	0.28	0.13
3	0.72	0.49	0.60	0.18

Table 5.4: Values of various segregation indices for the three clusterings on the three sets of variables.

The new index we propose takes advantage of the topology preservation property of the SOM algorithm. Indeed, since the proximity on the SOM grid preserves the proximity in the multidimensional space of the variables used for clustering, the closer two inputs are on the SOM grid, the more similar they are for the variables under consideration. Now, if the city were spatially well mixed, one would not observe any particular spatial pattern: geographical distances would be independent of the distances on the grid. Thus, any correlation between geographical and SOM distances signals spatial patterns, i.e. the presence of segregation, the level of which is well quantified by the actual value of the correlation.

We illustrate the link between geographical distances and grid distances in Figure 5.5, and give the values of the correlations for the three datasets in Table 5.4. Two remarks may be immediately made. First, the values of these correlations appear to be significant, although much weaker than the segregation indices based on the entropy. Second, although the second set of variables appears here also as the less subject to spatial segregation, there is a switch between the first and the third set of variables in terms of relative importance of segregation. The first set of variables appears as being the one producing the most significant spatial patterns, and in some sense this is consistent with the fact that the first set of variables displayed also the smallest rate of cluster volatility. According to our SOM-based approach, it is mostly income and revenue that yield robust clusters, well separated, which have also a significant spatial structure. This conclusion is yet to be nuanced, since our study is for the moment very preliminary. We still need to investigate, mostly empirically and analytically whenever this is possible, the possible ranges for the correlations as a function of the number of clusters, the form of the spatial patterns, etc. We already know, for example, that if one considers four equal groups fully separated on the four quadrants of a square city, the correlation may be computed exactly and is equal to 0.61. We also know that if the city has a circular form and the clusters are distributed in concentric patters, the correlation measure will not be able to quantify them.

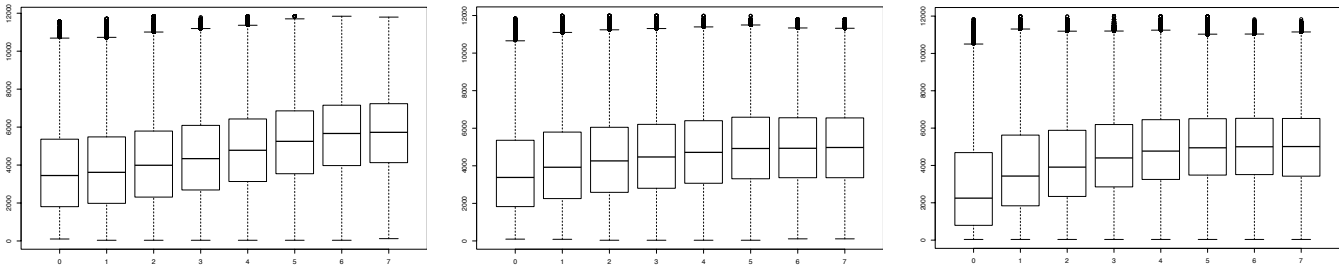


Figure 5.5: Visual representation of a SOM-based segregation index for the units clustered with the variables in Set 1 (left), 2 (middle), and 3 (right). The x-axis contains the shortest-path distance on the Kohonen grid for any pair of spatial units (the maximum value is 7, since the maps are 8×8), while the y-axis contains the geographical distance between the centroids.

5.3 Residential segregation perception through a multiscalar lens

Let me summarize next the different concepts related to the individual perception of segregation, that we recently introduced in [MO2] and [MO3]. They are based upon the observation that the image one has of the city is all the more *blurred* that the neighborhood she lives in is segregated, at all possible scales. We formalized this as *individual trajectories* in the city, *focal distances* and *distortion coefficients*, and eventually compared the real city, from the actual data, with a null model, of an unsegregated city.

5.3.1 Individual trajectories as a multiscalar fingerprint of the city

Since segregation patterns arise from a myriad of individual situations and perceptions, we build upon the idea of individual experiences and the fact that the longer an individual has to go from her home to seize what the city looks like as a whole, the more cut-off from the rest of the city she will feel. In the extreme case of a city where two equal groups A and B live in total separation, thus forming two ghettos, an individual living at the heart of one of the ghettos would have to explore the whole city to find out that it actually comprises equal proportions of each group. On the contrary, starting from the boundary between the two ghettos, there would be no need to cover so large an area to come to the same realization. This basic observation led us to introduce a new mathematical object that allows to capture and measure spatial dissimilarities as an individual and multiscalar phenomenon across the city.

One may indeed compute individual trajectories encoding the perception of the walker while visiting first her direct neighborhood, then the next closest one, and so on, gradually, until having visited the entire city. This exploration process considers ever larger neighborhoods around a starting point, all the way up to the whole city.

More formally, suppose the data is known as a set of spatial units or polygons $(u_i)_{i=1,\dots,N}$, at an already *basic* aggregated level (other configurations such as geo-localized individual data may be similarly dealt with). To each spatial unit u_i is associated an empirical distribution of some random variable, measured on the n_i individuals belonging to unit u_i. Next, for a given starting unit u_i, one may sequentially aggregate the rest of the units, using for instance a nearest neighbor rule (other aggregation procedures, based on enlarging radii etc, may equally be used). At step k of the aggregation procedure, k spatial units have been merged, including the starting one, and one may compute both the empirical distribution $\hat{f}_{i,1:k}$ of the aggregated population on the k units, as well as any dissimilarity or divergence with respect to the distribution of the whole city, $d(\hat{f}_{i,1:k}, f_0)$.

Once one has aggregated all N units around u_i, she obtains the individual trajectories in terms of distributions, $(n_{i,1:k}, \hat{f}_{i,1:k})_{k=1,\dots,N}$, or in terms of divergence, $(n_{i,1:k}, d(\hat{f}_{i,1:k}, f_0))_{k=1,\dots,N}$, where $n_{i,1:k}$ is the size of the population in the first k aggregated units around u_i. One may also remark that $\hat{f}_{i,1:N} = f_0$ and $d(\hat{f}_{i,1:N}, f_0) = 0$, which is equivalent to saying that each trajectory eventually converges to the city as a whole. Nevertheless, trajectories may differ widely from one starting point to another. Some converge quickly, corresponding to areas where even relatively small aggregates present a reasonably good picture of the whole city. Other trajectories, on the contrary, converge very slowly, corresponding to areas where segregation effects build up across scale and accumulate far beyond the local level.

For illustration, we use the Parisian data described in Section 5.2.1 and focus on the social housing distribution. In this case, the variable of interest has a Bernoulli distribution, and the empirical distributions $\hat{f}_{i,1:k}$ may be summarized by $\hat{p}_{i,1:k}$, the proportions of social housing within the first k aggregated units around u_i. As one may see in Figure 5.6, social housing is unevenly distributed across the city: while the actual average is 17,86% (close to the 20% official target), the median is equal to 7% only, and the rates in each spatial unit go from 0% to 97%, with a concentration of the high-rate units on the borders.

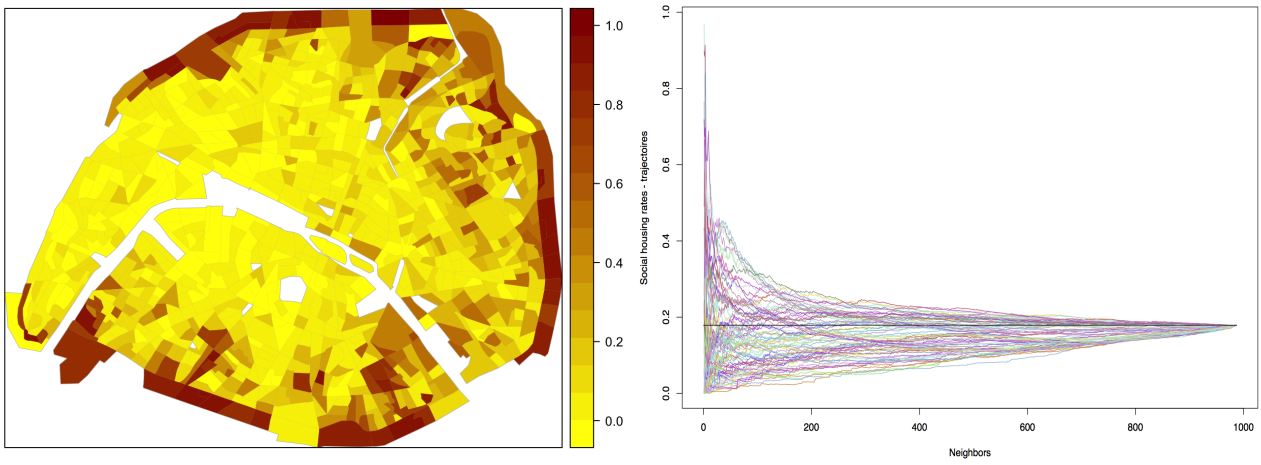


Figure 5.6: Left: social housing rate per spatial unit. Right: Trajectories for the social housing rate, starting from some (10%) of the 937 spatial units in Paris. The solid line corresponds to the city's average (17.9%). The x-axis is the number of aggregated units at each *instant* of the trajectory.

We also display in Figure 5.6 a sample of the trajectories computed from the social housing rate. They display an important heterogeneity, some starting from below the average and converging very slowly, others starting well above the average, and converging very fast.

At this point, I may say already that combining statistical data with geographical information and computing all possible individual trajectories in the city, allows one to create a complete and powerful mathematical object, a *fingerprint* of the city containing all the information about the variable of interest, at all scales and from the finest possibly available point of view. Nevertheless, despite its richness, the size and the complexity of this object make it difficult to use as such for large datasets or to explore in practice. One needs to summarize it into new indices and features, which should preserve a maximum amount of the information and bring to light the patterns of perceived segregation.

Before introducing the indices we proposed, let me emphasize the link between the notion of individual trajectories as defined above, and multiscalar approaches used in recent papers such as [172], [173] and [165], who use cross-sections of the trajectories defined here, i.e. values at certain points only. From this point of view, our approach may be seen as a general framework for the previous multiscalar proposals.

5.3.2 Focal distances

When summarizing the trajectories, one may be interested, for instance, in quantifying the speed at which their convergence occurs. If one fixes a convergence threshold δ, it is then possible to compute, for each trajectory, the *instant* when it enters (and remains thereafter) in the interval $[0, \delta]$. This *instant* is the size of the population that one needs to aggregate, for a given starting point, to have a distribution of groups that remains close, within a given threshold δ, of the reference distribution. In other terms, this is, for a given starting point, how far one needs to go in the aggregation process to *see* (with precision δ) the city's population as it is in its globality – the distance one needs to cover in order to get a relatively *clear* picture of the city. We call this convergence size the *focal distance*. Formally, the focal distance at point i, with precision δ, is defined as

$$\tau_i(\delta) = \inf_{k=1,\dots,N} \left\{ n_{i,1:k} \,|\, \forall \tilde{k} \geq k, d(\hat{f}_{i,1:\tilde{k}}, f_0)) \leq \delta \right\}, \tag{5.5}$$

where $n_{i,1:k}$ is the size of the population in the first k aggregated units around u_i. If the variable of interest is distributed according to a Bernoulli summarized by its parameter p, as in the example above on social housing rates, then focal distances may in particular be defined as

$$\tau_i^{\mathcal{B}}(\delta) = \inf_{k=1,\dots,N} \left\{ n_{i,1:k} \,|\, \forall \tilde{k} \geq k, |\hat{p}_{i,1:\tilde{k}} - p_0| \leq \delta \right\}. \tag{5.6}$$

Focal distances translate the distortion in the perception of the city than an individual living in unit u_i has: the smaller the *focal distance*, the less cut-off from the city she will feel; the larger the *focal distance*, the more persistent the feeling of segregation. An illustration of this concept for the data on the social housing rate in Paris in given in Figure 5.7, for two specific Paris districts. The convergence threshold δ was fixed at 5%. On the left, the Champs-Elysées and their surrounding area display trajectories converging very slowly and from below the city average, the 8th district being the "core" of the rich neighborhoods of Paris, with almost no social housing. On the right, in the 11th district, a working-class area, trajectories converge very fast.

Focal distances allow to draw an already significant picture of individual segregation patterns, although there is some arbitrariness in the choice of the convergence threshold. This may be circumvented by considering all its possible values. For any starting unit u_i and for $\delta = 0$, convergence occurs only at the very end of the trajectory, so that $\tau_i(\delta) = n_{i,1:N}$ ($n_{i,1:N}$ is the total population). When δ is large, convergence occurs immediately so that $\tau_i(\delta) = 0$. One may then study the variation of $\tau_i(\delta)$ as δ increases. The higher the curve representing $\tau_i(\delta)$, the longer the focal distances for u_i even at large values of δ (i.e. when convergence is easier). This helps identifying points in a city where spatial dissimilarities accumulate on multiple scales to create veritable "hotspots" of segregation. From these points, what one perceives of the city is very much altered, even on large scales, compared to what the city looks like in actuality. We formalized this concept into what we termed *distortion coefficients*.

5.3.3 Distortion coefficients

We formally introduced the notion of *distortion coefficients* in the particular case of multinomial distributions and for trajectories encoding the Kullback-Leibler divergence of the distribution $\hat{f}_{i,1:k}$ from f_0, where $\hat{f}_{i,1:k}$

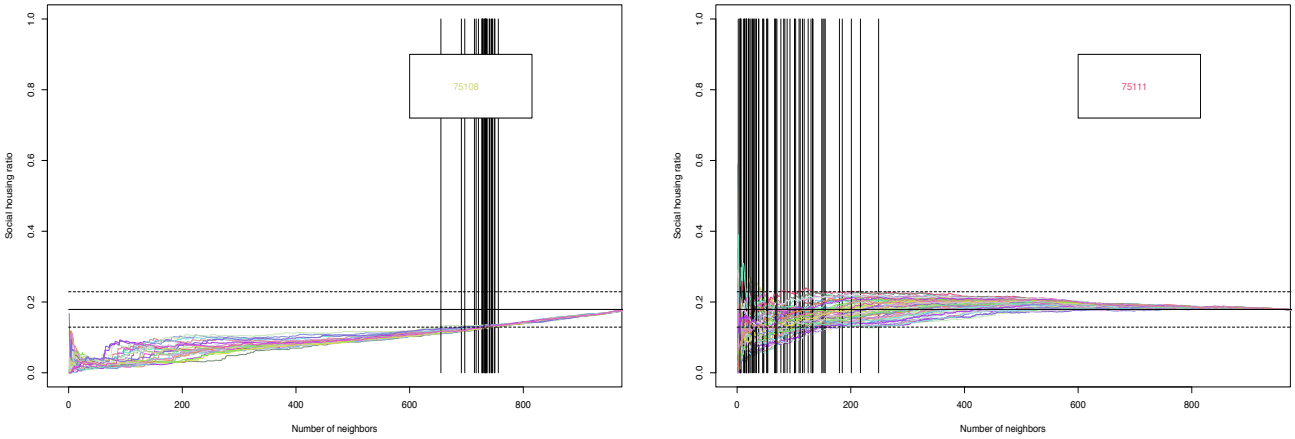


Figure 5.7: Trajectories for the social housing rate, starting from each census block in Paris 8th (left) and 11th (right) districts. The solid flat line is the city average, the dashed lines correspond to ±5% around it. Solid vertical lines correspond to *focal distances*. The x-axis is the number of aggregated units at each *instant* of the trajectory.

summarizes the empirical frequencies of the groups in the first k aggregated units around u_i, and f_0 is the distribution in the whole city.

For each unit u_i, we averaged the *focal distances* $\tau_i(\delta)$ for all δ, and obtained a measure of how distorted the perception of the city is, from location u_i. Formally, we simple integrated the focal distance curves, that is, we defined the *distortion coefficient* associated to the spatial unit u_i as

$$\Delta_i = \int_0^{\delta_{i,\max}} \tau_i(\delta)\, d\delta \ , \tag{5.7}$$

where $\delta_{i,\max} = \max_{k=1,\dots,N} d_{KL}(\hat{f}_{i,1:k}, f_0)$, where d_{KL} denotes the Kullback-Leibler divergence and

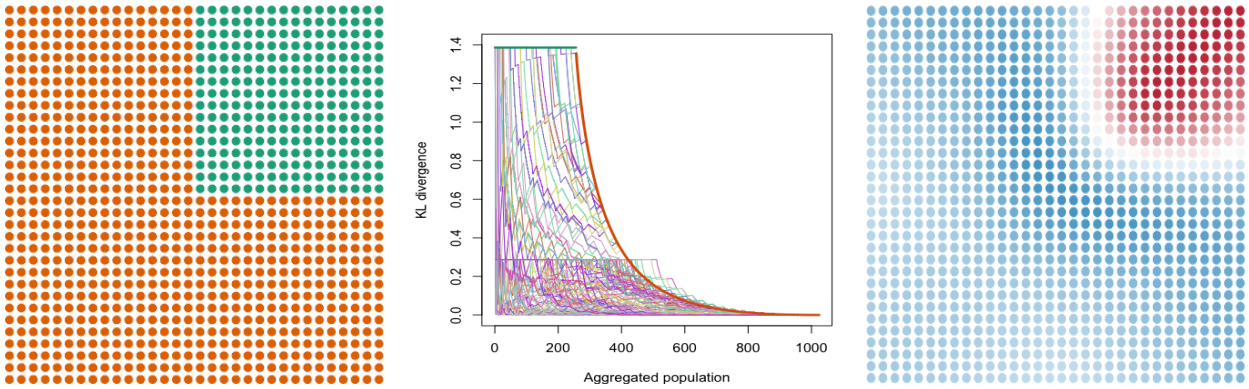
$$\tau_i(\delta) = \inf_{k=1,\dots,N} \left\{ n_{i,1:k} \,|\, \forall \tilde{k} \geq k, d_{KL}(\hat{f}_{i,1:\tilde{k}}, f_0)) \leq \delta \right\}. \tag{5.8}$$

In practice, since one will ultimately wish to make comparisons between different configurations (different variables measured for the same city, a same variable measured for different cities or at various time instants), *distortion coefficients* should be independent of the size of the city, and also of the reference distribution. We proposed a first attempt for a normalization procedure in [MO2], designed for multinomial distributions without any ordering of the categories. The *normalized distortion coefficients* are defined as $\tilde{\Delta}_i = \Delta_i / \mathcal{N}$, where the normalizing constant $\mathcal{N}$ is chosen as the maximum distortion coefficient in a theoretical extreme case of segregation. Theoretically, the maximal-segregation distortion coefficient is achieved when sorting the m groups into m ghettos, ordered by increasing frequencies, and then computing the coefficient for the most isolated person in the smallest group. This person would first meet all the individuals of his own group, then all those of the second most unfrequent group, and so on, until having seen the entire population of the city. The normalized distortion coefficients $\tilde{\Delta}_i$ take values between 0 and 1, and express the levels of distortion as a fraction of the perspective one has from the theoretical maximally-segregated unit (for given group proportions).

In the case of a Bernoulli distribution, with a proportion $p_0 < 0.5$ of group A in the whole city, the theoretical trajectory maximizing the distortion coefficient is that consisting in first aggregating exclusively all individuals from group A, and then all individuals of group $\overline{A}$. In this case, $\mathcal{N}$ may be explicitly computed as:

$$\mathcal{N} = -p_0 \log(p_0) + \int_{p_0}^{1} \left[\frac{p_0}{x} \log\left(\frac{1}{x}\right) + \frac{x - p_0}{x} \log\left(\frac{x - p_0}{x(1 - p_0)}\right) \right] dx \ . \tag{5.9}$$

In order to have a better grasp of *distortion coefficients* and of their normalization procedure, we simulated various examples, available in the supplementary information of [MO2]. Figure 5.8 illustrates what we previously called the most segregated theoretical configuration for a city with two groups. We highlighted the spatial unit from which segregation is the most acutely perceived, its corresponding Kullback-Leibler divergence trajectory, and the resulting map of distortion coefficients, normalized with respect to the distortion coefficient of this extreme unit.



(a) (b) (c)

Figure 5.8: A two-group simulated scenario with extreme segregation. (a) The actual city configuration: the green group represents 25% of the population. (b) The Kullback-Leibler divergence trajectories for all units. The trajectory of the most segregated unit (top right corner in (a) is plotted in a solid green and orange line: a walker starting from this unit would first meet all the green units, and the orange ones only afterwards). (c) The normalized distortion coefficients map computed as in Equation 5.7 and normalized with respect to $\mathcal{N}$ as defined in Equation 5.9, for $p_0 = 0.25$. The color scale goes from 0 (dark blue) to 1 (dark red).

Figure 5.9 also illustrates some simulated configurations for a population with two groups, less and less segregated, and the evolution of their normalized distortion coefficients maps. Again, the normalization is always done with respect to the extreme trajectory in the extreme configuration in Figure 5.8.

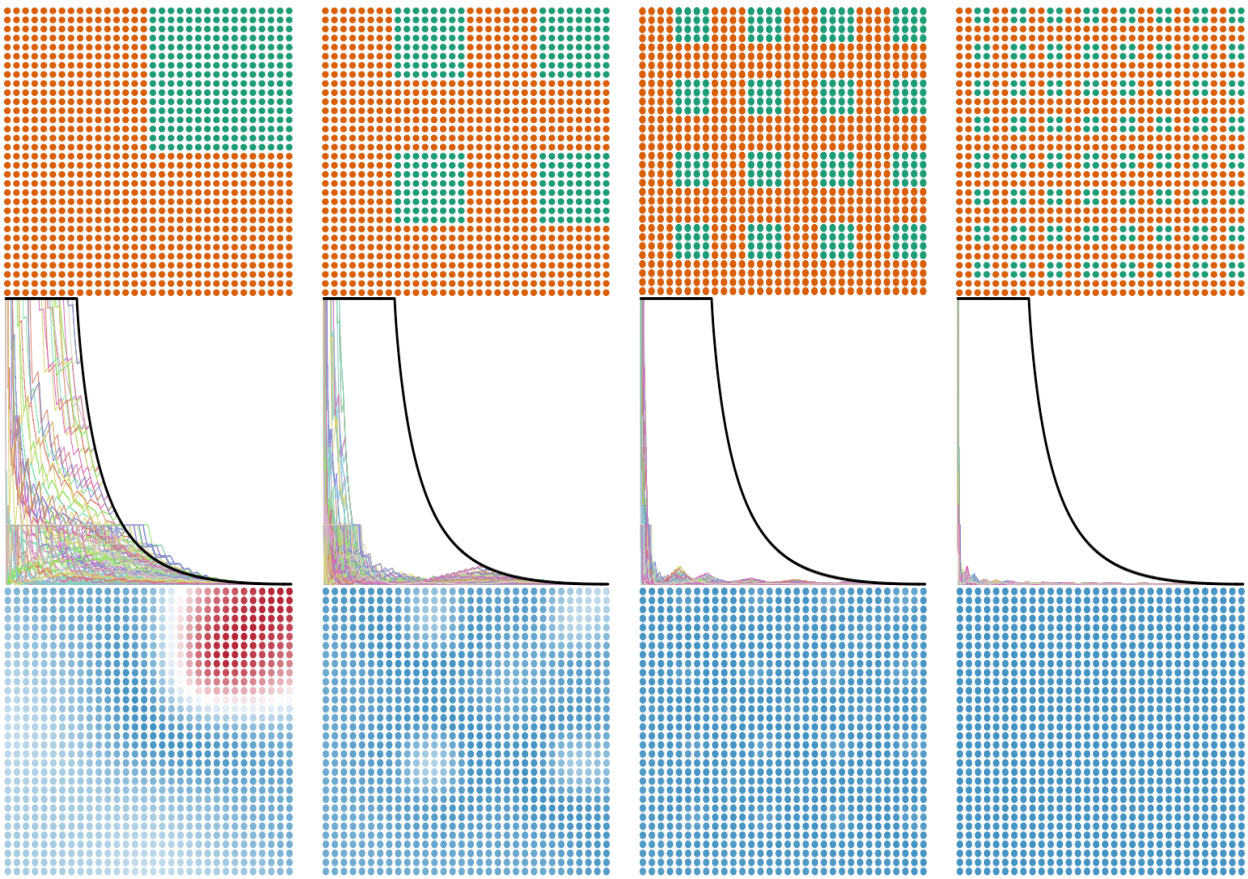


Figure 5.9: Various simulated city configurations. Top row: Actual city configurations, with the green group (25%) more or less segregated. Middle row: The Kullback-Leibler divergence trajectories. The trajectory of the most segregated unit (top right corner in the first scenario) is added on all graphs as a solid black line. Bottom row: The normalized distortion coefficients map computed as in Equation 5.7 and normalized with respect to $\mathcal{N}$ as defined in Equation 5.9, for $p_0 = 0.25$. The color scale goes from 0 (dark blue) to 1 (dark red).

I will briefly go back to the real data and illustrate the normalized distortion coefficients distribution and map

for the social housing rate in Paris. Figures 5.10 and 5.11 are excerpted from [MO24]. As one may easily see, most spatial units have low distortion coefficients, with a median value of 2.5% and 75% of the units below 5%. The tail of the distribution is however rather heavy, and when looking into details, one may identify a *hotspot* of segregation around the Champs-Elysées. Furthermore, the spatial unit with the highest distortion coefficient, 10%, hence maximum segregation in the actual city configuration, is situated Place Vendôme, while the spatial unit reaching the lowest distortion coefficient, 0.1%, is situated in the 10th district of the city, the Hôpital Saint Louis neighborhood, a historically well mixed area. One may also note the two orders of magnitude between the two.

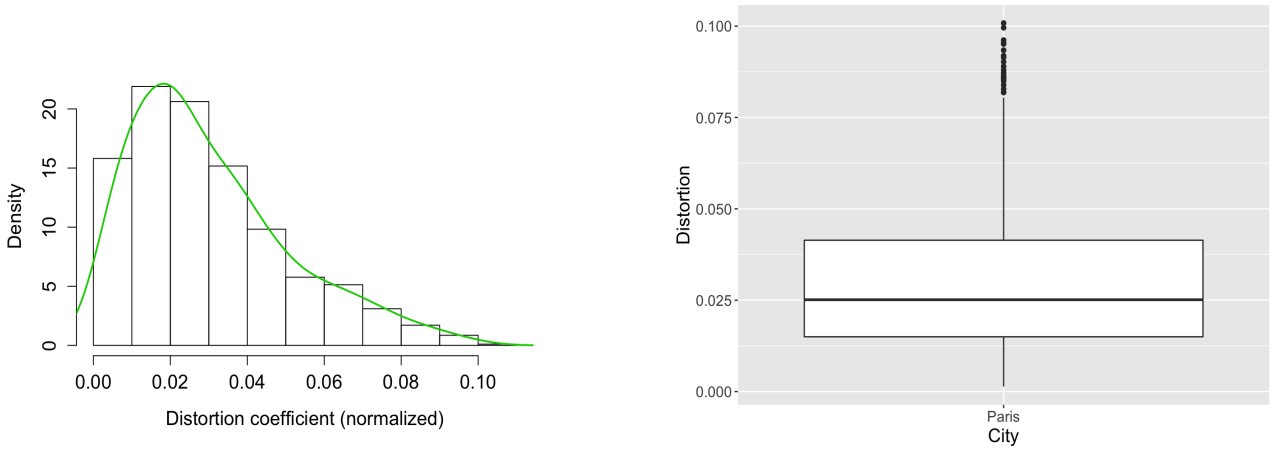


Figure 5.10: The empirical distribution of the normalized distortion coefficients computed from the Kullback-Leibler divergence trajectories on the social housing rate in Paris.

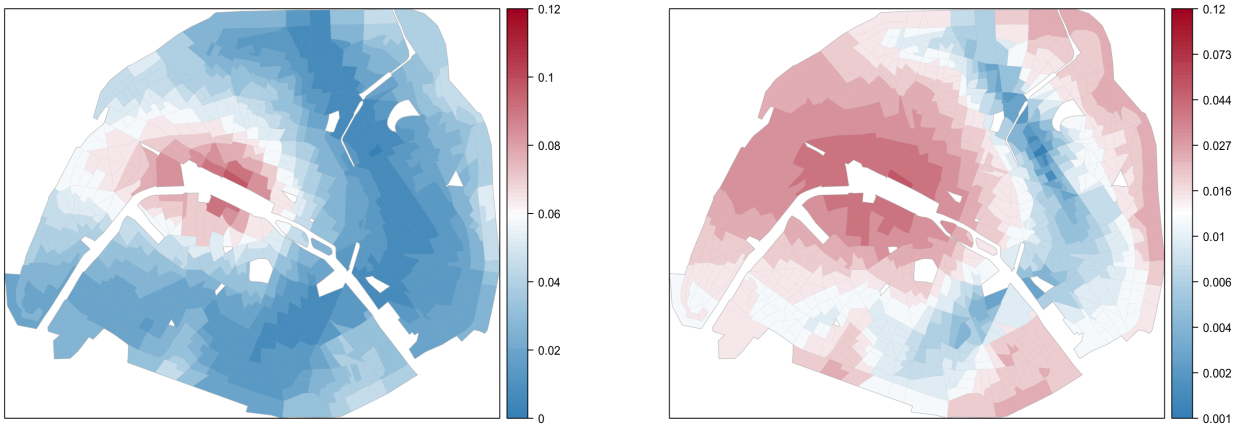


Figure 5.11: Spatial distribution of the normalized distortion coefficients computed from the Kullback-Leibler divergence trajectories on the social housing rate in Paris. Left: linear color scale; right: log color scale (color scales were normalized with respect to the maximum computed value of the coefficients).

5.3.4 Testing segregation with respect to a null "unsegregated" model

Assessing segregation may also be considered as a statistical hypothesis testing question: one may wish to compare the actual configuration with respect to some *null* model, which should be the *unsegregated* city. In the case of a well-mixed city, the probability distribution should be uniform across space. This means that when aggregating spatial units according to some spatial proximity rule, as in the computation of individual trajectories above, no sign of this rule should be reflected by the sequence of aggregates: everything would be as if one was drawing units at random in an urn. Any deviation from such a random behavior will be a sign that an underlying structure leads to biased shuffling of the units.

We used this concept of null model twice in our contributions, as I will briefly explain next.

Measuring deviations from random sequences using the ballot theorem We first used the notion of well-mixed city as a reference null model in [MO3], for Bernoulli-distributed variables of interest. More specifically, we placed ourselves in the particular case where the data should be known at individual level, hence where the trajectories in Section 5.3.1 were computed by aggregating individuals one by one instead of already basic level aggregated spatial units.

Let me denote by $N_{pop} = n_{i,1:N}$ the total size of the population in the city, and p_0 the proportion of group A in the whole city. For each individual i, consider $X_i = 1$ if i belongs to group A and 0 otherwise, and let

$$S_i(n) = \sum_{j=1}^{n} X_i(j) \ , n = 1, ..., N_{pop} \ , \tag{5.10}$$

be the number of individuals belonging to group A, among the n individuals spatially closest to i.

According to the ballot theorem [174], if two candidates have been submitted to a vote and candidate A wins with a final score $\rho > 0.5$, then the probability that during the counting the number of votes for A is at all times greater than the number of votes for $\overline{A}$ is equal to $2\rho - 1$.

The natural interpretation of the ballot theorem in terms of urban trajectories is the following. In one builds aggregated trajectories of social housing rates as in Section 5.3.1 and Figure 5.6, what is the probability that she will always see a minority of social housing, given that the fraction of social housing in the whole city is $p_0 = 0.18$? According to the ballot theorem, this probability should be equal to 0.64. In other words, if the city were perfectly mixed, 64% of the individual trajectories should always stay below 0.5. In actuality 85% of the trajectories are always beneath 0.5.

Obviously, one needs to test whether this deviation from the value in the ballot theorem is statistically significant or is due to a random sampling effect. By defining the following set of variables,

$$Y_i = \begin{cases} 1 & , \text{ if the trajectory of the } i\text{-th individual, } (\frac{S_i(n)}{n})_{n=1,...,N_{pop}}, \text{ stays always below 0.5} \\ 0 & , \text{ otherwise.} \end{cases} \tag{5.11}$$

$Y_1, ..., Y_N$ are distributed according to a Bernoulli with parameter q and one will test $H_0 : q = 0.64$ (the city is well mixed) versus $H_1 : q \neq 0.64$. By supposing the N_{pop} trajectories independent and identically distributed, one immediately rejects H_0 with a p-value of order 10^{-16}. This p-value is explained by the fact that a perfectly well-mixed city is an extreme unrealistic case. Nevertheless, it provides an absolute scale, very useful for reliable comparisons.

Let me also point out here the fact that some notations in this paragraph may seem quite messy, and this is partly due to the fact that the ballot theorem holds for individual count data, whereas the spatial data we have and on which we carried the analysis was available at a basic aggregated level. The trajectories we computed, such as those illustrated in Figure 5.6, may be seen as sampled in the individual-based trajectories, had one access to individual data. This situation could be formalized by introducing a subordination step, which will allow to properly write the aggregated spatial units trajectories in Section 5.3.1 as sampled from particular forms of individual trajectories (and random walks) in Equation 5.10. This formalization is not immediate however, and its study is part of Anoine Luquiaud's PhD work, under the supervision of Julien Randon-Furling and as a joint work following our paper [MO3].

A null model based on permutations Another way of using the concept of *null* or *"unsegregated"* model and of testing how different the actual configuration of the city is with respect to it, is to use random permutations and empirical testing. If one supposes that the null model corresponds to a completely random configuration of the city, she may aggregate the spatial units of the city by considering a large number random permutations, and thus obtain a set of artificial trajectories. For each of these trajectories, a normalized distortion coefficient may be then computed, as defined in Section 5.3.3. In practice, one thus has N normalized distortion coefficients corresponding to the actual configuration, where N is the number of spatial units in the city, with an empirical distribution as illustrated in Figure 5.10 for the social housing rate in Paris, and B normalized distortion coefficients, $N \ll B$, corresponding to B artificial trajectories in a perfectly well-mixed city.

As an illustration, Figure 5.3.4 excerpted from [MO2] displays the two distributions, the actual one, and the one of the artificial distortion coefficients, obtained by random permutations. This example is on data related to ethnic mixing in the Los Angeles area, but the methodology used for the computations is exactly the same as for the Parisian data. The histogram of the normalized distortion coefficients corresponding to the actual city, as well as their estimated density, are represented together with the mean and the 95% confidence interval (with a Gaussian hypothesis) of the normalized distortion coefficients computed for random trajectories. One may easily see that the hypothesis of a completely random distribution is rejected. There is a spatial structure in

the data, which means there is segregation. Again, a perfectly well-mixed city is a very unrealistic hypothesis, but here again this situation constitutes a reliable basis for comparisons.

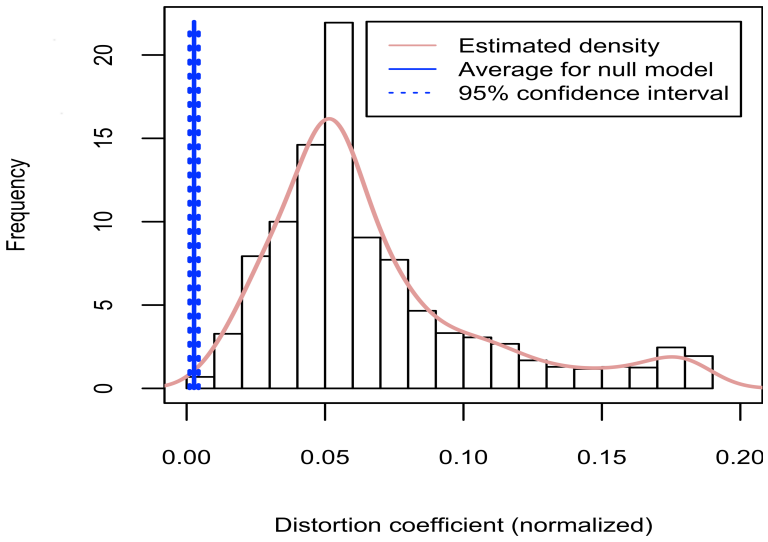


Figure 5.12: Histogram and estimated density (pink solid line) of normalized distortion coefficients in Los Angeles area. The solid blue vertical line corresponds to the average value of distortion coefficients in a *null model* obtained by random (spatial) permutations. Dashed lines indicate a 95% confidence interval around the average in the null model.

5.4 A conclusion and some (numerous) perspectives

Upon reflection, this chapter is somehow both an unexpected and a natural continuation of the work presented in the previous ones. Over the years, as I got more and more interested in complex data in humanities and social sciences, I chose to focus on its temporal aspects. However, space is just one step away from time, and at some point its integration becomes an evidence. All the more so with the definition we proposed of individual trajectories in the city, which encompasses a notion of displacement, be it temporal or spatial.

To summarize this chapter, I would say that we have investigated various exploratory techniques, which appear to bring insightful perspectives in the study of residential segregation. The notions of focal distances and distortion coefficients in particular attracted a great deal of attention from the geographers' community, also thanks to the very fruitful collaboration with William Clark. My intention is to continue with the two projects described in Sections 5.2 and 5.3, with a parallel development of the methodologies, and with some bridges established mostly by applications on real data.

With the colleagues involved in the study based on self-organizing maps, we are currently working on mixing together the information on the structure in the data, brought by the volatility index of the clustering, with the information on the spatial structure, brought by the correlation between geographical distances and distances in the mapping space. We believe the combination of the two could produce a global, powerful tool, for assessing multidimensional spatial patterns. We have already obtained some theoretical results for synthetic configurations, and performed various simulations. Some situations, such as concentric patterns, appear to be problematic for the moment, and we are investigating possible solutions for overcoming this.

The research project on the urban trajectories mushroomed in various directions, with a broad spectrum of theoretical, methodological, and applications possibilities, and I think this mainly comes both from the novelty of the notions we introduced, and from the enthusiasm they raised among colleagues. I will list some of our ongoing reflections and plans of work.

First, I will say that working on the urban trajectories is not conceptually or computationally straightforward and one needs to specify the exact framework she deals with. In practice, residential data may come in two forms: either as individual information with a geo-localization label (this is the case for some Northern European countries), or as already aggregated basic spatial units with more or less homogeneous population densities, such as blocks or tracts in the US, IRIS in France, ... Working on individual data is more interesting from a theoretical point of view, since the framework is more adapted to using random-walks modeling techniques and deriving properties of the trajectories such as first passage times, sojourn times etc, and also since all the information is

available at the finest possible level. In practice, the computation of these trajectories, and more particularly their storage are rapidly troublesome for massive data, and one needs to use parallel computing on the one hand, and meaningful indices for summarizing the trajectories on the other. Working on aggregated spatial units is more interesting in practice since the complexity of the algorithms and their computational burden are much lower, but this gives rise to other issues, such as the bias in the definition of the spatial units, or the fact that trajectories should now be modeled as subordinated random walks and variants, which is more difficult to handle from a theoretical perspective.

Trajectories may thus be explored either by extracting meaningful information and indices from them, or by clustering them and combining clustering with feature selection criteria. First, one may wish to extract other *scales of interest* from the trajectories, besides focal distances and distortion coefficients. With Mark Handcock and Julien-Randon Furling, we have started very recently to look at other indices starting from random walks and expected Kullback-Leibler divergences. A clustering of the trajectories combined with a regularization term such as a lasso-penalty or a fused-lasso penalty (by an the analogy of the trajectories with a time series) may both allow to extract some *critical instants* or *implicit scales* of the trajectories thanks to feature selection, and also to produce homogeneous groups of trajectories, that could be further used for practical analysis and interpretation. With Jean-Charles Lamirel, we've already published a very preliminary study [MO24] which uses growing neural gas as an extension, more flexible alternative to SOM, and a feature importance criteria computed as an analogy with the F-measure. We obtained very encouraging results, consistent with other exploratory results. This specific collaboration will be strengthened starting with the Fall of 2020, since we'll be co-supervising together a PhD student with a background in quantitative geography and demography. The aspects related to lasso-penalized clustering and variants will be investigated during Alex Mourer's PhD on variable importance, funded by Safran and co-supervised with Marie Chavent (INRIA, Université de Bordeaux).

Another theoretical aspect to be investigated relates to the theoretical properties of a null model built by considering random permutations. Establishing the theoretical properties of this model is not straightforward, and depends on what the urban trajectories are exactly made of. In the case of trajectories of proportions for Bernoulli distributions, one may approximate the corresponding trajectories by generalized Brownian bridges [175, 176] and transform the problem into a first-passage one for Brownian motion [177]. The more general case of multinomial distributions and trajectories computed from divergences between distributions, does not seem amenable to the same techniques, and poses an interesting mathematical challenge for future research. This work is done in collaboration with Julien Randon-Furling and Antoine Luquiaud, who is also currently investigating, as part of his PhD, the extension of the ballot theorem to subordinated processes, so that the test I presented in Section 5.3.4 has a complete theoretical foundation.

Other extensions that we will seek to develop include exploring models (Schelling-type models, Markov random fields and extensions, Poisson point processes) able to reproduce stylized aspects of the observed urban trajectories. Some of this work should be done in collaboration with Julien Randon-Furling and William Clark, some in collaboration with Radu Stoica, with whom we plan to co-supervise a master's internship and possibly a PhD thesis starting with the Summer of 2020.

In terms of data and applications, we are quite ambitious and hopefully will come up with meaningful studies in urban geography. With William Clark and Julien Randon-Furling, we are currently investigating a broader empirical context. We aim at characterizing ethnic mixing in the US for multiple cities and at various time instants, and contribute to a fine understanding of segregation, of how does it change in time or across space, of how micro and/or macroscopic interventions on the dynamics or on the spatial distribution may lead to significant changes in the trajectories, where these changes are the most prominent, and how do they propagate over space. This will allow me to complete the circle, and go back to the study of the temporal issues. Chris Fowler, who will be visiting the SAMM team as an invited professor during the Spring of 2020, will most certainly join us in these reflections.

Eventually, we wish to extend our research beyond residential segregation, and tackle more complex data such as geo-mobility, social networks, ... Indeed, one is not restricted to computing trajectories by aggregating spatially according to a nearest neighbor rule. One may also consider transportation times, for example, or "personal distances" based on information where one lives, works, goes out ... These personal trajectories transform the spatial network of blocks into personal networks reflecting personal paths and lives in the city. For some, the initial spatial network may turn into a small-world network, thanks to an ease of mobility across the different parts of the city. For others, personal trajectories may well be extremely concentrated around their local block. We'll start to investigate these aspects by using Cuban mobile phone and census data, within our PHC project and Dafne Garcia's PhD, and in collaboration with Alejandro Lage, a statistical physicist at Universidad de la Habana.

Since we wish our methods to be known and used beyond the computational statistics or machine learning community, we do make an effort into making the methodologies and indices described in this chapter available

in a user-friendly format. Dafne Garcia de Armas is currently finishing the implementation and documentation of a Python script that will allow one to compute urban trajectories and distortion coefficients when feeding census data together with shape files for the spatial units. An R-package should be also released next Summer and regularly updated with our latest findings.

I will end this chapter by mentioning a project that I'm currently putting up, that is important to me both from a professional and a personal perspective. As I said in the introduction of this chapter, my colleagues and I participated in a data-challenge on migrants integration in Europe and I am currently a fellow of the Migrations Convergence Institute. This institutional framework, as well as all the work done on residential segregation these last two years, and the associated perspectives for assessing segregation from a broader point of view, encouraged me to consider the issue of the Romanian diaspora, from a research perspective. Indeed, Romania has been facing a massive emigration phenomenon these last years, the official statistics being unable to assess how many inhabitants (10%? 20%? 40%? - depending on the sources) are living abroad. The phenomenon is a complex one from many points of view - including the integration in the arrival country -, and relatively few studies (other than qualitative) are available. Together with a group of Romanian sociologists and geographers, in Bucharest, Zurich and Paris, we wish to draw attention on this phenomenon, and use our knowledge in statistics, machine learning, humanities and social sciences for building a dedicated team and raise the necessary funding for theoretical and applied research on this topic.

Chapter 6

Conclusion: where do I stand and what next?

"*The idea of the future, pregnant with an infinity of possibilities,*
is thus more fruitful than the future itself,
and this is why we find more charm in hope than in possession,
in dreams than in reality. "
H. Bergson

Before concluding this manuscript, I would like to take two more pages for wrapping up. Although each of the previous chapters contains a conclusion and particularly the related perspectives about the works presented throughout them, I will briefly summarize here my current research situation and the associated on-going and forthcoming projects, in order to give a global picture.

My research crystallized as a particular interest in using data for understanding transitions, changes and rhythms in time-evolving phenomena, and especially those at study in humanities and social sciences. At present, all my projects are designed and being carried out in interdisciplinary settings. In terms of collaborations and practical applications, I am currently developing three lines of research.

First, I am interested in temporality issues for historical data, whether one speaks about data extracted from (possibly digitized) traditional archives, or from new sources of knowledge such as Wikipedia. The underlying questions are thus related to the rhythms of past and present societies, both in terms of organization and production of knowledge. Most of this work is being done in collaboration with historian researchers from the PIREH team (Julien Alerini, Stéphane Lamassé), some colleagues from the Archeology Department being also interested to join us. The techniques and methods we are developing are based on hidden Markov models and change-point detection algorithms and take into account the various specificities of the data, univariate or multivariate time series (integer-valued, bounded, with missing values or unequally sampled, with uncertainties on the time instants). Thanks to a research grant from the University scientific board, we will be able to fund a summer internship next year, and move forward at least the data analysis phase for the Wikipedia study. We are at present perfectly aware of the importance and of the interest of this topic, as well as of the scarcity of internal human and financial ressources, hence we are now convinced by the necessity of applying for a national or European grant on these issues.

Second, I aim at studying temporal networks and understanding their underlying structures, whether in terms of clusters of vertices or temporalities. During the last two years, I have been hosted by the MaIAGE team at INRA, where I had the opportunity of participating to a ANR (French National Research Agency) funded project. This project, called Cadence, is investigating the behavior of the entire French cattle-trading network over more than fifteen years, with a daily frequency, and focusing particularly on epidemics spreading issues. Being in charge with one of the working packages, and together with Elisabeta Vergu (principal investigator), I looked at how relational self-organizing maps (bagged version) could be adapted and trained on this kind data, while co-supervising an internship during the summer of 2019. A first paper is currently being written and will be submitted soon, but we have a significant amount of empirical results that need to be organized, summarized, and hopefully further published. This work raised a series of questions related to the definition and computation of temporal dissimilarities, the adaptation of relational SOM to sparse and not-symmetric dissimilarities, the

meaningfulness of the resulting clusters with respect to an epidemics spreading, ... The data being investigated in this project being very large (in terms of number of vertices, temporal span and frequency), its analysis has been quite tedious, and at some times even frustrating. It is only now, that we have finally obtained some clustering structures, that it becomes interesting to compare our approach with other methods and test the sensitivity with respect to epidemics spreading.

Third, my work is driven by the question of detecting spatial patterns in multivariate geo-localized data, and, more generally, of assessing the homogeneity of the data at all scales. The first results were motivated by the issue of residential segregation, but the framework we introduced could be easily generalized to networks, mobility data, ... Eventually, on a longer term, I would reflect upon these issues both from a temporal and spatial perspective, hence relating the notions of change-points and transitions in the temporal analysis to those of frontiers when space steps in the game. This work is being currently developed in collaboration with colleagues in geography (William Clark, UCLA), statistical physics (Alejandro Lage, University of Havana, Julien Randon-Furling, Université Paris 1), statistics (Cécile Hardouin, Université Paris X, Radu Stoica, Université de Nancy, Mark Handcock, UCLA) and machine learning (Aurélien Hazan, Université Paris Créteil, Jean-Charles Lamirel, LORIA Nancy). I am currently co-supervising the PhD of a Cuban student, and several interns and PhD should join this project next Summer and Fall. The conclusion in Chapter 5 contains a detailed summary of our ongoing projects.

Let me mention at this point that I am also currently co-supervising two industrial PhD's, the first related to aircraft engine health-monitoring (funded by Safran Aircraft Engines, and co-supervised with Marie Chavent, INRIA, Bordeaux), and the second to phytopharmacovigilence data mining (funded by Anses, and co-supervised with Fabrice Rossi, Université Paris Dauphine). The first project is aimed at studying variable importance for clustering and also at selecting meaningful variables, using regularization techniques and on data stemming from test benches. The second project is focused on change-point and anomaly detection when monitoring polluting substances concentrations in various environments. My interest in this kind of collaborations is double. On the one hand, the temporal issue is present in the data and the data poses several challenges in terms of modeling approaches which may be translatable to other fields of applications. On the other hand, I am particularly interested in how private companies handle data, how are the new techniques and methodologies disseminating, what are their current needs in data mining ... This last point is highly related to my teaching activities also. I have been teaching data mining and data-science for more than ten years now, and I have also been in charge of a data-science Master's program for six years. As a trainer of future data-scientists, a close collaboration with industry has always been valuable and important to me, and I've always tried to step out of our academic world (and bubble) and get some feed-back from the industry, on both student training and dissemination of scientific results.

As far as I see things, my interests and my activity – both in terms of supervision and research – are now well established, with ongoing, mid-term and long-term projects, and with consistent collaborations. If I were to identify some impediments, I will most probably name two. The first is related to the delicate balance one should find between research, student supervision, teaching, and the ever increasing amount of administrative tasks. The latter is to be considered also from an evermore digitized society perspective: new technologies are not always simplifying our lives in our universities, in some cases they are just a mean to add more bureaucracy and invent new chores. The second issue I see raises the question of scarcity of human and financial ressources, and of an imposed way of doing "planned" research while writing plenty of time-consuming applications for grants. I am not particularly convinced this is the ideal environment for a level-headed reflection, but I assume I have some more years to come ahead of me to try to figure this out.

Short curriculum vitae

of Madalina Olteanu, born on September 14th, 1978 (aged 41), French and Romanian citizenships.

Academic positions

2007 - ...	Associate professor in Applied Mathematics at Université Paris 1 Panthéon Sorbonne and member of the SAMM research team.
2017 - 2019	Sabbatical leave and research position at MaIAGE, INRA, Jouy en Josas.
2007	Post-doctoral researcher at MIAJ, INRA, Jouy en Josas.
2005 - 2007	Teaching and research assistant (ATER) at Université Paris 1 Panthéon Sorbonne.
2002 - 2005	Teaching assistant at Université Paris 5 in the Statistics department Research associate at Université Paris 1 Panthéon Sorbonne, SAMOS team.

Education

2006	PhD in Applied Mathematics, Université Paris 1 Panthéon Sorbonne, with Honors (*Très honorable*).
2002	M.S. in Applied Mathematics, Université Paris 1 - Université Paris 7, with Honors (B).
2001	B.S. in Applied Mathematics, University of Bucharest, with Honors (TB).
1997	Bachelor's degree in Mathematics and Physics, with Honors (TB).

Student Supervision

A. PhD students

Alex Mourer (2019 - ...). Co-supervised with Marie Chavent (INRIA and Université de Bordeaux), and Jérôme Lacaille (Safran Aircraft Engines). This PhD is funded by Safran Aircraft Engines, through a CIFRE grant.

Dafne Garcia de Armas (2018 - ...). Co-supervised with Sira Allende Alonso (Universidad de la Habana, Cuba).

Clément Laroche (2018 - ...). Co-supervised with Fabrice Rossi, Université Paris Dauphine. This PhD is funded by ANSES.

Cynthia Faure (2015-2018). Co-supervised with Jean-Marc Bardet (SAMM, Université Paris 1), and Jérôme Lacaille (Safran Aircraft Engine). This PhD was funded by Safran Aircraft Engines. Thesis defended in September 2018, on *Change-point detection and identification of causes in turbojet engine operation during flights and test benches.*

B. Postdoc's

September 2013 - March 2014 Co-supervision with Nathalie Vialaneix (MIA, INRA, Toulouse) of J. Boelaert's postdoc *Implementing the R package SOMbrero.*

C. Interns

March-August 2018 Co-supervision with Elisabeta Vergu (MaIAGE, INRA) of Kevin Pame's internship (Master of Statistics, Université Paris 5), *Self-organizing clustering in a large dynamical network.*

March-July 2013 Co-supervision with Nathalie Vialaneix (MIA, INRA, Toulouse) of Laura Bendhaiba's internship (GIS Engineer, PolyTech'Lille), *Implementing the SOMbrero R-package.*

June-October 2011 Supervision of James Ridgway's internship (Master of Statistics, ENSAE), *Hidden Markov models for integer-valued times series.*

June-September 2008 Supervision of Sébastien Massoni's internship (Master of Econometrics, Université Paris 1), *Analyzing career paths with self-organizing maps for categorical time series.*

D. PhD jurys

December 2017 Examiner, PhD defended by I. Gorynin (Université Paris Saclay), co-supervised by W. Pieczynski and E. Monfrini, *Bayesian state estimation in partially observed Markov processes.*

April 2014 Referee, PhD defended by E. Garcia-Garaluz (Universidad de Malaga, Spain), co-supervised by M. Atencia and G. Joya, *Mathematical modeling of dynamical systems in epidemiology.*

September 2010 Referee, PhD defended by A. Sorjamaa (Aalto University School of Science and Technology, Finland), supervised by A. Lendasse, *Assessment of spatio-temporal data bases: time series prediction and missing value problem.*

E. Other

2006 - First year (M1 MAEF, *Applied mathematics in economics and finance*) and second year (M2 TIDE, *Data-science for business decision and analytics*) master thesis supervision. On average, two dissertations per year.

2007 - Apprentice supervision in the master program M2 TIDE. On average, three students per year.

Fundings and grants

2019-2020 *Socio-spatial dynamics in large cities: from new mathematical models to new multidisciplinary perspectives.* Project funded by Partenariats Hubert Curien (PHC), Campus France.
Co-coordinator with Sira Allende Alonso (Universidad de la Habana, Cuba).

2019-2022 *Variable importance and variable selection for high-dimensional clustering in an industrial context*
Project funded by Safran Aircraft Engines.
Co-coordinator with Marie Chavent (INRIA, Bordeaux).

2018-2020 *Temporality as a multidisciplinary approach*
Project funded by Université Paris 1 Panthéon Sorbonne.
Co-coordinator with Stéphane Lamassé (Université Panthéon Sorbonne).

2018-2020 *Migrations, segregation(s), integration(s)*
Project funded by Université Panthéon Sorbonne.
Co-coordinator with Julien Randon-Furling (Université Panthéon Sorbonne).

2018-2021 *Feasibility study: phytopharmacovigilance (PPV) data mining. Creation of a tool for detecting emersions for the PPV.* Project funded by ANSES.
Co-coordinator with Fabrice Rossi (Université Paris Dauphine).

2017-2021 *CADENCE: spread of epidemic processes on dynamical networks of animal movements with application to cattle in France*
ANR grant. Coordinator : Elisabeta Vergu (MaIAGE, INRA).

2014-2017 *Change-point detection and identification of causes in turbojet engine operation during flights and test benches.* Project funded by Safran Aircraft Engines.
Co-coordinator with Jean-Marc Bardet (Université Panthéon Sorbonne).

2013-2014 *Temporality as a multidisciplinary approach*
Project funded by Université Paris 1 Panthéon Sorbonne.
Co-coordinator with Stéphane Lamassé (Université Panthéon Sorbonne).

Scientific activities

Reviewer for scientific journals Annals of Applied Statistics, Journal of Applied Probability/Advances in Applied Probability, Computational Statistics and Data Analysis, Journal of Multivariate Analysis, Sociological Methods and Research, IEEE Transactions on Neural Networks and Learning Systems, Neural Computation and Applications, Neural Processing Letters, Neural Networks, International Journal of Forecasting, Urban Studies, Computational Economics, Neurocomputing.

Technical committees of international conferences IJCNN (International Joint Conference on Neural Networks), NC^2(New Challenges in Neural Computation), LaCOSA II (International Conference on Sequence Analysis and Related Methods), WSOM+ (Workshop on Self Organizing Maps), ESANN (European Symposium on Artificial Neural Networks), IWANN (International Work Conference on Artificial Neural Networks), ICOR (International Conference on Operations Research), MASHS (Modèles et apprentissage en sciences humaines et sociales).

Conference and seminar organization

December 2019 *Challenging the time issue in modeling complex data from humanities and social sciences*, Special session for the CM Statistics, London, UK.

June 2017 *12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM+)*, Co-organized with J-C. Lamirel (LORIA, Nancy) and M. Cottrell (Université Paris 1 Panthéon Sorbonne) in Nancy, France.

2014 *Temporality: perceptions and analysis*, a series of six interdisciplinary seminars. Co-organized with S. Lamassé (Université Paris 1 Panthéon Sorbonne).

June 2012 *Modeling and statistical learning in humanities and social sciences (MASHS 2012)*, Co-organized with N. Vialaneix (MIA, INRA) and M. Cottrell (Université Paris 1).

October 2011 *Trajectories'11* workshop. Co-organized with P. Rousset (CEREQ, France).

September 2010 *Statistical learning*, Special session for the "Journées MAS de la SMAI" meeting, Bordeaux, France.

June 2009 *Modeling and data analysis in biomedical systems*, Special session at the International Workshop on Artificial Neural Networks (IWANN 2009). Co-organized with E. Garcia (Universidad de Malaga, Spain) in Salamanca, Spain.

June 2008 *Modeling and statistical learning in humanities and social sciences (MASHS 2008)*, Co-organized with P. Gaubert (Université Paris Créteil).

Teaching activities

Invited lectures

October 2019 *Hypothesis testing.* Lecture for undergraduate level (24h). University of Habana, Cuba.

April 2017 *Hypothesis testing.* Lecture for undergraduate level (24h). University of Habana, Cuba.

March 2016 *Time series: change point detection and regime switching.* Lecture for graduate level (12h). University of Habana, Cuba.

July 2015 *Time series: change point detection and regime switching.* Lecture for graduate level (10h). Summer school in Applied Mathematics, organized by the French-Romanian CNRS European Laboratory. Sinaia, Romania.

March 2015 *Machine learning.* Lecture for graduate level (8h). Spring school *Interdisciplinary College*, Gunne, Germany.

May 2009 *Data mining and machine learning.* Lecture for graduate level (10h). University of Alger, Algeria.

Graduate level

2003-2016
2019-... On average 90 hours per year. Lectures and tutorials in *Data mining, Statistics, Statistical Learning, Quantitative techniques and applied statistics, Statistics with SAS, Survival Analysis, Time series*, for master students in Economics, Applied Mathematics, Data-science, Bioinformatics. Université Paris 5, Université Paris 1 Panthéon Sorbonne, Université de Rouen.

2011-2014 *Quantitative methods in humanities and social sciences* (6h per year). Lecture for PhD students in History and Archeology, Université Paris 1 Panthéon Sorbonne.

2008-2009 *Data mining with SAS* (18h per year). Lecture for PhD students in Economics, Université Paris 1.

2007-2016 Training in mathematics for Economics teaching-degree candidates (*CAPES SES, Agrégation SES*). 54h per year. Université Paris 1 Panthéon Sorbonne.

Undergraduate level

2003-2016
2019-... On average 60 hours per year. Lectures and tutorials in *Statistics, Probability theory, Linear algebra, Time series*, for students in Economics, Applied Mathematics, Biology. Université Paris 5, Université Paris 1 Panthéon Sorbonne.

Administrative responsibilities and tasks

2018-2020 Elected member of the University senate (Université Paris 1 Panthéon Sorbonne).

2016-... Member of the board of the apprentice training center CFA Formasup Paris-Ile de France.

2013-2016 Elected member of the Academic Council of the University (Université Paris 1 Pantéon Sorbonne).

2012-2018 Director of the data-science Master program *TIDE (Techniques d'information et de décision dans l'entreprise)*, Université Paris 1 Panthéon Sorbonne. Sandwich program.

2012-2016 Elected member of the Education Committee of the University (Université Paris 1 Pantéon Sorbonne) and member of its permanent commission.

2010-... Elected member of the research committee of the Mathematics and Computer Science department, Université Paris 1 Panthéon Sorbonne.

2010-... Member of several selection committees for Associate Professorship (Université Paris 1 Panthéon Sorbonne - 2010, 2011, 2015, 2019; Université Toulouse 3 - 2015; Université Paris 13 - 2009).

List of publications

A. Journal articles

Articles cited in the manuscript

MO1. M. Olteanu, A. Hazan, M. Cottrell, J. Randon-Furling, "Multidimensional urban segregation: towards a neural network measure", 2019, *Neural Computing and Applications*, Vol. 31(6), p.1-13.

MO2. M. Olteanu, J. Randon-Furling, W. A. Clark, "Segregation through the multiscalar lens", 2019, *Proceedings of the National Academy of Sciences*, Vol. 116(25), p.11250-12254.

MO3. J. Randon-Furling, M. Olteanu, A. Lucquiaud, "From urban segregation to spatial structure detection", 2018, *Environment and Planning B: Urban Analytics and City Science*.

MO4. M. Cottrell, M. Olteanu, F. Rossi, N. Villa-Vialaneix, "Self-organizing maps, theory and applications", 2018, *Revista Investigacion Operacional*, Vol. 39(1), p. 1-22.

MO5. J. Alerini, M. Olteanu, J. Ridgway, "Markov and the Duchy of Savoy: segmenting a century with regime-switching models", 2017, *Journal de la Société Française de Statistique*, Vol. 158(2), p.89-117.

MO6. J. Mariette, M. Olteanu, N. Villa-Vialaneix, "Efficient interpretable variants of online SOM for large dissimilarity data", 2017, *Neurocomputing*, Vol. 225, p. 31-48.

MO7. M. Olteanu, N. Villa-Vialaneix, "Using SOMbrero for clustering and visualizing graphs", 2015, *Journal de la Société Française de Statistique*, Vol. 156(3), p.95-119.

MO8. M. Olteanu, N. Villa-Vialaneix, "On-line relational and multiple relational SOM", 2015, *Neurocomputing*, Vol. 147, p.15-30.

MO9. M. Cottrell, M. Olteanu, F. Rossi, J. Rynkiewicz, N. Villa-Vialaneix, "Neural networks for complex data", 2012, *KI - Künstliche Intelligenz*, Vol.26, p.373-380.

MO10. M. Olteanu, J. Rynkiewicz, "Asymptotic properties of autoregressive regime-switching models", 2012, *ESAIM P&S*, Vol. 16, p.25-47.

MO11. M. Olteanu, J. Rynkiewicz, "Asymptotic properties of mixture-of-experts models", 2011, *Neurocomputing*, Vol.74(9), p.1444-1449.

MO12. M. Olteanu, J. Rynkiewicz, "Estimating the number of components in a mixture of multilayer perceptrons", 2008, *Neurocomputing*, Vol.71(7-9, p.1321-1329.

Other articles

MO13. M. Olteanu, V. Nicolas, B. Schaeffer, C. Denys, A. Missoup, J. Kennis and C. Laredo, "Nonlinear projection methods for visualizing Barcode data and application on two data sets", 2013, *Molecular Ecology Resources*, Vol.13(6), p.976-990.

MO14. F. Austerlitz, K. Bleakley, M. Olteanu, O. David, C. Laredo, R. Leblois, B. Schaeffer, M. Veuille, "DNA barcode analysis : comparing phylogenetic and statistical classification methods", 2009, *BMC Bioinformatics*, Vol. 10(14).

MO15. M.T. Boyer-Xambeu, G. Deleplace, P. Gaubert, L. Gillard, M. Olteanu, "The periodization of the international bimetallism : 1821-1873", 2007, *Revista Investigacion Operacional*, Vol.28(2), p.143-156.

MO16. M. Olteanu, "A descriptive method to evaluate the number of regimes in a switching autoregressive model", 2006, *Neural Networks*, Vol.19, p. 963-972.

MO17. B. Maillet, M. Olteanu, J. Rynkiewicz, "Caractérisation des crises financières à l'aide de modèles hybrides (HMC-MLP)", 2004, *Revue d'économie politique*, Vol. 4, p. 489-506.

B. Editorials

MO18. M. Cottrell, M. Olteanu, J. Rouchier, N. Villa-Vialaneix, Editorial of the special issue of RNTI - MASHS 2011/2012 : Modèles et Apprentissage en Sciences Humaines et Sociales. Revue Des Nouvelles Technologies De l'Information, SHS-1, p. 97–110, 2012.

C. Book chapters

MO19. J. Gravier, L. Nahassia, D. Michel, N. Verdier, M. Olteanu, "Processus - Trajectoire", To appear in 2020, In A. Bretagnolle, P. Brun, M.-V. Ozouf-Marignier, L. Sanders, N. Verdier (Eds.), *Les mots-clefs des systèmes de peuplement dans le temps long : regards croisès*, Editions de la Sorbonne.

MO20. M. Olteanu, J. Alerini, "Quelques réflexions sur la périodisation en histoire", 2019, In G. Bonnot, St. Lamassé (Ed.), *Dans les dédales du web: Historiens en territoires numériques*, Editions de la Sorbonne, p.57-85.

MO21. E. Garcia Garaluz, M. Atencia, G. Joya, M. Olteanu, "Modeling dengue epidemics with autoregressive switching Markov models", 2009, In A. Cabestany, F. Sandoval, A. Prieto, J.M. Corchado (Eds.), *Bio-inspired systems: Computational and Ambient Intelligence*, p. 886-892.

MO22. M.-T. Boyer-Xambeu, G. Deleplace, P. Gaubert, L. Gillard, M. Olteanu, "Kolonnen maps and time-series algorithms: a clear convergence", 2008, In J.R. Rabunal, J. Dorado, A. Pazos (Eds.), *Encyclopedia of Artificial Intelligence*.

MO23. M.-T. Boyer-Xambeu, G. Deleplace, P. Gaubert, L. Gillard, M. Olteanu, "Mixing Kohonen algorithm, Markov switching model and detection of multiple change-points : an application to monetary history", In F. Sandoval, A. Preto, J. Cabestany, M. Grana (Eds.), *Computational and Ambient Intelligence*.

D. International conferences with peer-reviewed proceedings

Proceedings cited or related to the manuscript

MO24. M. Olteanu, J.-C. Lamirel, "When clustering the multiscalar fingerprint of the city reveals its segregation patterns", 2019, In A. Vellido, K. Gibert, C. Angulo, and J. D. Martin Guerrero (Eds.), *Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization (Proceedings of WSOM+ 2019)*, Springer International Publishing, p. 140–149.

MO25. M. Olteanu, J. Randon-Furling, W. Clark, "Spatial analysis in high resolution geo-data", 2019, In M. Verleysen (Ed.), *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2019)*, p. 559-564.

MO26. J. Alerini, M. Cottrell, M. Olteanu, "Hidden Markov models for time series of continuous proportions with excess zeros", 2017, In I. Rojas, G. Joya and A. Catala (dir.), *Advances in Computational Intelligence. 14th International Work-Conference on Artificial Neural Networks, IWANN 2017. Proceedings, Part II*, New York, Springer, p. 198-209.

MO27. M. Cottrell, M. Olteanu, J. Randon-Furling, A. Hazan, "Multidimensional urban segregation: an exploratory case study", 2017, *12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM 2017+)*, IEEE Xplore.

MO28. J. Mariette, M. Olteanu, F. Rossi, N. Villa-Vialaneix, "Accelerating stochastic kernel SOM", 2017, In M. Verleysen (Ed.), *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2017)*, p. 269-274.

MO29. M. Olteanu, N. Villa-Vialaneix, "Sparse online self-organizing maps for large relational data", 2016, In E. Merényi, M. J. Mendenhall, & O.D.P. (Eds.), *Advances in Self-organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2016)*, Springer International Publishing, Vol. 428, p. 27-37.

MO30. N. Bourgeois, M. Cottrell, S. Lamassé, M. Olteanu, "Search for meaning through the study of co-occurrences in texts", 2015, In I. Rojas, G. Joya and A. Catala (Eds.), *Advances in Computational Intelligence (Proceedings of IWANN 2015), Lecture Notes in Computer Science*, Springer, p. 578-591.

MO31. L. Bendhaiba, J. Boelaert, M. Olteanu, N. Villa-Vialaneix, "SOMbrero: an R package for numeric and non-numeric self-organizing maps", 2014, In Th. Villmann, F.-M. Schleif, M. Kaden, M. Lange, *Advances in Self-organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2014)*, Berlin Springer Verlag, p. 219-228.

MO32. J. Mariette, M. Olteanu, N. Villa-Vialaneix, "Bagged kernel SOM", 2014, In Th. Villmann, F.-M. Schleif, M. Kaden, M. Lange, *Advances in Self-organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2014)*, Berlin Springer Verlag, p. 45-54.

MO33. C. Cierco-Ayrolles, M. Olteanu, N. Villa-Vialaneix, "Multiple kernel self-organizing maps", 2013, In M. Verleysen (Ed.), *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2013)*, p. 83-88.

MO34. S. Massoni, M. Olteanu, N. Villa-Vialaneix, "Which dissimilarity is to be used when extracting typologies in sequence analysis? A comparative study", 2013, In I. Rojas, G. Joya and A. Cabestany (Eds.), *Advances in Computational Intelligence (Proceedings of IWANN 2013), Lecture Notes in Computer Science*, Springer, p. 69-79.

MO35. M. Cottrell, M. Olteanu, N. Villa-Vialaneix, "Online relational SOM for dissimilarity data", 2012, In P. Estevez, J. Principe, P. Zegers (Eds.), *Advances in Self-organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2012)*, Springer, p. 13-22.

MO36. M. Olteanu, J. Ridgway, "Hidden Markov models for time series of counts with excess zeros", 2012, In M. Verleysen (Ed.), *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2012)*, p. 133-138.

MO37. M. Olteanu, J. Rynkiewicz, "Asymptotic properties of mixture-of-experts models", 2010, In M. Verleysen (Ed.), *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2010)*, p. 207-212.

MO38. S. Massoni, M. Olteanu, P. Rousset, "Career-path analysis using optimal matching and self-organizing maps", 2009, In J. Principe, R. Miikkulainen (Eds.), *Advances in Self-organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2009), Lecture Notes in Computer Science*, Springer, p. 154-162.

MO39. S. Massoni, M. Olteanu, P. Rousset, "Analyse des trajectoires d'insertion professionnelle avec un algorithme de Kohonen pour données catégorielles", 2009, *Proceedings of MASHS 2009 (Modelling and leArning in Social and Human Sciences)*.

MO40. M. Olteanu, J. Rynkiewicz, "Estimating the number of components of a mixture autoregressive model", 2007, *Proceedings of the European Symposium on Time Series Prediction (ESTSP 2007)*, p. 143-154.

MO41. M. Olteanu, J. Rynkiewicz, "Estimating the number of components in a mixture of multilayer perceptrons", 2007, In M. Verleysen (Ed.), *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2007)*, p. 403-408.

Other proceedings

MO42. M. Cottrell, C. Faure, J. Lacaille, M. Olteanu, "Anomaly detection for bivariate signals", 2019, In I. Rojas, G. Joya and A. Catala (Eds.), *Advances in Computational Intelligence (Proceedings of IWANN 2019), Lecture Notes in Computer Science*, Springer, p. 162-173.

MO43. M. Cottrell, C. Faure, J. Lacaille, M. Olteanu, "Detection of abnormal flights using fickle instances in SOM maps", 2019, In A. Vellido, K. Gibert, C. Angulo, and J. D. Martin Guerrero (Eds.), *Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization (Proceedings of WSOM+ 2019)*, Springer International Publishing, p. 120-129.

MO44. C. Faure, M. Olteanu, J.-M. Bardet, J. Lacaille, "Using self-organizing maps for clustering and labelling aircraft engine data phases", 2017, *12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM 2017+)*, IEEE Xplore.

MO45. J.-M. Bardet, C. Faure, J. Lacaille, M. Olteanu, "Comparison of three algorithms for parametric change-point detection", 2016, In M. Verleysen (Ed.), *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2016)*, p. 89-94.

MO46. S. Massoni, M. Olteanu, P. Rousset, "Career-path analysis using drifting Markov models (DMM) and self-organizing maps", 2010, *Proceedings of MASHS 2010 (Modelling and leArning in Social and Human Sciences)*, p. 171-179.

MO47. Ch. Bouveyron, S. Girard, M. Olteanu, "Supervised classification of categorical data with uncertain labels for DNA barcoding", 2009, In M. Verleysen (Ed.), *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2009)*, p. 22-34.

MO48. M. Olteanu, "Revisiting linear and nonlinear methodologies for time series prediction: application to ESTSP'08 competition data", 2008, *Proceedings of the European Symposium on Time Series Prediction (ESTSP 2008)*, p. 139-148.

MO49. M. Olteanu, "A descriptive method to evaluate the number of regimes in a switching autoregressive model", 2005, *Proceedings of the International Workshop on Self-Organizing Maps (WSOM 2005)*, p. 259-266.

MO50. B. Maillet, M. Olteanu, J. Rynkiewicz, "Nonlinear analysis of shocks when financial markets are subject to changes in regime", 2004, In M. Verleysen (Ed.), *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2004)*, p. 87-92.

E. Invited conferences

The speaker is in bold font.

MO51. **S. Lamassé**, M. Olteanu, "Detecting the evolution phases of a text production", In *12th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2019)*, London, UK, December 14-16, 2019.

MO52. J Alerini, **M. Olteanu**, "Markov and the Dukes of Savoy: A temporal analysis of the Piedmontese-Savoyard legislation", In *12th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2019)*, London, UK, December 14-16, 2019.

MO53. **M. Olteanu**, N. Vialaneix, J. Marietta, G. Beaunée, K. Pame, E. Vergu, "Clustering complex data with kernel and relational SOM. An application to cattle trading networks", In *The 22nd Conference of the Romanian Society of Probability and Statistics (SPSR 2019)*, Bucharest, Romania, May 10-11, 2019.

MO54. **J. Randon-Furling**, M. Olteanu, W. Clark, "The distorted city - capturing the complexity of perceived segregation", In *ECSR Workshop*, Florence, Italy, April 11, 2019.

MO55. M. Olteanu, **J. Randon-Furling**, "Assessing segregation in complex networks through a multi-focal approach", In *11th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2018)*, Pisa, Italy, December 14-16, 2018.

MO56. **M. Olteanu**, K. Pame, G. Beaunée, C. Bidot, E. Vergu, "Clustering and visualizing large cattle-trading networks using self-organizing maps", In *11th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2018)*, Pisa, Italy, December 14-16, 2018.

MO57. M. Olteanu, **J. Randon-Furling**, W. Clark, "Migrations and segregation in European cities", In *D4I Joint Research Center - European Commission Workshop*, Brussels, Belgium, November 28, 2018.

MO58. **M. Olteanu**, P. Rousset, "Using big data in order better to visualise the competences associated with jobs: the birth of an experimental project", In *CEREQ Workshop*, Marseille, France, September 27, 2017.

MO59. **M. Olteanu**, N. Villa-Vialaneix, "Using SOMbrero for clustering and visualizing complex data", In *9th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2016)*, Sevilla, Spain, December 9-11, 2016.

MO60. **M. Cottrell**, M. Olteanu, F. Rossi, N. Villa-Vialaneix, "Theoretical and applied aspects of the self-organizing maps", In E. Merényi, M. J. Mendenhall, & O.D.P. (Eds.), *Advances in Self-organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2016)*, Springer International Publishing, Vol. 428, p. 27-37, Houston, US, January 6-8, 2016.

MO61. J. Alerini, **M. Olteanu**, "Markov et les ducs de Savoie : analyse de la temporalité du droit piémonto-savoyard", In *Workshop on "Data-mining in human and social sciences : issues and perspectives"*, IHP, Paris, France, April, 2012.

MO62. **M. Olteanu**, "Etude des trajectoires d'insertion professionnelle à l'aide de chaines de Markov non-homogènes et de cartes auto-organisées", In *Ateliers d'Ouverture, CEREQ*, Marseille, France, December, 2011.

F. Other conferences

1. W. Clark, M. Olteanu, J. Randon-Furling, "Segregation beyond scale: assessing the individual perceptions of migrant residential segregation", In *European Colloquium on Theoretical and Quantitative Geography (ECTQG)*, Mondorf-Les-Bains, Luxembourg, September 5-9, 2019.
2. M. Olteanu, J. Randon-Furling, W. Clark, "Focal distances and distortion coefficients: assessing the individual perception of multiscalar segregation", In *useR! 2019*, Toulouse, France, July 9-12, 2019.
3. M. Olteanu, K. Pame, G. Beaunée, C. Bidot, E. Vergu, "Clustering and visualizing large cattle-trade networks using relational self-organizing maps", In *51èmes Journées de Statistique de la SFdS (JDS 2019)*, Nancy, France, June 3-7, 2019.
4. W. Clark, J. Randon-Furling, M. Olteanu, "A new method for analyzing ethnic mixing: Southern California as an exemplar", In *American Association of Geographers Meeting (AAG 2019)*, Washington DC, US, April 3-7, 2019.
5. M. Olteanu, J. Randon-Furling, "Converging to the city: a myriad trajectories", In *Conference in Complex Systems (CCS 2018)*, Thessaloniki, Greece, September 23-28, 2018.
6. W. Clark, J. Randon-Furling, M. Olteanu, "A new method for analyzing ethnic mixing: studies from Southern California", In *23rd International Conference on Computational Statistics (CompStat 2018)*, Iasi, Romania, August 28-31, 2018.
7. W. Clark, J. Randon-Furling, M. Olteanu, "A new method for analyzing ethnic mixing: studies from Southern California", In *European Network for Housing Research (ENHR) Conference*, Uppsala, Sweden, June 26-29, 2018.
8. M. Olteanu, G. Beaunée, C. Bidot, K. Pame, E. Vergu, "Clustering and visualizing large cattle-trading networks using self-organizing maps", In *International School and Conference on Network Science (NetSci 2018)*, Paris, France, June 11-15, 2018.
9. M. Olteanu, J. Randon-Furling, "Multiscalar socio-spatial dynamics in the city", In *BIFI International Conference on Complex Systems*, Zaragoza, Spain, February 6-8, 2018.
10. M. Olteanu, G. Beaunée, C. Bidot, C. Laredo, E. Vergu, "Using SOMbrero for clustering and visualizing large cattle-trading networks", In *BIFI International Conference on Complex Systems*, Zaragoza, Spain, February 6-8, 2018.
11. M. Olteanu, J. Randon-Furling, "Analyzing spatial dissimilarities via effective time-series", In *International Work-Conference on Time Series Analysis (ITISE 2017)*, Granada, Spain, September 18-20, 2017.
12. J. Alerini, M. Olteanu, "Exploring a century of Savoy history using hidden-Markov models with Beta-inflated distributions", In *International Work-Conference on Time Series Analysis (ITISE 2017)*, Granada, Spain, September 18-20, 2017.
13. J.-M. Bardet, C. Faure, J. Lacaille, M. Olteanu, "Design aircraft engine bivariate data phases using change-point detection methods and self-organizing maps", In *International Work-Conference on Time Series Analysis (ITISE 2017)*, Granada, Spain, September 18-20, 2017.
14. M. Olteanu, N. Villa, "Using SOMbrero for clustering and visualizing complex data", In *International Workshop on Operations Research (IWOR 2017)*, Habana, Cuba, March 14-17, 2017.
15. P. Alpi, M. Olteanu, J. Yilmaz, "Unsupervised learning for panel data", In *LaCOSA II International conference on sequence analysis and related methods*, Lausanne, Switzerland, June 8-10, 2016.
16. M. Olteanu, N. Villa-Vialaneix, "Classification et visualisation de graphes avec SOMbrero", In *4èmes Rencontres R*, Grenoble, France, June 24-26, 2015.
17. M. Olteanu, N. Villa-Vialaneix, "Multiple dissimilarity SOM for clustering and visualizing graphs with node and edge attributes", *International Conference on Machine Learning (ICML 2015), Workshop FEAST*, Lille, France, July 10, 2015.

18. M. Olteanu, N. Villa-Vialaneix, "Self-organizing maps for clustering visualization of bipartite graphs", In *46èmes Journées de Statistique de la SFdS (JDS 2014)*, Rennes, France, June 2-6, 2014.
19. L. Bendhaiba, M. Olteanu, N. Villa-Vialaneix, "SOMbrero : cartes auto-organisatrices stochastiques pour l'intégration de données décrites par des tableaux de dissimilarités", In *2èmes Rencontres R*, Lyon, France, June 27-28, 2013.
20. C. Cierco-Ayrolles, M. Olteanu, N. Villa-Vialaneix, " Carte auto-organisatrice pour graphes étiquetés", In *Ateliers Fouille de Grands Graphes, Extraction et Gestion des Connaissances (EGC 2013)*, Toulouse, France, January 29, 2013.
21. M. Olteanu, J. Ridgway, "DiscreteTS : two hidden-Markov models for time series of count data", In *1ères Rencontres R*, Bordeaux, France, July 2-3, 2012.
22. J. Alerini, M. Olteanu, J. Ridgway, "Modélisation de séries temporelles à valeurs entières par des modèles autorégressifs à changements de régime", In *44èmes Journées de Statistique de la SFdS (JDS 2012)*, Brussels, Belgium, May 21-25, 2012.
23. J. Alerini, M. Olteanu, J. Ridgway, "An application of regime-switching models to historical data", In *10th International Conference on Operations Research (ICOR 2012)*, Havana, Cuba, March 6-9, 2012.
24. C. Laredo, V. Nicolas, M. Olteanu, "On the use of self-organizing maps for the representation of Barcoding data : an application to *Hylomyscus* data", In *4th International Barcode of Life Conference*, Adelaide, Australia, November 28 - December 3, 2011.
25. S. Massoni, M. Olteanu, P. Rousset, "Career-path analysis using drifting Markov models (DMM) and self-organizing maps", In *9th International Conference on Operations Research (ICOR 2010)*, Havana, Cuba, February, 2010.
26. M. Olteanu, J. Rynkiewicz, "Consistency of the Bayesian Information Criterion for a class of mixture autoregressive models", In *The 11th Conference of the Romanian Society of Probability and Statistics (SPSR 2008)*, Bucharest, Romania, April, 2008.
27. M.T. Boyer-Xambeu, G. Deleplace, P. Gaubert, L. Gillard, I. Kammoun, M. Olteanu, "Combining Markov switching models and the detection of change-points with the SOM algorithm to explain a temporal process", In *8th International Conference on Operations Research (ICOR 2008)*, Havana, Cuba, March, 2008.
28. F. Austerlitz, K. Bleakley, M. Olteanu, O. David, C. Laredo, R. Leblois, B. Schaeffer, M. Veuille, "Comparing phylogenetic and statistical classification methods for DNA barcoding", In *2nd International Barcode of Life Conference*, Taipei, Taiwan, September, 2007.
29. M.T. Boyer-Xambeu, G. Deleplace, P. Gaubert, L. Gillard, M. Olteanu, "The periodization of the international bimetallism: 1821-1873", In *7th International Conference on Operations Research (ICOR 2006)*, Havana, Cuba, March, 2006.
30. M. Olteanu, J. Rynkiewicz, "Estimating the number of regimes in an autoregressive model with Markov switching", In *7th International Conference on Operations Research (ICOR 2006)*, Havana, Cuba, March, 2006.
31. B. Maillet, M. Olteanu, J. Rynkiewicz, "Nonlinear analysis of shocks when financial markets are subject to changes in regime", In *Colloque Econométrie des Valeurs Mobilières (AEA 2004)*, Paris, France, April, 2004.
32. M. Olteanu, J. Rynkiewicz, "Prévision d'un indice des chocs du marché avec des modèles hybrides HMM-MLP", In *Approches Connexionnistes en Sciences Economiques et de Gestion (ACSEG 2003)*, Nantes, France, November 2003.
33. T.M. Hoang, M. Olteanu, "Coupled self-organizing maps for the bi-clustering of microarray data", In *Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP 2003)*, Protaras, Cyprus, October, 2003.

G. Software

1. N. Vialaneix, J. Mariette, M. Olteanu, F. Rossi, L. Bendhaiba, J. Boelaert, *SOMbrero: SOM Bound to Realize Euclidean and Relational Outputs*, R package version 1.2-4., 2019.

Bibliography

[1] F.A. Duboin. *Raccoolta per ordine di materie delle leggi cioè editti, manifesti, ecc., pubblicati negli stati della Real Casa di Savoia fino all'8 dicembre 1798.* Torino, 1818-1869.

[2] T. Couzin. Contribution piémontaise à la genèse de l'État italien. L'historicité de la « Raccolta per ordine di materie delle leggi » (1818-1868). *Bolettino Storico-Bibliografico Subalpino*, CVI:101–120, 2008.

[3] Ch. Truong, L. Oudre, and N. Vayatis. A review of change point detection methods. *CoRR*, abs/1801.00718, 2018.

[4] W. Zucchini, I.L. MacDonald, and R. Langrock. *Hidden Markov models for time series: an introduction using R.* Chapman and Hall/CRC, 2017.

[5] O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models (Springer Series in Statistics).* Springer-Verlag, Berlin, Heidelberg, 2005.

[6] J.D. Hamilton. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, 57(2):357–84, March 1989.

[7] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[8] L.E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

[9] T. Rydén. Em versus markov chain monte carlo for estimation of hidden markov models: a computational perspective. *Bayesian Anal.*, 3(4):659–688, 12 2008.

[10] A.J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, IT-13(2):260–269, April 1967.

[11] F. Braudel and E. Labrousse. *Histoire économique et sociale de la France.* Presses universitaires de France, Paris, 1977.

[12] N. Bonneuil. Temporalités en démographie historique. *Histoire & Mesure*, 6(1-2):137–148, 1991.

[13] A. Mathis and J.-Y. Grenier. Séries temporelles, structure et conjoncture : le prix du blé à l'époque moderne. *Histoire & Mesure*, 6(1):51–76, 1991.

[14] A. Guerreau. *Statistique pour historiens.* Ecole Nationale des Chartes, 2004.

[15] L. Sun. *Statistical methods for serially correlated zero-inflated proportions.* PhD thesis, Oregon State University, 2014.

[16] S.M. DeSantis and D. Bandyopadhyay. Hidden markov-models for zero-inflated poisson counts with an application to substance use. *Statistics in Medicine*, 30(14):1678–1694, 2011.

[17] E. Gassiat, A. Cleynen, and S. Robin. Inference in finite state space non parametric hidden markov models and applications. *Statistics and Computing*, pages 1–11, 2015.

[18] M.A. Al-Osh and A.A. Alzaid. First-order integer-valued autoregressive (inar(1)) process. *Journal of Time Series Analysis*, 8:261–275, 1987.

[19] M.A. Al-Osh and A.A. Alzaid. An integer-valued p-th order autoregressive (inar(p)) process. *Journal of Applied Probability*, 27(2):314–324, 1990.

[20] D. Jin-Guan and Li. Yuan. The integer-valued autoregressive (inar(p)) model. *Journal of Time Series Analysis*, 12(2):129–142, 1991.

[21] F. Steutel and K. van Harn. Discrete analogues of self-decomposability and stability. *The Annals of Probability*, 7:893–899, 1979.

[22] K.F. Wallis. Time series analysis of bounded economic variables. *Journal of Time Series Analysis*, 8(1):115–123, 1987.

[23] S. Ferrari and F. Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004.

[24] M.l Smithson and J. Verkuilen. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, 11(1):54, 2006.

[25] R. Ospina and S.L.P. Ferrari. A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, 56(6):1609 – 1623, 2012.

[26] A.B. Simas, W. Barreto-Souza, and A.V. Rocha. Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, 54(2):348 – 366, 2010.

[27] J. Alerini and M. Olteanu. Exploring a century of savoy history using hidden markov models with beta-inflated distributions. Preprint, 2017.

[28] C. Rosso. *Una burocrazia di antico regime. I segretari di Stato dei duchi di Savoia, I (1559-1637).* Torino, 1992.

[29] J. Glete. *War and the State in Early Modern Europe. Spain, the Dutch Republic and Sweden as Fiscal-Military States, 1500–1600.* Routledge, Londres, 2002.

[30] Ch. Tilly. *Coercion, capital, and European states, AD 990-1990.* B. Blackwell, Cambridge (Mass.), 1990.

[31] J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

[32] J. Rynkiewicz. *Modèles hybrides intégrant des réseaux de neurones artificiels à des modèles de chaînes de Markov cachées : application à la prédiction de séries temporelles.* PhD thesis, 2000. Thèse de doctorat dirigée par Cottrell, Marie Mathématiques appliquées Paris 1 2000.

[33] R. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

[34] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307 – 327, 1986.

[35] F. Rosenblatt. *Principles of neurodynamics: perceptrons and the theory of brain mechanisms.* Report (Cornell Aeronautical Laboratory). Spartan Books, 1962.

[36] R. A. Davis, D. Huang, and Y.-C. Yao. Testing for a change in the parameter values and order of an autoregressive model. *Ann. Statist.*, 23(1):282–304, 02 1995.

[37] E. Gombay. Change detection in autoregressive time series. *Journal of Multivariate Analysis*, 99(3):451–464, 2008.

[38] R. A. Jacobs, M. Jordan, S.J. Nowlan, G.E. Hinton, et al. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[39] M. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

[40] Ch. Francq and M. Roussignol. Ergodicity of autoregressive processes with markov-switching and consistency of the maximum-likelihood estimator. *Statistics: A Journal of Theoretical and Applied Statistics*, 32(2):151–173, 1998.

[41] V. Krishnamurthy and T. Ryden. Consistent estimation of linear and non-linear autoregressive models with markov regime. *Journal of Time Series Analysis*, 19(3):291–307, 1998.

[42] R. Douc, E. Moulines, and T. Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime. *The Annals of Statistics*, 32(5):2254–2304, 2004.

[43] B. E. Hansen. The likelihood ratio test under nonstandard conditions: Testing the markov switching model of gnp. *Journal of Applied Econometrics*, 7(S1):S61–S82, 1992.

[44] B.E. Hansen. Erratum: The likelihood ratio test under nonstandard conditions: Testing the markov switching model of gnp. *Journal of Applied Econometrics*, 11(2):195–198, 1996.

[45] B.E. Hansen. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, 64(2):413–430, 1996.

[46] R. Garcia. Asymptotic null distribution of the likelihood ratio test in markov switching models. *International Economic Review*, 39(3):763–788, 1998.

[47] R. Rios and L.-A. Rodriguez. Penalized estimate of the number of states in gaussian linear ar with markov regime. *Electron. J. Statist.*, 2:1111–1128, 2008.

[48] J. Henna. On estimating of the number of constituents of a finite mixture of continuous distributions. *Annals of the Institute of Statistical Mathematics*, 37(2):235–240, Dec 1985.

[49] K. Roeder. A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association*, 89(426):487–495, 1994.

[50] A.J. Izenman and Ch. Sommer. Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association*, 83(404):941–953, 1988.

[51] B.G. Lindsay. Moment matrices: Applications in mixtures. *The Annals of Statistics*, 17(2):722–740, 1989.

[52] D. Dacunha-Castelle and E. Gassiat. Testing the order of a model using locally conic parametrization: Population mixtures and stationary arma processes. *The Annals of Statistics*, 27(4):1178–1209, 1999.

[53] B.G. Leroux and M.L. Puterman. Maximum-penalized-likelihood estimation for independent and markov-dependent mixture models. *Biometrics*, 48(2):545–558, 1992.

[54] C. Keribin. Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 62(1):49–66, 2000.

[55] X. Liu and Y. Shao. Asymptotics for likelihood ratio tests under loss of identifiability. *The Annals of Statistics*, 31(3):807–832, 2003.

[56] E. Gassiat. Likelihood ratio inequalities with applications to various mixtures. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6):897 – 906, 2002.

[57] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[58] P. Doukhan. *Mixing: Properties and Examples*. Lecture Notes in Statistics. Springer New York, 2012.

[59] K. Fukumizu. Likelihood ratio of unidentifiable models and multilayer neural networks. *Ann. Statist.*, 31(3):833–851, 06 2003.

[60] N. A. Gershenfeld and A. S. Weigend. The future of time series. Technical report, Xerox Corporation, Palo Alto Research Center, 1993.

[61] A. S. Weigend, M. Mangeas, and A. Srivastava. Nonlinear gated experts for time series: Discovering regimes and avoiding overfitting. *International Journal of Neural Systems*, 6(04):373–399, 1995.

[62] W. Jiang and M. A. Tanner. On the asymptotic normality of hierarchical mixtures-of-experts for generalized linear models. *IEEE Transactions on Information Theory*, 46(3):1005–1013, 2000.

[63] E. Gassiat and C. Keribin. The likelihood ratio test for the number of components in a mixture with markov regime. *ESAIM: Probability and Statistics*, 4:25–52, 2000.

[64] J. Rynkiewicz. General bound of overfitting for mlp regression models. *Neurocomputing*, 90:106–110, 2012.

[65] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

[66] T. Kohonen. Analysis of a simple self-organizing process. *Biological cybernetics*, 44(2):135–140, 1982.

[67] T. Kohonen, editor. *Self-Organizing Maps*. Springer, volume 30 of springer series in information science edition, 1995.

[68] J.C. Fort, P. Letremy, and M. Cottrell. Advantages and drawbacks of the batch kohonen algorithm. In M. Verleysen, editor, *Proceedings of 10th European Symposium on Artificial Neural Networks (ESANN 2002)*, pages 223–230, Bruges, Belgium, 2002.

[69] T. Kohonen. *Self-Organizing Maps, 3rd Edition*, volume 30. Springer, Berlin, Heidelberg, New York, 2001.

[70] J. C. Fort. Som's mathematics. *Neural Netw.*, 19(6):812–816, July 2006.

[71] M. Cottrell, P. Letremy, and E. Roy. Analysing a contingency table with kohonen maps: A factorial correspondence analysis. In J. Mira, J. Cabestany, and A. Prieto, editors, *New Trends in Neural Computation*, pages 305–311, Berlin, Heidelberg, 1993. Springer Berlin Heidelberg.

[72] T. Kohohen and P.J. Somervuo. Self-organizing maps of symbol strings. *Neurocomputing*, 21:19–30, 1998.

[73] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[74] I. Schoenberg. Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert". *Annals of Mathematics*, 36:724–732, 1935.

[75] G. Young and A. Householder. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3:19–22, 1938.

[76] N. Krislock and H. Wolkowicz. *Handbook on Semidefinite, Conic and Polynomial Optimization*, volume 166 of *International Series in Operations Research & Management Science*, chapter Euclidean distance matrices and applications, pages 879–914. Springer, 2012.

[77] J.A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Information Science and Statistics. Springer, New York; London, 2007.

[78] L. Goldfarb. A unified approach to pattern recognition. *Pattern Recognition*, 17(5):575–582, 1984.

[79] E. Pękalska and R.P.W. Duin. *The Dissimilarity Representation for Pattern Recognition. Foundations and Applications*. World Scientific, 2005.

[80] S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[81] A. Abbott and J. Forrest. Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, 16:471–494, 1986.

[82] Y. Chen, E.K. Garcia, M.R. Gupta, A. Rahimi, and L. Cazzanti. Similarity-based classification: concepts and algorithm. *Journal of Machine Learning Research*, 10:747–776, 2009.

[83] M. Martín-Merino and A. Muñoz. Extending the som algorithm to non-euclidean distances via the kernel trick. In N.R. Pal, N. Kasabov, R.K. Mudi, S. Pal, and S.K. Parui, editors, *Neural Information Processing*, pages 150–157, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[84] B. Hammer, A. Hasenfuss, F. Rossi, and M. Strickert. Topographic processing of relational data. In Bielefeld University Neuroinformatics Group, editor, *Proceedings of the 6th Workshop on Self-Organizing Maps (WSOM 07)*, Bielefeld, Germany, September 2007.

[85] D. Mac Donald and C. Fyfe. The kernel self organising map. In *Proceedings of 4th International Conference on knowledge-based intelligence engineering systems and applied technologies*, pages 317–320, 2000.

[86] T. Fruchterman and B. Reingold. Graph drawing by force-directed placement. *Software, Practice and Experience*, 21:1129–1164, 1991.

[87] D. Harel and Y. Koren. Drawing graphs with non-uniform vertices. In *Proceedings of the Working Conference on Advanced Visualization Interfaces (AVI'02)*, pages 157–166, New York, NY, USA, 2002. ACM Press.

[88] D. Tunkelang. *A Numerical Optimization Approach to General Graph Drawing*. PhD thesis, School of Computer Science, Carnegie Mellon University, January 1999. CMU-CS-98-189.

[89] X.W. Wang and I.M. Miyamoto. Generating customized layouts. In F. Brandenburg, editor, *Graph Drawing*, volume 1027 of *Lecture Notes in Computer Science*, pages 504–515. Springer (Berlin/Heidelberg), 1996.

[90] R. Boulet, B. Jouve, F. Rossi, and N. Villa. Batch kernel SOM and related laplacian methods for social network analysis. *Neurocomputing*, 71(7-9):1257–1273, 2008.

[91] F. Rossi and N. Villa-Vialaneix. Optimizing an organized modularity measure for topographic graph clustering: a deterministic annealing approach. *Neurocomputing*, 73(7-9):1142–1163, 2010.

[92] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review, E*, 69:026113, 2004.

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

[94] A.J. Smola and R. Kondor. Kernels and regularization on graphs. In M. Warmuth and B. Schölkopf, editors, *Proceedings of the Conference on Learning Theory (COLT) and Kernel Workshop*, Lecture Notes in Computer Science, pages 144–158, 2003.

[95] F. Fouss, A. Pirotte, J.M. Renders, and M. Saerens. Random-walk computation of similarities between nodes of a graph, with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):355–369, 2007.

[96] R.I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th International Conference on Machine Learning*, pages 315–322, 2002.

[97] M.E.J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review, E*, 74(036104), 2006.

[98] D.E. Knuth. *The Stanford GraphBase: A Platform for Combinatorial Computing.* Addison-Wesley, Reading, MA, 1993.

[99] D. Combe, C. Largeron, E. Egyed-Zsigmond, and M. Géry. Getting clusters from structure data and attribute data. In *Proceedings of International Conference on Advances in Social Networks Analysis and Mining (IEEE/ACM)*, pages 731–733, 2012.

[100] N. Villa-Vialaneix, M. Olteanu, and C. Cierco-Ayrolles. Carte auto-organisatrice pour graphes étiquetés. In *Actes des Ateliers FGG (Fouille de Grands Graphes), colloque EGC (Extraction et Gestion de Connaissances)*, Toulouse, France, 2013.

[101] A. Abbott and Tsay. Sequence analysis and optimal matching methods in sociology. *Review and Prospect. Sociological Methods and Research*, 29(1):3–33", 2000.

[102] L. Wu. Some comments on "Sequence analysis and optimal matching methods in sociology, review and prospect". *Sociological Methods and Research*, 29(1):41–64, 2000.

[103] G. Reza, M.D. Nasir, I. Hamidah, and N. Norwti. A survey: clustering ensembles techniques. In *Proceedings of World Academy of Science, Engineering and Technology*, volume 38, pages 644–653, 2009.

[104] G. Cleuziou, M. Exbrayat, L. Martin, and J.H. Sublemontier. CoFKM: a centralized method for multi-view clustering. In *Proceedings of International Conference on Data Mining*, 2009.

[105] B. Zhao, J.T. Kwok, and C. Zhang. Multiple kernel clustering. In *Proceedings of the 9th SIAM International Conference on Data Mining (SDM)*, Sparks, Nevada, USA, 2009.

[106] J. Zhuang, J. Wang, S.C.H. Hoi, and X. Lan. Unsupervised multiple kernel clustering. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 20:129–144, 2011.

[107] X. Chen, Y. Ye, X. Xu, and J.Z. Huang. A feature group weighting method for subspace clustering of high-dimensional data. *Pattern Recognition*, 45:434–446, 2012.

[108] M. Lebbah, A. Chazottes, F. Badran, and S. Thiria. Mixed topological map. In M. Verleysen, editor, *Proceedings of the 13th European Symposium on Artificial Neural Networks (ESANN)*, pages 357–362, Bruges, Belgium, 2005.

[109] A. Rakotomamonjy, F.R. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[110] T. Villmann, H. Sven, and M. Kästner. Gradient based learning in vector quantization using differentiable kernels. In P.A. Estevez, J. Principe, P. Zegers, and G. Barreto, editors, *Advances in Self-Organizing Maps (Proceedings of WSOM 2012)*, volume 198 of *AISC (Advances in Intelligent Systems and Computing)*, pages 193–204, Santiago, Chile, 2012.

[111] D. Luenberger. *Linear and Nonlinear Programming.* Addison-Wesley, 1984.

[112] F. Bonnans. *Optimisation Continue.* Dunod, 2006.

[113] A. Gabadinho, G. Ritschard, N.S. Müller, and M. Studer. Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4):1–37, 2011.

[114] N.S. Müller, G. Ritschard, M. Studer, and A. Gabadinho. Extracting knowledge from life courses: clustering and visualization. In *10th International Conference DaWaK (Data Warehousing and Knowledge Discovery)*, volume 5182 of *Lecture Notes in Computer Science*, pages 176–185, Turin, Italy, September 2008. Berlin: Springer.

[115] L. Lesnard. Setting cost in optimal matching to uncover contempo-raneous socio-temporal patterns. *Sociological Methods et Research*, 38(3):389–419, 2010.

[116] G. Pölzlbauer. Survey and comparison of quality measures for self-organizing maps. In *Proceedings of the Fifth Workshop on Data Analysis (WDA'04)*, pages 67–82. Elfa Academic Press, 2004.

[117] D. Hofmann and B. Hammer. Sparse approximations for kernel learning vector quantization. In *ESANN*, 2013.

[118] C.T. Chu, S.K. Kim, Y.A. Lin, Y.Y. Yu, G. Bradski, A.Y. Ng, and K. Olukotun. Map-Reduce for machine learning on multicore. In J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NIPS 2010)*, volume 23, pages 281–288, Hyatt Regency, Vancouver, Canada, 2010.

[119] X. Chen and M.G. Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 24:1655–1684, 2014.

[120] S. del Rio, V. López, J.M. Beniítez, and F. Herrera. On the use of MapReduce for imbalanced big data using random forest. *Information Sciences*, 285:112–137, 2014.

[121] M. Bădoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. In J. Reif, editor, *Proceedings of the 34th annual ACM Symposium on Theory of Computing*, number 250-257, Montreal, QC, Canada, 2002. ACM New York, NY, USA.

[122] D. Yan, L. Huang, and M.I. Jordan. Fast approximate spectral clustering. In J. Elder, F. Soulié-Fogelman, P. Flach, and M. Zaki, editors, *Proceedings of the 15th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*, pages 907–916. ACM New York, NY, USA, 2009.

[123] A. Kleiner, A. Talwalkar, P. Sarkar, and M.I. Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014.

[124] N. Laptev, K. Zeng, and C. Zaniolo. Early accurate results for advanced analytics on mapreduce. In *Proceedings of the 28th International Conference on Very Large Data Bases*, volume 5 of *Proceedings of the VLDB Endowment*, Istanbul, Turkey, 2012.

[125] X. Meng. Scalable simple random sampling and stratified sampling. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, volume 28 of *JMLR: W&CP*, Georgia, USA, 2013.

[126] A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof. On-line random forests. In *Proceedings of IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1393–1400. IEEE, 2009.

[127] M. Denil, D. Matheson, and N. de Freitas. Consistency of online random forests. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, pages 1256–1264, 2013.

[128] C.K.I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems (Proceedings of NIPS 2000)*, volume 13, Denver, CO, USA, 2000. Neural Information Processing Systems Foundation.

[129] F. Rossi, A. Hasenfuss, and B. Hammer. Accelerating relational clustering algorithms with sparse prototype representation. In *Proceedings of the 6th Workshop on Self-Organizing Maps (WSOM 07)*, Bielefeld, Germany, 2007. Neuroinformatics Group, Bielefeld University.

[130] D. Hofmann, F.M. Schleif, B. Paaß en, and B. Hammer. Learning interpretable kernelized prototype-based models. *Neurocomputing*, 141:84–96, 2014.

[131] D. Hofmann, A. Gisbrecht, and B. Hammer. Efficient approximations of robust soft learning vector quantization for non-vectorial data. *Neurocomputing*, 147:96–106, 2015.

[132] A. Gisbrecht, B. Mokbel, and B. Hammer. The Nyström approximation for relational generative topographic mappings. In *NIPS workshop on challenges of Data Visualization*, Whistler BC, Canada, 2010.

[133] X. Zhu, A. Gisbrecht, F.M. Schleif, and B. Hammer. Approximation techniques for clustering dissimilarity data. *Neurocomputing*, 90:72–84, 2012.

[134] B. Hammer and A. Hasenfuss. Topographic mapping of large dissimilarity data sets. *Neural Computation*, 22(9):2229–2284, September 2010.

[135] J. Leskovec and J. Mcauley. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pages 539–547, 2012.

[136] F. Fouss, A. Pirotte, J. Renders, and M. Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):355–369, 2007.

[137] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008, 2005.

[138] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[139] S. Kumar, M. Mohri, and A. Talwalkar. Sampling techniques for the Nyström method. *Journal of Machine Learning Research*, 13:981–1006, 2012.

[140] J. Mariette and N. Villa-Vialaneix. Aggregating self-organizing maps with topology preservation. In E. Merényi, M.J. Mendenhall, and O'Driscoll P., editors, *Advances in Self-Organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2016)*, volume 428 of *Advances in Intelligent Systems and Computing*, pages 27–37, Houston, TX, USA, 2016. Springer International Publishing Switzerland.

[141] P. Drineas, M. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.

[142] A. Gittens and M.W. Mahoney. Revisiting the nystrom method for improved large-scale machine learning. *Journal of Machine Learning Research*, 28(3):567–575, 2013.

[143] M. Cottrell and P. Letrémy. Missing values : processing with the kohonen algorithm. In *Proceedings of Applied Stochastic Models and Data Analysis, ASMDA*, pages 489–496. Springer (Berlin/Heidelberg), 2005.

[144] M. Batty. *Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals*. The MIT Press, 2007.

[145] T.C. Schelling. Models of segregation. *The American Economic Review*, 59(2):488–493, 1969.

[146] T.C. Schelling. Dynamic models of segregation. *The Journal of Mathematical Sociology*, 1(2):143–186, 1971.

[147] I. Benenson, E. Hatna, and E. Or. From schelling to spatially explicit modeling of urban ethnic and economic residential dynamics. *Sociological Methods and Research*, 37(4):463–497, 5 2009.

[148] E. Hatna and I. Benenson. The schelling model of ethnic residential dynamics: Beyond the integrated - segregated dichotomy of patterns. *Journal of Artificial Societies and Social Simulation*, 15(1):6, 2012.

[149] E. Hatna and I. Benenson. Combining segregation and integration: Schelling model dynamics for heterogeneous population. *Journal of Artificial Societies and Social Simulation*, 18(4):15, 2015.

[150] S.F. Reardon and D. O'Sullivan. Measures of spatial segregation. *Sociological methodology*, 34(1):121–162, 2004.

[151] S. F. Reardon and G. Firebaugh. Measures of multigroup segregation. *Sociological methodology*, 32(1):33–67, 2002.

[152] M. R. Kramer, H. L. Cooper, C. D. Drews-Botsch, L. A. Waller, and C. R. Hogue. Do measures matter? comparing surface-density-derived and census-tract-derived measures of racial residential segregation. *International Journal of Health Geographics*, 9(1):29, Jun 2010.

[153] S. Hong, D. O'Sullivan, and Y. Sadahiro. Implementing spatial segregation measures in R. *PLOS ONE*, 9:1–18, 11 2014.

[154] O. D. Duncan and B. Duncan. A methodological analysis of segregation indexes. *American sociological review*, 20(2):210–217, 1955.

[155] D. WS Wong. Spatial indices of segregation. *Urban studies*, 30(3):559–572, 1993.

[156] R. L Morrill. On the measure of geographic segregation. In *Geography research forum*, volume 11, pages 25–36, 2016.

[157] M. J White. The measurement of spatial segregation. *American journal of sociology*, 88(5):1008–1018, 1983.

[158] M. Poulsen, R. Johnson, and J. Forrest. Plural cities and ethnic enclaves: introducing a measurement procedure for comparative study. *International Journal of Urban and Regional Research*, 26(2):229–243, 2002.

[159] S. Openshaw. *The modifiable areal unit problem.* University of East Anglia, 1984.

[160] S. F. Reardon, S. A. Matthews, D. O'Sullivan, B. A. Lee, G. Firebaugh, C. R. Farrell, and K. Bischoff. The geographic scale of Metropolitan racial segregation. *Demography*, 45(3):489–514, Aug 2008.

[161] F.F Feitosa, G. Camara, A. M. V. Monteiro, T. Koschitzki, and M. PS Silva. Global and local spatial indices of urban segregation. *International Journal of Geographical Information Science*, 21(3):299–323, 2007.

[162] P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1-2):17–23, 1950.

[163] G. Leckie, R. Pillinger, K. Jones, and H. Goldstein. Multilevel modeling of social segregation. *Journal of Educational and Behavioral Statistics*, 37:3–30, 02 2012.

[164] R. Louf and M. Barthelemy. Patterns of residential segregation. *PloS one*, 11(6):e0157476, 2016.

[165] W. A V Clark, E. Andersson, J. Östh, and B. Malmberg. A multiscalar analysis of neighborhood composition in Los Angeles, 2000–2010: A location-based approach to segregation and diversity. *Annals of the Association of American Geographers*, 105(6):1260–1284, 2015.

[166] C. Fowler. Segregation as a multiscalar phenomenon and its implications for neighborhood-scale research: the case of south seattle 1990–2010. *Urban geography*, 37(1):1–25, 2016.

[167] G. Tintori, A. Alessandrini, and F. Natale. Diversity, residential segregation, concentration of migrants: a comparison across eu cities. findings from the data challenge on integration of migrants in cities (d4i). JRC 115159, European Commission, Publications Office of the European Union, Luxembourg, 2018.

[168] E. Diday. The dynamic clusters method and optimization in non hierarchical-clustering. In R. Conti and A. Ruberti, editors, *5th Conference on Optimization Techniques Part I*, pages 241–258, Berlin, Heidelberg, 1973. Springer Berlin Heidelberg.

[169] E. de Bodt, M. Cottrell, and M. Verleysen. Statistical tools to assess the reliability of self-organizing maps. *Neural Networks*, 15, 8-9:967–978, 2002.

[170] N. Bourgeois, M. Cottrell, B. Deruelle, S. Lamassé, and P. Letrémy. How to improve robustness in kohonen maps and display additional information in factorial analysis: Application to text mining. *Neurocomputing*, 147:120–135, 2015.

[171] N. Bourgeois, M. Cottrell, St. Lamasse, and M. Olteanu. Search for Meaning Through the Study of Co-occurrences in Texts. In I. Rojas, G. Joya, and A. Catala, editors, *Advances in Computational Intelligence, IWANN 2015, Part II*, volume 9095 of *Lecture Notes in Computer Science*, pages 578–591, Palma de Mallorca, Spain, June 2015. Springer-Verlag.

[172] J. Östh, W.A.V. Clark, and B. Malmberg. Measuring the scale of segregation using k-nearest neighbor aggregates. *Geographical Analysis*, 47(1):34–49, 2015.

[173] E.K. Andersson, B. Malmberg, et al. Contextual effects on educational attainment in individualised, scalable neighbourhoods: Differences across gender and social class. *Urban Studies*, 52(12):2117–2133, 2015.

[174] J. Bertrand. Solution d'un probleme. *CR Acad. Sci. Paris*, 105(1887):369, 1887.

[175] B. Rosén. Limit theorems for sampling from finite populations. *Arkiv för Matematik*, 5(5):383–424, 1964.

[176] P. K. Sen. Finite population sampling and weak convergence to a Brownian bridge. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 85–90, 1972.

[177] P. Salminen and M. Yor. On hitting times of affine boundaries by reflecting Brownian motion and Bessel processes. *Periodica Mathematica Hungarica*, 62(1):75–101, 2011.