Journal de la Société Franaise de Statistique
Vol. 158 No. 2 (2017)

Markov and the Duchy of Savoy: segmenting a century with regime-switching models

Titre: Markov et les ducs de Savoie : périodisation d'un siècle d'histoire avec des modèles à changements de régime

Julien Alerini[1], Madalina Olteanu[2] and James Ridgway[3]

Abstract: Studying time is at the core of the historian's research. For the statistician, time is usually a supplementary parameter or a supplementary variable which the models he develops have to take into account. This manuscript is the result of a collaboration between historians and mathematicians, having time as the starting point of a joint work. We are interested in studying a specific time-series, reporting the production of law related to military logistics of the Duchy of Savoy, during the XVIth and XVIIth centuries. The expected outcome is a better understanding of the temporality and of the functioning of the state. Two models based on hidden Markov chains and taking into account the specificities of the data are introduced. They are then estimated on the historical data and provide interesting results, which either confirm the existing historical hypothesis on the subject or bring new insights on the studied period.

Résumé : Le temps est au coeur du travail de recherche de l'historien. Pour le statisticien, le temps n'est en général qu'un paramètre ou une variable supplémentaires, que les modèles développés doivent intégrer ou prendre en compte. Ce manuscrit est le fruit d'une collaboration entre historiens et mathématiciens, prenant le temps comme point de départ de ce travail commun. Nous étudions une série temporelle particulière, recensant la législation liée à la logistique militaire, émise par le Duché de Savoie pendant les XVIème et XVIIème siècles. Le résultat attendu est une meilleure compréhension de la temporalité et du fonctionnement de l'état. A cette fin, deux modèles basés sur des chaines de Markov cachées et prenant en compte les spécificités des données sont introduits. Ils sont ensuite estimés sur les données historiques et fournissent des résultats intéressants, qui soit confirment les hypothèses historiques existantes, soit apportent de nouvelles perspectives sur la période étudiée.

Keywords: Integer-valued time-series, hidden Markov models, autoregressive models, zero-inflated distributions.
Mots-clés : Séries temporelles à valeurs entières, chaîne de Markov cachée, modèles auto-régressifs.
AMS 2000 subject classifications: 62P25,62M10,62M05

1. Introduction

The study of phenomena evolving in time is a common topic in statistics, and the field of stochastic processes covers a wide range of models and techniques for describing and pointing out specific temporal behaviors. Generally, in most of these models, time is considered as a supplementary parameter or a supplementary variable only, and is not subject to further modeling or investigation.

[1] IHMC-PIREH, UMR 8589, Université Paris 1 Panthéon-Sorbonne, France
E-mail: julien.alerini@univ-paris1.fr
[2] SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne, France
E-mail: madalina.olteanu@univ-paris1.fr
[3] CEREMADE, UMR 7534, Université Paris-Dauphine, France
E-mail: james.ridgway@ensae.fr

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

On the contrary, time is the main research topic and the main issue for historians. Time is the first historian's paradigm.

History as a science is similar, in terms of epistemological object, to all humanities and social sciences. The difference comes from the consideration and the use of temporal structures in the scientific investigation (Koselleck, 2000). History aims not only at understanding the societies of the past, but also at studying their evolutions. The date allows to set the before, the while and the after the event, but dates alone do not establish a narrative. M. de Certeau claims that the narrative is set through the chronology. The dates, together with the chronology, are the necessary conditions for the possibility of splitting time into periods (de Certeau, 2002). Hence, time is a production of the historical science and not solely an astronomical object.

Furthermore, for historians, time does not have the same meaning as for physicists. For I. Newton, time is the binder between phenomena succeeding one another, while for A. Einstein, time is only a dimension of the universe. On the contrary, following E. Durkheim, and later G. Gurvitch, historians work with and think of *social time* (Gurvitch, 1961). The social time may be viewed as a time which serves for "tagging" the human society, each phenomenon having its own temporality (Prost, 1996). If the historians of the twentieth century reject a narrative based essentially on dates, they use chronologies and periodicities instead. Social time describes periods and establishes centuries. Thereby, historians speak of the "century of Louis XIV", a notion established in Voltaire's historical work, or of the "Iron century" when referring to the XVIIth century, which is opposed to the "beau XVIe siècle". For E. Hobsbawn, the "long XIXth century", which started in 1789 and ended in 1914, is the age of revolutions, capitalism and imperialism. In 1914 started the "short XXth" century, which represents the age of extremes, ended in 1991 (Hobsbawm, 2008). Thus, since the French Revolution, time has been segmented into centuries. This segmentation appears as necessary for the historical investigation, which requires a collective homogeneity in a limited duration (Milo, 1991). At the same time, this partition reveals to what extent the perception of time and particularly of the past may be subjective. This is an issue for memory phenomena. Both individual memory and institutional memory select the facts from the past, and reorder the information (Douglas, 2004). P. Ricoeur stresses this point. People grasp time through the narrative. The narrative of the past organises and relates time (Ricœur, 1991). C. Ginzburg points out the role of the clues and of the memory in defining time. The clues from the past (i.e. the historical sources) are more than instants or events. They narrate the before and the while of the event. They are a part of the after, which lasts the time of their appropriation by the societies (Ginzburg, 2010).

F. Braudel has another vision of time. Time would be the assembly of three distinct, yet mingled, times. The first one is the long term, *longue durée*, also called structural time or geographical time. It concerns the phenomena whose evolutions are so slow that they may not be noticed by men. The second one is the circumstantial time. It is the time of cyclic phenomena, of social changes on longer term. Finally, the third one is the short term, *courte durée*, the time of events. This is time lived by the individuals. For a long time, F. Braudel and his successors considered the time of events as trivial, not suited for historical investigation (Braudel, 1949, 1958). Nevertheless, this conception of different levels of time is mainly metaphoric. Although F. Braudel claims the necessity of articulating the studies of the different temporalities in order to understand the historical dynamics, he does not explain how one switches from one time to another. Despite this, F. Braudel influenced a whole generation of historians, who worked with these three temporalities.

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

During the sixties and the seventies, it was the circumstantial time that was mostly used. In the introduction of *Histoire économique et sociale de la France*, E. Labrousse states that the goal of the historian is to identify cycles and trends (Labrousse, 1970). He thus neglects the other two temporalities introduced by F. Braudel. E. Labrousse restricts his research to the increases and decreases observed in economic time series.

F. Braudel's conception of time was frequently criticized. On the one hand, M. Foucault insists upon the necessity of researching discontinuities and of identifying the events which are either revealing or causing the ruptures (Foucault, 2008). The microhistory asserts the importance of the event in highlighting a society's transformation (Ginzburg, 2010; Levi, 1989). On the other hand, rejecting the dichotomy event *vs.* structure, R. Koselleck introduces another approach of time. Time, or more precisely the temporal structure, is a tension between the experience of the past and the society's expectation of the future. The event and the structure are differentiated, but the event may become structurally significant and the structure may become an event (Koselleck, 2000). Hence, R. Koselleck refutes F. Braudel's architecture of time. Furthermore, time cannot be conceived as a chronology only. Thus, R. Koselleck proposes next the concept of regimes of historicity. A regime of historicity is a way of linking together past, present and future, and of giving them meaning. It may be also defined as the relation a society has with its past and how it reflects upon it (Hartog, 2003). The regimes of historicity are thereby used by F. Hartog as substitutes of the big historical segmentations mentioned previously or of the Braudelian vision of time.

By stating that the event defines a before and an after, one may then build periods of time which would represent phases of the human activity between two successive events (Bensa and Fassin, 2002). The difficulty lies in the identification of the event which establishes the alteration. Is the event producing the change or is it rather the expression of it? More particularly, it is possible that the change may not be assessed with respect to the event only. It may be the outcome of an accumulation of small items, a slow modification of the environment (Kuhn, 1993).

If the above theories on time allow to categorize the temporality, they do not provide the means for building appropriate periods for the phenomena studied by the historian. All these theories are useful to think history, but they are not much efficient for case studies and local analysis. Specific tools and practical methodologies devoted to time segmentation may then be expected to emerge from interdisciplinary work involving statistical learning.

This manuscript is the result of a collaboration involving historians and statisticians. Time and the temporal evolution of a specific historical phenomenon were the starting point and the main issue of this joint work. The aim of our collaboration was to study the Duchy of Savoy during the XVIth and the XVIIth centuries. These two centuries were marked by deep political changes and by several long and intense wars. It was a period during which the Duchy of Savoy changed and shaped its structure and its functioning as a state. More specifically, in the present manuscript we were interested in the rhythms of issuing legislative texts, related to military logistics. The corpus of data comes from the massive work of F-A. Duboin (Duboin, 1869; Couzin, 2008). This is an opus intended for the restoration of the Sabaudian law, following the Napoleonic age. It was meant to be of use to the new state. 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 Couzin (2008), this edition would be exhaustive, and few texts would be missing. Hence, we may consider this source as a complete edition of the

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

Sabaudian law, from the XIIIth to the XVIIIth centuries. The data which will be used hereafter appear as a time series, containing the number of legislative texts related to military logistics, issued by the Duchy between 1559 and 1661.

As will be illustrated in the next sections, empirical observations showed 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. Furthermore, during the studied period, the state generally increased and reinforced its grip on the whole society. But at what rhythm and which were the steps of this action? A segmentation of the data should bring some elements of answer: are there synchronous periods when comparing the production of law with the contexts of war? Or are there discordances between the two, which may inform on the state politics? If the legislative issuance accelerates before the beginning of a period of belligerence, one may infer that the state is preparing for war. Similarly, a prolonged period of an important production of law during peace can mean 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 remains low can mean either a neglect on behalf of the state, or that its military infrastructure is already advanced enough. Therefore, for the historian, the interest in segmenting the data lies in the necessity to detach himself from the periodization that is usually offered in the reading of what is traditionally called marking events.

All these questions guided the processing and the modeling of the data. The classical approach, largely used in quantitative history (Braudel and Labrousse, 1977; Guerreau, 2004) and which consists of smoothing the series, may have certainly brought out some insights on the functioning of the state, but it wouldn't have probably been the most adapted for answering the previous queries. Instead of smoothing, two new models based on hidden Markov chains were introduced. They were both designed in order to take into account the specificities of the data, which will be detailed hereafter.

The remaining of the manuscript is organized as follows: Section 2 contains a detailed description of the data and some historical and statistical considerations. Sections 3 and 4 introduce each of the proposed models, give the estimation procedures as well as some experimental results on simulated data. In Section 5, the two models are estimated on the historical data. The results are then explained and interpreted from the historian's perspective. The last section concludes with a discussion on the perspectives and the remaining work on these topics. The historical considerations in the next section and the commented results in Section 5 are drawn from the PhD manuscript of J. Alerini (2012).

2. The legislation on military logistics in the Duchy of Savoy

2.1. The data

As mentioned above, the corpus of data was extracted from the work of F.-A. Duboin (1869). The study of his work allows 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

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

selected for further investigation. The role and the purpose of a document were used as selection criterion. Thus, the legislation related to roads and bridges was included, although it had not been specifically issued for military purposes. On the other hand, the texts related to taxes which do not apply directly to the supply, food or accommodation of the troops were left out, although some of these taxes were used to pay the soldiers. Also, the edicts on military discipline were included, since they link the issue of indiscipline of the soldier to problems of food and lodging. The final dataset consists of 472 documents, representing 8.17% of the whole legislation. This ratio may seem relatively small and this is the first surprise of our study, since we can state now that the question of the operational maintenance of the army represents about 8% of the administrative and legislative effort of the state. However, since no other comparable studies in other countries or other periods exist, we cannot evaluate the importance of it. A full description of the corpus and of its construction is available in Alerini (2012).

The dataset also includes a second series, which will be called "global", containing the entire production of law between 1559 and 1661, and a third series which is computed as the ratio between the number of texts on military logistics and the total number of law texts. These two series are also important for studying the transformation of the state, but they will not be analyzed in detail in this manuscript.

2.2. *What time scale for the series extracted from the corpus?*

In order to evaluate the temporality of the state as well as the contribution of military logistics to its activity, the selected corpus of documents was represented as a time series. Three different time scales were considered, as illustrated in Figures 1, 2 and 3.

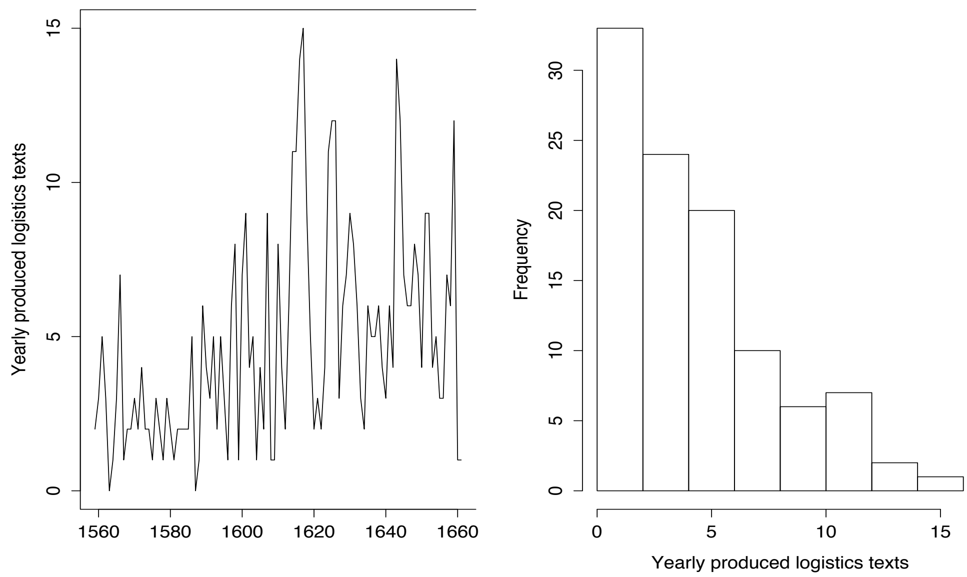

FIGURE 1. *Yearly representation (time series and histogram) for the legislation issued on military logistics*

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

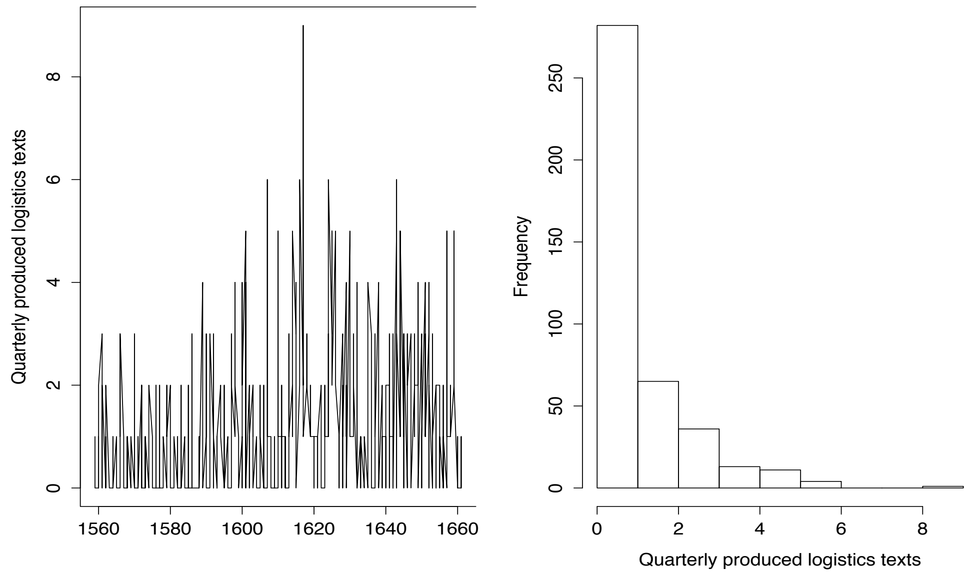

FIGURE 2. *Quarterly representation (time series and histogram) for the legislation issued on military logistics*

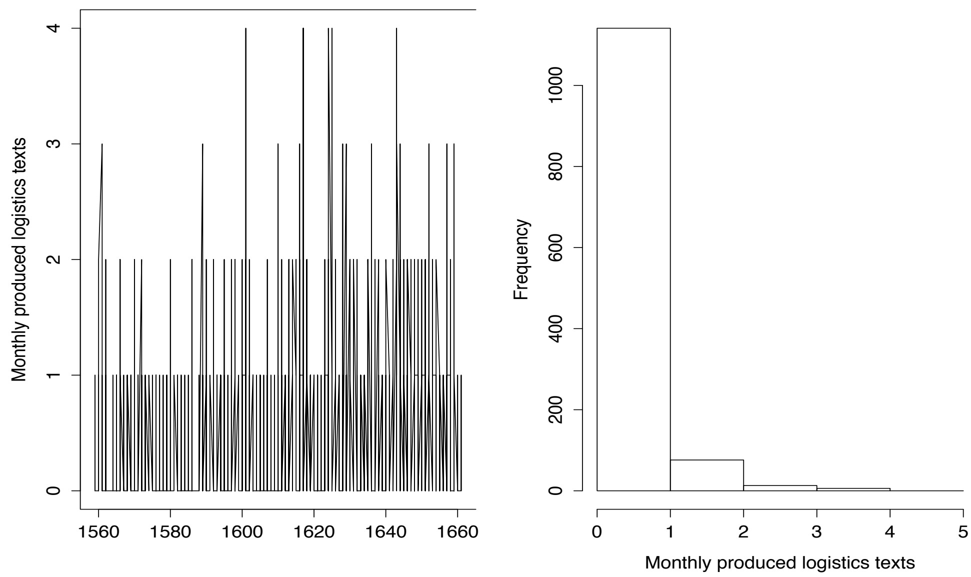

FIGURE 3. *Monthly representation (time series and histogram) for the legislation issued on military logistics*

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

The yearly representation, Figure 1, has the advantage of providing a more robust time series, less sensitive to possible missing data, but this choice is obviously too rough for emphasizing interesting events during the studied period. Moreover, with a yearly representation, the length of the time series (122 observations) is not sufficient for obtaining reliable estimates of the models introduced in the next sections.

A quarterly analysis, Figure 2, provides other benefits. At that time, tax collecting was quarterly organized and this gave a periodicity to the administration activity. From this point of view, the quarterly representation may appear as the most suited, since it shares the same time perception with the contemporaries. However, this partitioning has also some drawbacks. The state needed to collect more at some periods and this led to an increase of "extraordinary" quarters, the number of quarters varying from four to twelve according to the year. The quarterly segmentation is also troubled by the effective length of the quarters. For example, in the 1650s, the winter quarters ran from December to mid-May, thereby during five months and a half. In this case, the quarters do not longer define a fraction of the collected tax, but the duration of the tax collection.

The monthly approach, Figure 3, is finer than the previous two. Within this framework, one may observe the closeness between making the decision and issuing the associated document, although the cut-off between two documents, one issued the last day of a month and the other issued the first day of the following month may sometimes be arbitrary (from this point of view, the quarterly analysis is smoother and takes into account this issue better). Furthermore, using a monthly partitioning leads to the observation of many months without any production of documents (69.9% of the observed period).

Between 1559 and 1661, the state as a whole issued 55.5 law texts per year (in average), of which 4.8 in connection with military logistics (the global series varies between 19 and 107 per year with a median of 52 , while the series of texts on military logistics varies between 0 and 15 with a median of 4). The ratio between the texts on military logistics and the total number of law texts varies between 0 and 0.23, and, for more than half of the series, it takes values between 0.12 and 0.46. The importance of military logistics is thus far from negligible.

The data represented as a time series offer a first hint of periodization. A significant peak at the beginning of the period (1560 and 1561) can be explained by the restoration conducted by Emmanuel Philibert, and the subsequent tax and administrative reforms. From then on, the legislative documents are issued on a relatively constant rhythm until the 1600s. A sudden increase in the production of law occurs in 1610, when the Duchy of Savoy enters the XVIIth century war-cycle. Texts on military logistics are then steadily produced during the following period, and their issuance appears to be less sensitive to periods of conflict.

In the light of the previous comments and remarks and despite the discussed drawbacks, the monthly representation of the data was considered preferable among the others and used throughout the rest of the paper.

2.3. War and peace

Let us now take a closer look at the monthly diplomatic situation. Some basic summary statistics on the production of law are available in Table 1. According to these, there is a significant difference in mean when comparing war *vs.* peace periods, and more particularly in the corpus of military-logistics documents. The analysis of variance (see the results in Appendix A) clearly

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

indicates that being at peace or being at war has a strong impact on legislative issuance. These results confirm 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.

TABLE 1. *Monthly production of law*

	War (45.7% of the data)			Peace (54.3% of the data)			Whole data		
	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
"Global" series	0	4.87	17	0	4.05	18	0	4.43	18
"Logistics" series	0	0.54	4	0	0.28	4	0	0.4	4

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 currently used to our knowledge for studying time series in quantitative history and based on ARIMA-type models (Bonneuil, 1991; Mathis and Grenier, 1991), 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.

2.4. *Some statistical issues concerning the corpus*

Two main specificities are to be taken into account by the models which will be used hereafter: on the one hand, an important over-dispersion due the excess of zeros, and, on the other hand, an approximate six-month seasonality, particularly linked to military traditions. The latter was suggested by the historian, but it may also be observed on the partial autocorrelation function in Figure 4. Furthermore, according to the historical expertise, it is natural to suppose the existence of several regimes controlling the data (peace *vs.* war, political changes, ...).

Before introducing new models for these specific data, some of the existing methods on segmenting time-series were tested. First, several hidden-Markov models in the **R**-package **HiddenMarkov** (Harte, 2015) were estimated. More specifically, the binomial and the Poisson distributions were considered as emission distributions. Unfortunately, the results (not reported here) were not very encouraging in terms of stability of the transition matrix and in terms of interpretability and, in our opinion, this was mainly due to the excess zeros in the dataset. Second, the strategy of change-point detection was considered. Change-points in mean or/and variance were computed, with either an underlying Poisson distribution or a Gaussian one, and using the **R**-package **changepoint** (Killick and Eckley, 2014; Killick et al., 2014). Here also, the excess of zeros does not provide interpretable results from the historian's point of view (an example

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

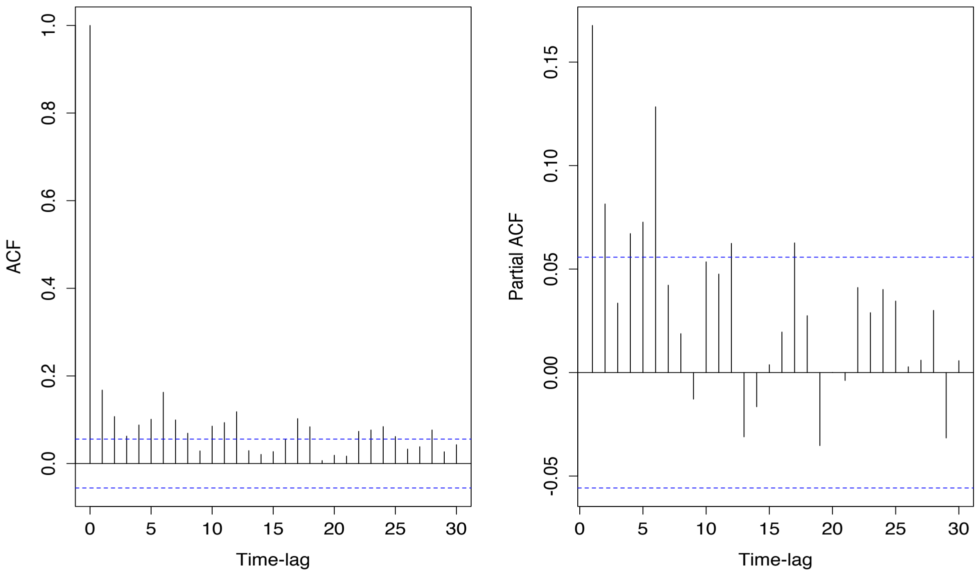

FIGURE 4. *Autocorrelation and partial-autocorrelation functions for the monthly series*

of change-point detection is illustrated in Appendix B). Moreover, the algorithms looking for change points in time series are probably not the best suited for this kind of data, since they do not allow for transitions between two periods. Eventually, none of the above methods was taking into account the seasonality of the data.

The key ideas and the main mathematical contributions of the present paper are outlined next and will be detailed in the following sections. Since the main goal here is to segment the data according to some latent "regimes" (which may be linked to war *vs.* peace periods but not necessarily identical) and also to highlight the existence of transitions between these regimes, the use of hidden Markov models was a natural choice. Let us recall that time series of counts are often modeled with hidden Markov models (HMM hereafter, Baum and Petrie, 1966). These are especially interesting in the context of the presumed existence of several regimes controlling the parameters of the model (coding *vs.* non-coding regions for DNA data, crisis *vs.* stable periods for financial data, ...).

The first model proposed in this manuscript was designed for taking into account the over-dispersion of the data. This was achieved by considering zero-inflated Poisson (ZIP hereafter) as emission distributions in the basic HMM model. Introduced in the late 60's (Cohen, 1963), ZIP(η, λ) distributions allow for excess zeros as follows:

$$\mathbb{P}(X = k) = \begin{cases} \eta + (1-\eta)e^{-\lambda} & , \quad k = 0 \\ (1-\eta)\frac{e^{-\lambda}\lambda^k}{k!} & , \quad k \in \mathbb{N}^\star \end{cases} ,$$

where $\eta \in]0,1[$ and $\lambda > 0$. ZIP distributions were mainly used for regression purposes (Mullahy, 1986; Lambert, 1992), but one can extend their use in the context of HMM. Let us mention

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

that the idea of mixing ZIP distributions and hidden Markov models is not completely new. For example, DeSantis and Bandyopadhyay (2011) and Wang (2001) proposed close versions of the same model by introducing a partially observed Markov chain such that the two components in the zero-inflated Poisson occur according to the states of the partially observed process.

ZIP distributions may be preferred to Poisson distributions in some cases, since they allow for a certain amount of dispersion in the data (Ridout et al., 1998), in the sense that in the Poisson distribution the mean and the variance are both equal to λ, whereas for $X \sim ZIP(\eta,\lambda)$,

$$\mathbb{E}(X) = (1-\eta)\lambda = \mu,$$
$$\mathbb{V}(X) = \mu + \left(\frac{\eta}{1-\eta}\right)\mu^2.$$

Hence, this distribution allows for more flexibility in the modeling process. The complete description of the proposed ZIP-HMM model as well as the estimation procedure will be detailed in Section 3.

The second contribution of the manuscript was aimed at taking into account the seasonality and the autoregressive structure of the data, as pointed out by the summary statistics and as expected by the historian. This was achieved by combining HMM with an integer-valued autoregressive model. For the autoregressive part, we used the integer-valued autoregressive models of order p, INAR(p) hereafter, as introduced in Al-Osh and Alzaid (1987), Al-Osh and Alzaid (1990) and Jin-Guan and Yuan (1991).

The main reason justifying the choice of INAR(p) processes in the context of regime-switching time series is that the behavior of the autocorrelation function is similar to that of an AR(p) model (Jin-Guan and Yuan, 1991). Let us recall that autoregressive models for real-valued time-series have been extensively used and studied in the context of regime-switching Markov models since the seminal article Hamilton (1989). Hence, INAR(p) models may be used similarly to linear autoregressive models AR(p) and combined with a hidden Markov process. The complete description of the INAR(p)-HMM model as well as the estimation procedure are given in Section 4.

3. Hidden Markov models with zero-inflated Poisson distributions (ZIP-HMM)

In this section, a hidden Markov model is proposed for dealing with time series of counts with excess zeros. This framework has already been introduced in a short version in Olteanu and Ridgway (2012) and will be further developed in the subsequent paragraphs.

3.1. The model

Let $(X_t)_{t\in\mathbb{Z}}$ be the observed time series, valued in $\mathbb{N}$, and let $(S_t)_{t\in\mathbb{N}}$ be the unobserved process, controlling the parameters of the distribution of X_t. Throughout the rest of the paper, S_t is supposed to be a homogeneous Markov chain, valued in a finite state-space $E = \{e_1,...,e_q\}$ and defined by its transition matrix $(\pi_{ij})_{i,j=1,...,q}$,

$$\pi_{ij} = \mathbb{P}(S_t = e_j | S_{t-1} = e_i)$$

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

and by its initial probability distribution π^0, $\pi_i^0 = \mathbb{P}(S_1 = e_i)$, $\forall i = 1,...,q$.
Moreover, let us suppose that X_t are independent conditionally to S_t, and that X_t conditionally to S_t is distributed according to a ZIP(η_i, λ_i):

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

Hence, the parameter space may be written as:

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

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

3.2. *The estimation procedure*

Since the above model involves a hidden Markov chain, the estimation procedure is carried out using the EM algorithm (Baum and Petrie, 1966; Dempster et al., 1977). However, the regular Baum-Welsh method cannot be applied directly, since the ZIP is a mixture between a Poisson and a Dirac distributions. Hence, the EM algorithm needs to be slightly adapted to our problem.

In order to write the complete likelihood, an auxiliary sequence of variables is introduced. Let us define Z_t an underlying random process such that $Z_t = 1$ leads to a "structural" zero, with $Z_t|S_t = e_i \sim \mathscr{B}er(\eta_i)$, a Bernoulli distribution. By structural zero we mean a zero induced by the Dirac rather than the Poisson. Hereafter, we suppose that the pairs (X_t, Z_t) are independent, conditionally to the hidden process S_t. With the previous assumption, and with the notations $X_1^T = (X_1,...,X_T)$, $S_1^T = (S_1,...,S_T)$, $Z_1^T = (Z_1,...,Z_T)$, the complete likelihood is given by:

$$\mathscr{L}(Z_1^T, X_1^T, S_1^T; \theta) = \prod_{t=1}^{T}\prod_{i=1}^{q} f(X_t, Z_t|S_t = e_i;\theta)^{\mathbf{1}_{e_i}(S_t)} \prod_{t=2}^{T}\prod_{i,j=1}^{q} \pi_{ij}^{\mathbf{1}_{e_i,e_j}(S_{t-1},S_t)} \times C\,, \tag{3}$$

$$\text{where } f(X_t, Z_t|S_t = e_i;\theta) = \eta_i^{\mathbf{1}_{Z_t=1}}(1-\eta_i)^{\mathbf{1}_{Z_t=0}}\left(\frac{e^{-\lambda_i}\lambda_i^{X_t}}{X_t!}\right)^{\mathbf{1}_{Z_t=0}}, \tag{4}$$

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.
The algorithm consists in maximizing $\mathbb{E}_{\theta^\star}\left[\ln(\mathscr{L}(Z_1^T, X_1^T, S_1^T;\theta))|X_1^T\right]$ with respect to θ and updating $\theta^\star$ at each step.
E-Step. The expectation step is given by:

$$\begin{aligned} Q(\theta|\theta^\star) &= \mathbb{E}_{\theta^\star}\left[\ln(\mathscr{L}(Z_1^T, X_1^T, S_1^T;\theta))|X_1^T\right] \\ &= \textstyle\sum_{t=1}^{T}\sum_{i=1}^{q}\{\mathbb{P}_{\theta^\star}(S_t = e_i, Z_t = 1|X_1^T)\ln(\eta_i) \\ &+ \mathbb{P}_{\theta^\star}(S_t = e_i, Z_t = 0|X_1^T)(\ln(1-\eta_i) - \lambda_i + X_t\ln(\lambda_i) - \ln(X_t!))\} \\ &+ \textstyle\sum_{t=2}^{T}\sum_{i,j=1}^{q}\mathbb{P}_{\theta^\star}(S_{t-1} = e_i, S_t = e_j|X_1^T)\ln(\pi_{ij}). \end{aligned} \tag{5}$$

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

We take further interest in the part of Equation 5 containing the parameters associated to the ZIP distribution; we denote it ν_θ (the rest of the equation will be dealt with separately). This equation contains two joint probabilities that can be expressed as follows:

$$\mathbb{P}_{\theta^\star}(S_t = e_i, Z_t = 1 | X_1^T) = \mathbb{P}_{\theta^\star}(Z_t = 1 | X_1^T, S_t = e_i)\mathbb{P}_{\theta^\star}(S_t = e_i | X_1^T). \quad (6)$$

The latter probability of the right hand-side of (6) is obtained by the Baum-Welch forward-backward algorithm, while the first is given by:

$$\mathbb{P}_{\theta^\star}(Z_t = 1 | X_1^T, S_t = e_i) = \begin{cases} 0, \text{ if } X_t > 0 \\ \frac{\mathbb{P}_{\theta^\star}(X_t=0|S_t=e_i, Z_t=1)\mathbb{P}_{\theta^\star}(Z_t=1|S_t=e_i)}{\mathbb{P}_{\theta^\star}(X_t=0|S_t=e_i)}, \text{ if } X_t = 0. \end{cases}$$

By noting $\xi_i^\star = \frac{\eta_i^\star}{\eta_i^\star + (1-\eta_i^\star)e^{-\lambda_i^\star}}$, we get:

$$\mathbb{P}_{\theta^\star}(S_t = e_i, Z_t = 1 | X_1^T) = \begin{cases} 0, \text{ if } X_t > 0 \\ \xi_i^\star \mathbb{P}_{\theta^\star}(S_t = e_i | X_1^T), \text{ if } X_t = 0. \end{cases}$$

Similarly,

$$\mathbb{P}_{\theta^\star}(S_t = e_i, Z_t = 0 | X_1^T) = \begin{cases} \mathbb{P}_{\theta^\star}(S_t = e_i | X_1^T), \text{ if } X_t > 0 \\ (1-\xi_i^\star)\mathbb{P}_{\theta^\star}(S_t = e_i | X_1^T), \text{ if } X_t = 0. \end{cases}$$

Therefore, ν_θ can be expressed as:

$$\nu_\theta = \sum_{t:X_t>0} \sum_{i=1}^{q} p_t(e_i)\{\ln(1-\eta_i) - \lambda_i + X_t \ln(\lambda_i) - \ln(X_t!)\}$$
$$+ \sum_{t:X_t=0} \sum_{i=1}^{q} \xi_i^\star p_t(e_i)\ln(\eta_i) + (1-\xi_i^\star)p_t(e_i)\{\ln(1-\eta_i) - \lambda_i\}, \quad (7)$$

where $p_t(e_i) = \mathbb{P}_{\theta^\star}(S_t = e_i | X_1^T)$.

M-Step. The maximization step can be carried analytically. By denoting $\mathbb{P}_{\theta^\star}(S_{t-1} = e_i, S_t = e_j | X_1^T) = p_t(e_i, e_j)$, and by recalling that $p_t(e_i)$ and $p_t(e_i, e_j)$ may be computed using the forward-backward algorithm, the following updates are obtained:

$$\hat{\pi}_{ij} = \frac{\sum_{t=2}^T p_t(e_i, e_j)}{\sum_{t=1}^T p_t(e_i)};$$

$$\hat{\eta}_i = \frac{\xi_i^\star \sum_{t:X_t=0} p_t(e_i)}{\sum_{t=1}^T p_t(e_i)};$$

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

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

3.3. Simulation study

The quality of the estimates and the convergence of the algorithm are 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, 3 and 4. The results are globally very stable, showing relatively low MSE, decreasing with the sample size. The computational times are also quite reasonable. For example, one initialization of the EM algorithm takes (in average) 0.45 seconds for 500 observations, 0.60 for 1 000, 2.70 for 5 000 and 7.95 for 10 000.

TABLE 2. *MSE* ($\pi_{22} = 0.6, \eta_1 = 0.2, \lambda_1 = 0.5, \eta_2 = 0.2, \lambda_2 = 3$)

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 3. *MSE* ($\pi_{11} = 0.4, \pi_{22} = 0.6, \eta_1 = 0.2, \eta_2 = 0.2, \lambda_2 = 3$)

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 4. *MSE* ($\pi_{11} = 0.4, \pi_{22} = 0.6, \lambda_1 = 0.5, \eta_2 = 0.2, \lambda_2 = 3$)

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

4. Integer-valued autoregressive hidden Markov models, INAR(p) - HMM

In this section, we introduce a new model, which combines hidden Markov models with integer-valued autoregressive processes.

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

4.1. The model

Let us first recall the definition of an INAR(p) model, as introduced in Al-Osh and Alzaid (1987), Al-Osh and Alzaid (1990) and Jin-Guan and Yuan (1991). The sequence of integer-valued variables $(X_t)_{t\in\mathbb{Z}}$ is said to be an INAR(p) if the following relation is verified:

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \cdots + \alpha_p \circ X_{t-p} + \varepsilon_t \ , \tag{8}$$

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. For all $i = 1,\dots,p$,

$$\alpha_i \circ X_{t-i} = \sum_{k=1}^{X_{t-i}} \xi_{i,k} \ , \tag{9}$$

is the Steutel-van Harn thinning operator introduced in Steutel and van Harn (1979) 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 $X_{t-1} = x_{t-1},\dots,X_{t-p} = x_{t-p}$ may be written as follows:

$$\begin{array}{c} f(x_t|x_{t-p}^{t-1}) = \sum\limits_{i_1=0}^{x_t\wedge x_{t-1}} C_{x_{t-1}}^{i_1}\alpha_1^{i_1}(1-\alpha_1)^{x_{t-1}-i_1} \sum\limits_{i_2=0}^{x_t-i_1\wedge x_{t-2}} C_{x_{t-2}}^{i_2}\alpha_2^{i_2}(1-\alpha_2)^{x_{t-2}-i_2} \\ \cdots \sum\limits_{i_p=0}^{(x_t-i_1-\cdots-i_{p-1})\wedge x_{t-p}} C_{x_{t-p}}^{i_p}\alpha_p^{i_p}(1-\alpha_p)^{x_{t-p}-i_p}\frac{e^{-\lambda}\lambda^{x_t-i_1-\cdots-i_p}}{(x_t-i_1-\cdots-i_p)!} \end{array} , \tag{10}$$

where $x_{t-p}^{t-1} = (x_{t-p},\dots,x_{t-1})$ and supposing that ε_t is distributed according to a Poisson of parameter $\lambda > 0$.

The hybrid model INAR(p)-HMM may be now introduced. Let $(X_t)_{t\in\mathbb{Z}}$ be the observed time series, valued in $\mathbb{N}$, and let $(S_t)_{t\in\mathbb{N}}$ be a hidden Markov chain, as defined in Section 3.1. Furthermore, the observed time-series X_t is supposed to be INAR(p), conditionally to S_t:

$$(X_t|S_t = e_i) = \alpha_{1,i} \circ X_{t-1} + \alpha_{2,i} \circ X_{t-2} + \cdots \alpha_{p,i} \circ X_{t-p} + \varepsilon_{i,t}, \tag{11}$$

where $\varepsilon_{i,t} \sim \mathscr{P}(\lambda_i)$, 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:

$$\Theta = \{\theta = (\alpha,\lambda,\pi) \in]0,1[^{p\times q} \times (\mathbb{R}^+)^q \times]0,1[^{q^2} \text{ and } \forall i \in \{1\cdots q\}, \sum_{j=1}^{q} \pi_{ij} = 1\}, \tag{12}$$

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

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

4.2. Estimation procedure

Here again, as for the previous model in the previous section, parameter estimation is carried out using the EM algorithm. For a given time-series $X_{-p+1}^T = (X_{-p+1},...,X_T)$ and a given Markov-chain path $S_1^T = (S_1,...,S_T)$, the complete likelihood of the data may be written as:

$$\mathscr{L}(X_{-p+1}^T, S_1^T; \theta) = \prod_{t=1}^{T}\prod_{i=1}^{q} f(X_t|X_{t-p}^{t-1}, S_t = e_i; \theta)^{\mathbf{1}_{e_i}(S_t)} \times \prod_{t=2}^{T}\prod_{i,j=1}^{q} \pi_{ij}^{\mathbf{1}_{e_i,e_j}(S_{t-1},S_t)} \times C, \qquad (13)$$

where $C = \prod_{i=1}^q \left(\pi_i^0\right)^{\mathbf{1}_{e_i}(S_1)} \times f\left(X_{-p+1}^0\right)$ is the likelihood of the initial state of the Markov chain and of the first observations.

E-step First, the expectation of the complete log-likelihood of the data conditionally to the observed time series is computed.

$$\mathbb{E}_{\theta^\star}\left(\ln \mathscr{L}(X_{-p+1}^T, S_1^T; \theta)|X_{-p+1}^T\right) = \mathbb{E}_{\theta^\star}\left(\sum_{t=1}^T \sum_{i=1}^q \mathbf{1}_{e_i}(S_t) \ln f(X_t|X_{t-p}^{t-1}, S_t = e_i; \theta)|X_{-p+1}^T\right)$$

$$+\mathbb{E}_{\theta^\star}\left(\sum_{t=2}^T \sum_{i,j=1}^q \mathbf{1}_{e_i,e_j}(S_{t-1},S_t) \ln \pi_{ij}|X_{-p+1}^T\right) + C. \qquad (14)$$

By denoting as previously

$$p_t(e_i) = \mathbb{P}_{\theta^\star}\left(S_t = e_i|X_{-p+1}^T\right),$$

and

$$p_t(e_i,e_j) = \mathbb{P}_{\theta^\star}\left(S_{t-1} = e_i, S_t = e_j|X_{-p+1}^T\right),$$

the expectation of the complete log-likelihood may be written as:

$$\begin{aligned} \mathbb{E}_{\theta^\star}\left(\ln \mathscr{L}(X_{-p+1}^T, S_1^T; \theta)|X_{-p+1}^T\right) &= \sum_{t=1}^T \sum_{i=1}^q p_t(e_i) \ln f(X_t|X_{t-p}^{t-1}, S_t = e_i; \theta) \\ &+ \sum_{t=2}^T \sum_{i,j=1}^q p_t(e_i,e_j) \ln \pi_{ij} + C \qquad (15) \\ &= \nu_\theta + \mu_\theta + C. \end{aligned}$$

The difficulty here, when compared with the ZIP-HMM model in the previous section, resides in the practical computation of the log-likelihood for a general time-lag p. In the algorithmic implementation, this difficulty was solved by using recursive programming.

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

$$\hat{\pi}_{ij} = \frac{\sum_{t=2}^T p_t(e_i,e_j)}{\sum_{t=1}^T p_t(e_i)},$$

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

where $p_t(e_i, e_j)$ and $p_t(e_i)$ are being computed with the Baum-Welch forward-backward algorithm.

For ν_θ, 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 constrains on the α's and λ's are first removed by reparameterizing 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 of ν_θ 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 in the next section.

4.3. 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 data are 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 5.

TABLE 5. *Mean squared error - scenario A*

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

Scenario B The data are 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 6.

TABLE 6. *Mean squared error - scenario B*

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

Scenario C The data are simulated according to a HMM-INAR(1) with two states for the hidden Markov chain. The parameters kept constant are $\pi_{11} = 0.2, \pi_{22} = 0.4, \alpha_1 = 0.2, , \alpha_2 = 0.1, \lambda_2 = 4$.

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

The remaining parameter, λ_1 takes values in the interval $[0.1, 7]$. The results are given in Tables 7 and 8.

TABLE 7. *Mean squared error - scenario C*

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 8. *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

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.

5. Segmenting a century of Sabaudian history with hidden Markov models

Let us now present the results on the corpus of data described in Section 2. As a reminder, the dataset consists of a time series containing the monthly production of law, related to military logistics and issued between 1559 and 1661 (1 236 observations). Furthermore, the architectures of the models to be estimated (number of states for the hidden Markov chain, number of lags in the autoregressive parts) are 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), as explained in Section 2.3.

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). The estimated parameters and the estimated a posteriori probabilities for one of the states in the three-state ZIP-HMM are available in Appendix C.

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

5.1. Results with the ZIP-HMM model

First, a ZIP-HMM model with two hidden states was estimated from the data. The estimations of the various parameters are given below:

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

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. Next, the a-posteriori conditional probabilities of the Markov chain being in the second regime are computed and plotted in the second graph of Figure 5. Moreover, the values of these probabilities are thresholded at 0.5 (blue dotted line).

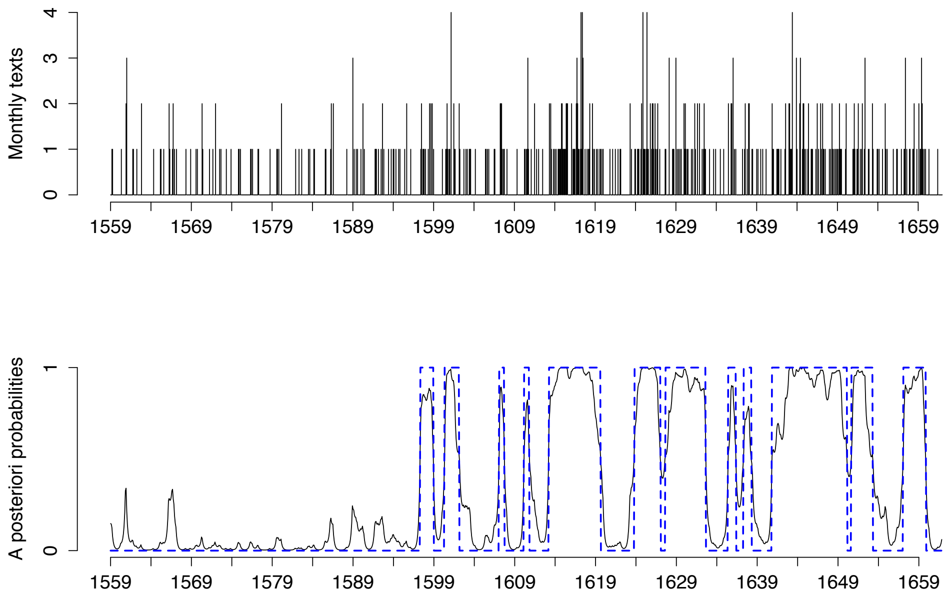

FIGURE 5. *Initial time series and a-posteriori probabilities for the second regime of the ZIP-HMM model*

A close study of Figure 5 allows a detailed analysis of the activity of the state. The relationship between war and legislative issuance is neither absolute nor synchronous. Indeed, the state can either anticipate, or draw the consequences of a war, and thus have an intense legislative activity during peacetime. Figure 7 is illustrating a timeline matching the a posteriori probabilities of the hidden Markov regimes and the periods of war and peace.
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 usual on matters of feeding or lodging the military. For a switch to regime B, one has to wait until June 1597, when an important series of documents redefining the military logistics of the Duchy starts being issued.

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

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 regime B 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 regime B 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 follows regime B. 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, regime B 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. Regime B, 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 does not dismiss his troops. On the contrary, he maintains an army two and half times larger than at the beginning of the century. The endeavor to lodge and supply these soldiers continues at the same pace as during the war, also the number and the frequency of the directives and orders does not change. 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 regime B 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 regime B (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, regime B 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. For instance, the switch from regime B to regime A in June 1653 is due to an alteration of the taxes which reduces the number of orders necessary to

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

collect them. On the contrary, in 1657, a new switch to regime B may be due to the extraordinary charges and to the attempts of the state to collect tax arrears from the communities.
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 despite this. 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.

5.2. Results with the INAR(6)-HMM model

The data were also 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 in Figure 4. 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, both close to being absorbant. The expected values of the Poisson distributions are relatively close and quite small. The main difference between the two regimes arises from 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 biannual 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 6, 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.

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

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

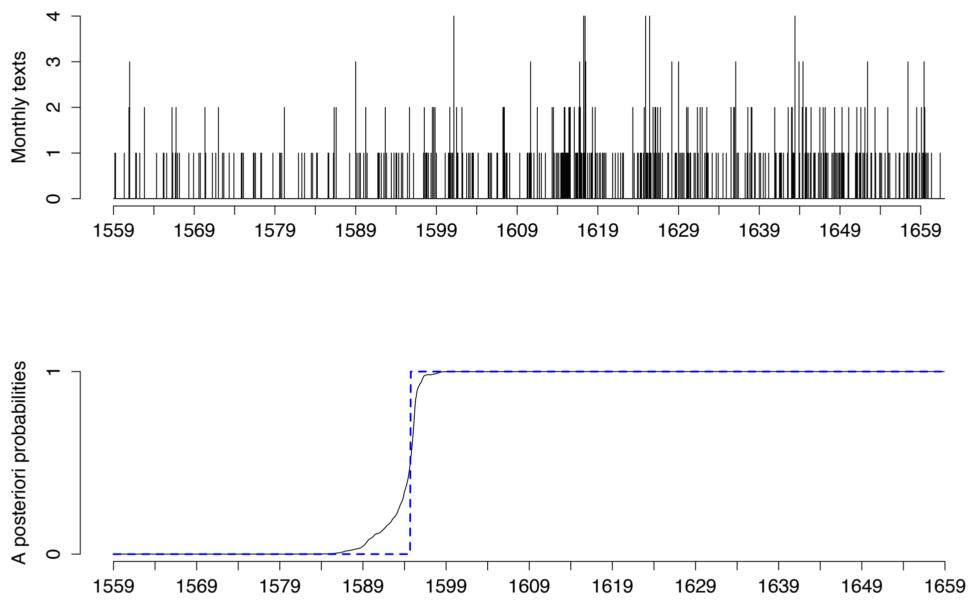

FIGURE 6. *Initial time series and a-posteriori probabilities for the first regime of the INAR(6)-HMM model*

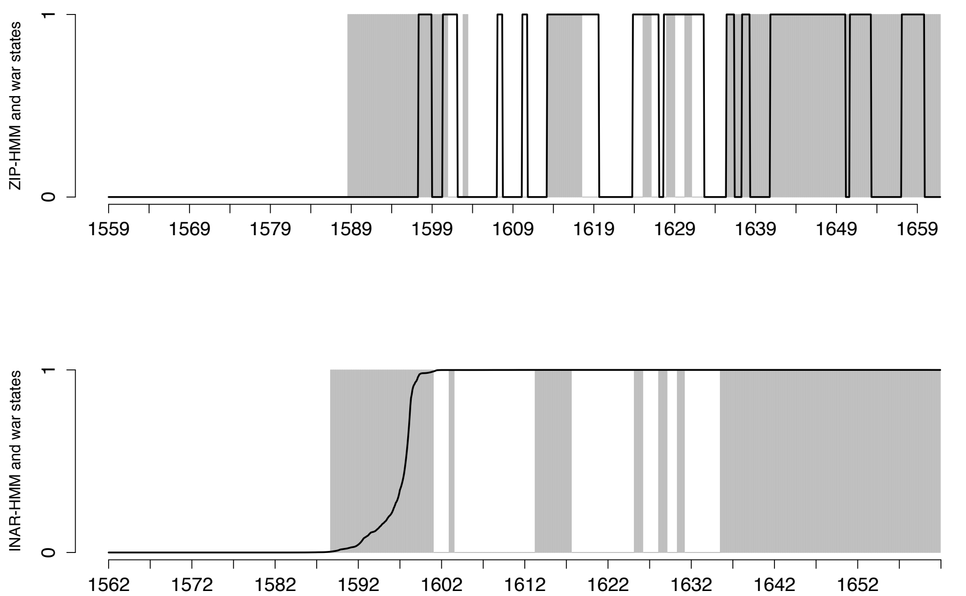

FIGURE 7. *Hidden-Markov models segmentations crossed with war (grey) vs. peace (white) periods*

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.

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

During this period of transition, one may witness an increase in the legislative activity of the Duchy. These results converge with the analysis of C. Rosso (1992) 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.

Hence, the two models prove to be complementary: the ZIP-HMM model appeares 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.

The above estimated models may be used for simulating artificial time series and then compare them with the real data. An example is provided in Appendix D. As illustrated in Figures 12 and 13, the artificial data sets generated with the estimated parameters appear to be quite similar to the real data, and comfort the idea that the use of expert knowledge for selecting the architecture of the models (number of states, lags of time) was suitable in this framework.

6. Conclusions and future work

Two models were introduced in the present manuscript, each of them capturing some specificities of a historical time series: on the one hand, the ZIP-HMM model was able to handle the excess of zeros and appeared as particularly suited for modeling short-term events; on the other hand, the INAR(p)-HMM model was able to take into account the seasonality and the autoregressive structure of the series and allowed to bring out a long period of transition between two normative systems of the Duchy of Savoy.

From the historian's point of view, the results are very encouraging and motivate a deeper study: we did not consider here the series of all legislative documents or the ratio between the documents on military logistics and the whole legislation. The analysis performed here is not a cyclic one, as E. Labrousse whished (Labrousse, 1970). In our study, time is not considered as a series of events, but as a series of states. The identification of different periods does not depend on the increase or the decrease of some variable. It is the rhythm at which the state changes that defines the periods. The political event (i.e. make peace or declare war) is only a means for segmenting time, but it is not the unique one. Issuing a text of law is also an event, but the difficulty lies in singling out the text which makes sense and which represents the event segmenting time. Selecting an event which accounts for a rupture in time is a subjective choice. Each text of law issued by the Duchy is of the same importance in the model, but the series of events (the legislation issuance) are different. The methodology used here aims at creating coherent ensembles, which do not depend on the a-priori interpretation of some events.

Three levels of reading may then be proposed and articulated together. First, there is the monthly time series indexing the production of law. It corresponds to an observation of the events, which provides local significance but no global interpretation. Next, a detailed analysis of the state's statuses ensues, reflecting a level of temporary activity. This second level allows the comparison between the diplomatic situation and the military preparedness. The Markov model provides a link between the event and the circumstance. Eventually, the third level of reading allows to understand the path from the events to a change in the structure of the state. Although they may appear close to the Braudelian architecture of time, these levels of interpretation do not go back to

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

the metaphor proposed by F. Braudel. Instead of this, they allow to understand the path from the event to the trend.

From the statistician's point of view, the two models proposed above need to be further studied: consistency, asymptotic properties of the estimates, ... The conditions for the identifiability of the ZIP-HMM have already been established in a general framework by Gassiat et al. (2015). The estimation procedure for the INAR(p)-HMM model, which is very slow and subject to local minima with the EM algorithm, needs to be improved also, and a MCMC approach may be considered (Rydén, 2008). Moreover, we are currently investigating the extension of the ZIP-HMM model to other probability distributions: negative Binomial (in view of an application to the same data set) and Beta (in view of an application to the ratio between the series on military logistics and the series on the global production of law).

At last, let us stress once more that the present manuscript is the result of a true interdisciplinary work, and not only an "application" of a statistical model to a series of data. The corpus of data and its specificities guided the construction of the two models throughout, while the results we obtained open new and promising areas of research for both the statistician and the historian.

References

Al-Osh, M. and Alzaid, A. (1987). First-order integer-valued autoregressive (inar(1)) process. *Journal of Time Series Analysis*, 8:261–275.

Al-Osh, M. and Alzaid, A. (1990). An integer-valued p-th order autoregressive (inar(p)) process. *Journal of Applied Probability*, 27(2):314–324.

Alerini, J. (2012). *La Savoie et le "Chemin espagnol", les communautés alpines à l'épreuve de la logistique militaire (1560-1659)*. PhD thesis, Université Paris 1 Panthéon-Sorbonne.

Baum, L. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563.

Bensa, A. and Fassin, E. (2002). Les sciences sociales face à l'évènement. *Terrain*, 38:5–20.

Bonneuil, N. (1991). Temporalités en démographie historique. *Histoire & Mesure*, 6(1-2):137–148.

Braudel, F. (1949). *La Méditerranée et le Monde méditerranéen à l'époque de Philippe II*. Armand Colin, Paris.

Braudel, F. (1958). La longue durée. *Annales ESC*, 13(4):725–753.

Braudel, F. and Labrousse, E. (1977). *Histoire économique et sociale de la France.* Presses universitaires de France, Paris.

Cohen, A. (1963). Estimation in mixtures of discrete distributions. In *Proc. Int. Symp. on Classical and Contagious Discrete Distributions*, pages 373–378. Montreal, Pergamon Press.

Couzin, T. (2008). 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.

de Certeau, M. (2002). *L'écriture de l'histoire*. Gallimard, Paris.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society (B)*, 39(1):1–38.

DeSantis, S. and Bandyopadhyay, D. (2011). Hidden markov-models for zero-inflated poisson counts with an application to substance use. *Statistics in Medicine*, 30(14):1678–1694.

Douglas, M. (2004). *Comment pensent les institutions*. La découverte, Paris.

Duboin, F.-A. (1818-1869). *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.

Foucault, M. (2008). *L'archéologie du savoir*. Gallimard, Paris.

Gassiat, E., Cleynen, A., and Robin, S. (2015). Inference in finite state space non parametric hidden markov models and applications. *Statistics and Computing*, pages 1–11.

Ginzburg, C. (2010). *Mythes, emblèmes, traces : morphologie et histoire*. Verdier, Lagrasse.

Guerreau, A. (2004). *Statistique pour historiens*. Ecole Nationale des Chartes.

Gurvitch, G. (1961). *La multiplicité de temps sociaux*. Centre de documentation universitaire, Paris.

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

Hamilton, J. D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, 57(2):357–84.

Harte, D. (2015). *HiddenMarkov: Hidden Markov Models*. Statistics Research Associates, Wellington. R package version 1.8-3.

Hartog, F. (2003). *Régimes d'historicité. Présentisme et expérience du temps*. Seuil, Paris.

Hobsbawm, E. J. (2008). *L'âge des extrêmes : histoire du court XXe siècle, 1914-1991*. Le Monde Diplomatique, Bruxelles.

Jin-Guan, D. and Yuan, L. (1991). The integer-valued autoregressive (inar(p)) model. *Journal of Time Series Analysis*, 12(2):129–142.

Killick, R., Eckley, I., and Haynes, K. (2014). *changepoint: An R package for changepoint analysis*. R package version 1.1.5.

Killick, R. and Eckley, I. A. (2014). changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19.

Koselleck, R. (2000). *Le futur passé : contribution à la sémantique des temps historiques*. Éd. de l'École des hautes études en sciences sociales, Paris.

Kuhn, T. S. (1993). *La structure des révolutions scientifiques*. Flammarion, Paris.

Labrousse, E. (1970). *Histoire économique et sociale de la France. Des derniers temps de l'âge seigneurial aux préludes de l'âge industriel (1660-1789).* Presses universitaires de France, Paris.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34:1–14.

Levi, G. (1989). *Le Pouvoir au village. Histoire d'un exorciste dans le Piémont du XVIIe siècle*. Gallimard, Paris.

Mathis, A. and Grenier, J.-Y. (1991). Séries temporelles, structure et conjoncture : le prix du blé à l'époque moderne. *Histoire & Mesure*, 6(1):51–76.

Milo, D. S. (1991). Le siècle, projet expérimental avorté. In Dumoulin, O. and Valéry, R., editors, *Périodes. La construction du temps historique. Actes du Ve colloque d'histoire au présent*, pages 123–128. Editions de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33:341–365.

Olteanu, M. and Ridgway, J. (2012). Hidden markov models for time series of counts with excess zeros. In *20th European Symposium on Artificial Neural Networks, ESANN 2012, Bruges, Belgium, April 25-27, 2012*.

Prost, A. (1996). *Douze leçons sur l'histoire*. Seuil, Paris.

Ricœur, P. (1991). *Temps et récit*. Seuil, Paris.

Ridout, M., Demetrio, C., and Hinde, J. (1998). Models for count data with many zeros. In *Proceedings of the XIXth International Biometric Conference*, pages 179–192.

Rosso, C. (1992). *Una burocrazia di antico regime. I segretari di Stato dei duchi di Savoia, I (1559-1637).* Torino.

Rydén, T. (2008). Em versus markov chain monte carlo for estimation of hidden markov models: a computational perspective. *Bayesian Anal.*, 3(4):659–688.

Steutel, F. and van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *The Annals of Probability*, 7:893–899.

Wang, P. (2001). Markov zero-inflated poisson regression models for a time series of counts with excess zeros. *Journal of Applied Statistics*, 28(5):623–632.

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

Appendix A: ANOVA results for the monthly belligerence status

Global series, whole production of law

	Analysis of variance table				
	DF	Sum of Sq	Mean Sq	F-value	$\mathbb{P}(>F)$
Global texts	1	208.1	208.092	23.688	1.279e-06 ***
Residuals	1234	10840.4	8.785		

Kruskal-Wallis chi-squared = 29.7693	df = 1	p-value = 4.866e-08

Logistics series, documents on military logistics

	Analysis of variance table				
	DF	Sum of Sq	Mean Sq	F-value	$\mathbb{P}(>F)$
Logistics texts	1	19.42	19.4218	41.383	1.786e-10 ***
Residuals	1234	579.14	0.4693		

Kruskal-Wallis chi-squared = 40.1696	df = 1	p-value = 2.328e-10

Appendix B: Examples of data segmentation using change-point detection

Three change-point detection algorithms, available in the **R**-package **changepoint** (Killick and Eckley, 2014; Killick et al., 2014), were tested on the time series containing the number of documents on military logistics monthly issued. The first one, illustrated in Figure 8, contains the change-points detected in mean and variance, and with the hypothesis of an underlying Poisson distribution within each segment. The estimated values of the parameters for the Poisson distributions are the following:

$\hat{\lambda}$	0.13	0.20	0.36	1.23	0.36	1.22	0.52	0.12

According to these estimations, some of the segments appear to be controlled by the same underlying distribution. At the same time, the interpretation one can draw from these results is less interesting than in the hidden-Markov framework, especially the first and the last segments, which are historically meaningless.

For illustration purposes only, Figures 9 and 10 contain the change points detected either in mean or in variance, but with the assumption that the underlying distribution is Gaussian (the only option currently available in the **changepoint** package). The latter hypothesis is clearly not suited for our data, which are integer valued, and the quality of the results is the consequence of it.

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

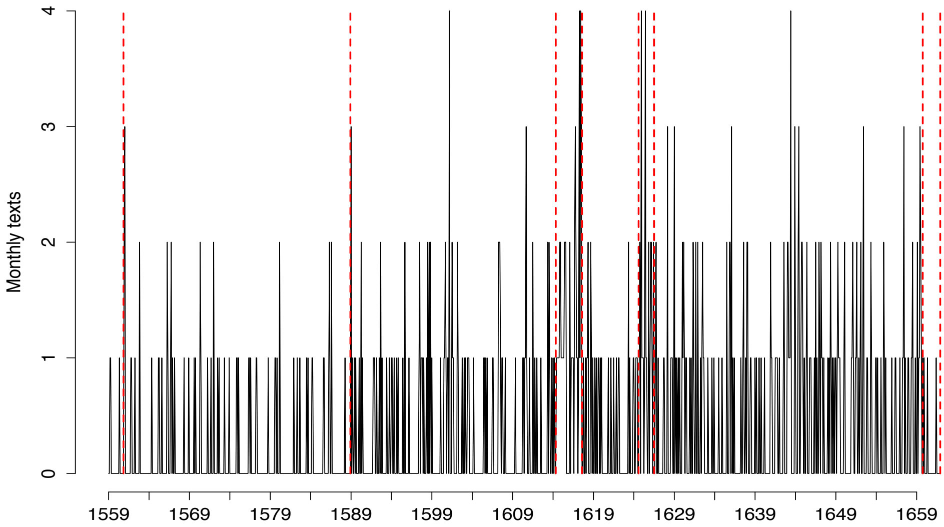

FIGURE 8. *Change-points detected in mean and variance with a Poisson underlying distribution*

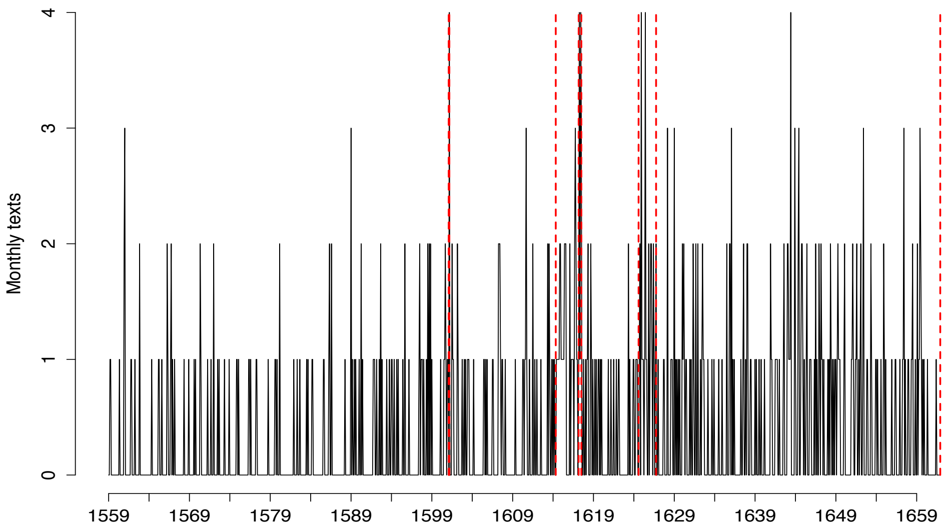

FIGURE 9. *Change-points detected in mean with a Gaussian underlying distribution*

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

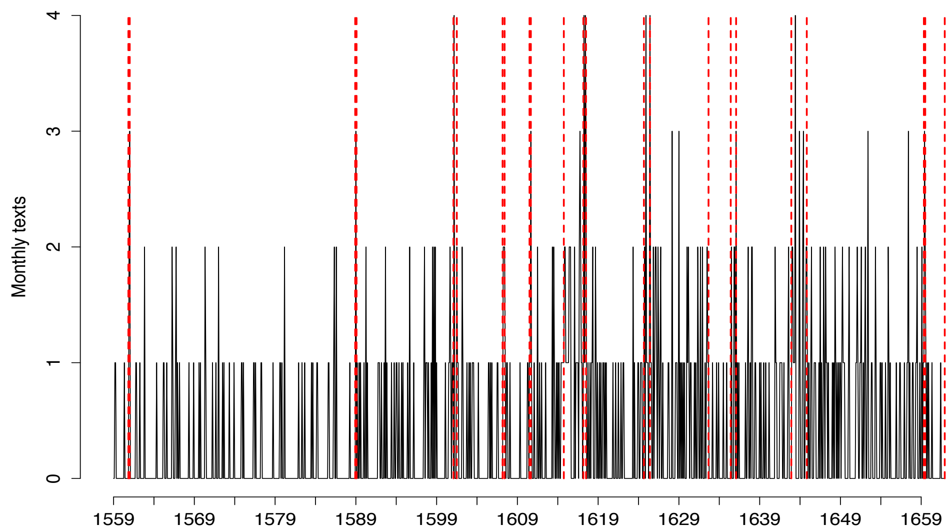

FIGURE 10. *Change-points detected in variance with a Gaussian underlying distribution*

Appendix C: Estimated parameters for the three-state ZIP-HMM

When training a ZIP-HMM model with three hidden states, the resulting estimated parameters are the following :

$$\hat{\pi} = \begin{pmatrix} 0.94 & 0.02 & 0.04 \\ 0.13 & 0.87 & 0.00 \\ 0.03 & 0.00 & 0.97 \end{pmatrix} \quad \hat{\lambda} = \begin{pmatrix} 0.53 \\ 1.18 \\ 0.33 \end{pmatrix} \quad \hat{\eta} = \begin{pmatrix} 0.00 \\ 0.00 \\ 0.40 \end{pmatrix}.$$

In Figure 11, the a posteriori probabilities of belonging to the third regime of the model, as well as their values thresholded at 0.5 are plotted. When entering into this regime, the activity of the State in issuing law related to military logistics is the less intensive. This regime is very similar to the first regime of the two-state ZIP-HMM estimated in the previous sections and illustrated in Figure 5. However, the three-state model is much more instable than the two-state model, the switches from one state to another occurring more often and being more difficult to interpret. The first and the second regime of the three-state ZIP-HMM share some common features: for both of them, the η parameter is estimated equal to zero, hence a Poisson distribution only is sufficient for describing these states. Also, the λ parameter in the second regime is more than twice the λ parameter in the first regime, meaning that the production of law is twice more important in the second regime. It is during this second regime, that the State is issuing most of the law.

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

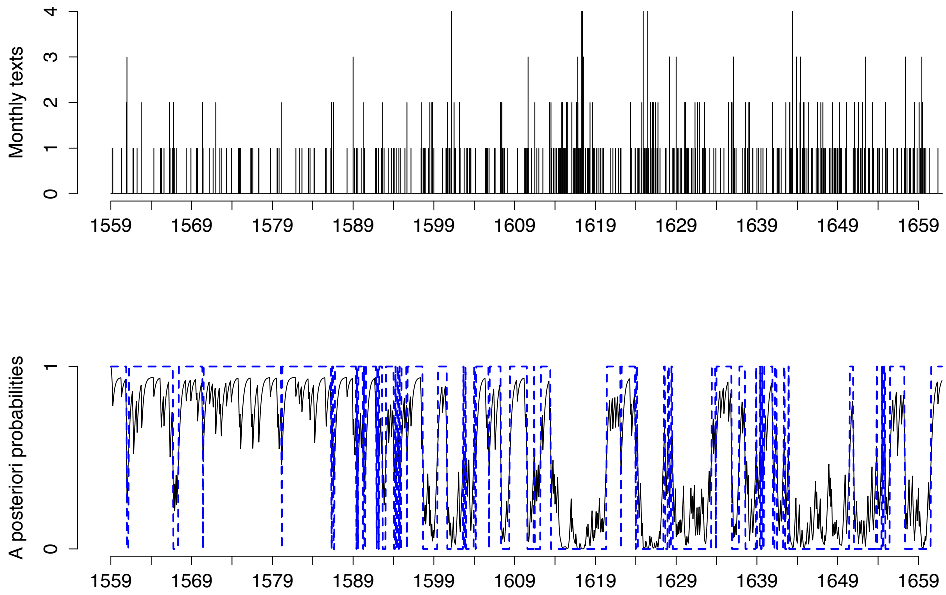

FIGURE 11. *A posteriori probabilities of the third regime for a three-state ZIP-HMM*

Appendix D: Simulated examples of time-series obtained with the estimated models in Section 5

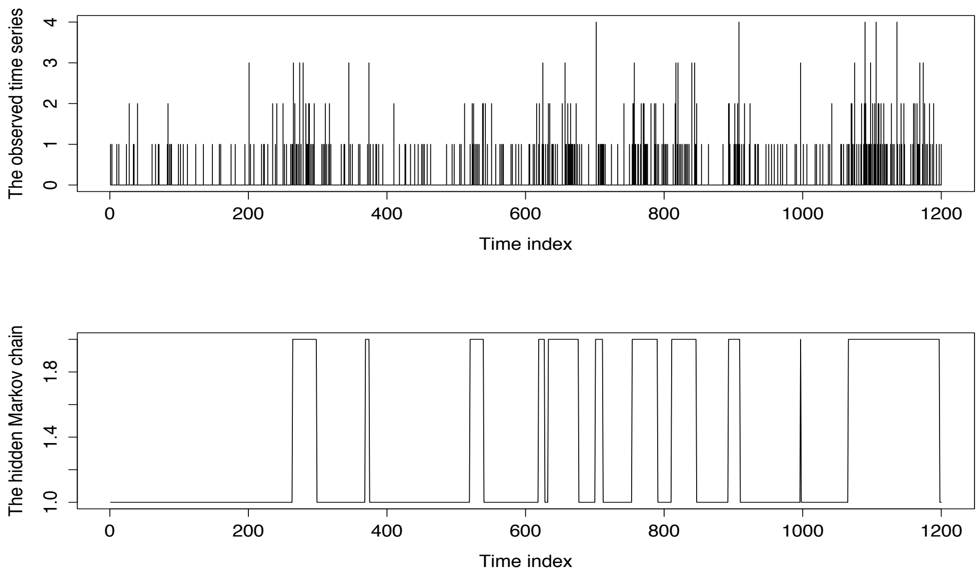

FIGURE 12. *A simulated time series and the corresponding regimes with the estimated ZIP-HMM model*

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238

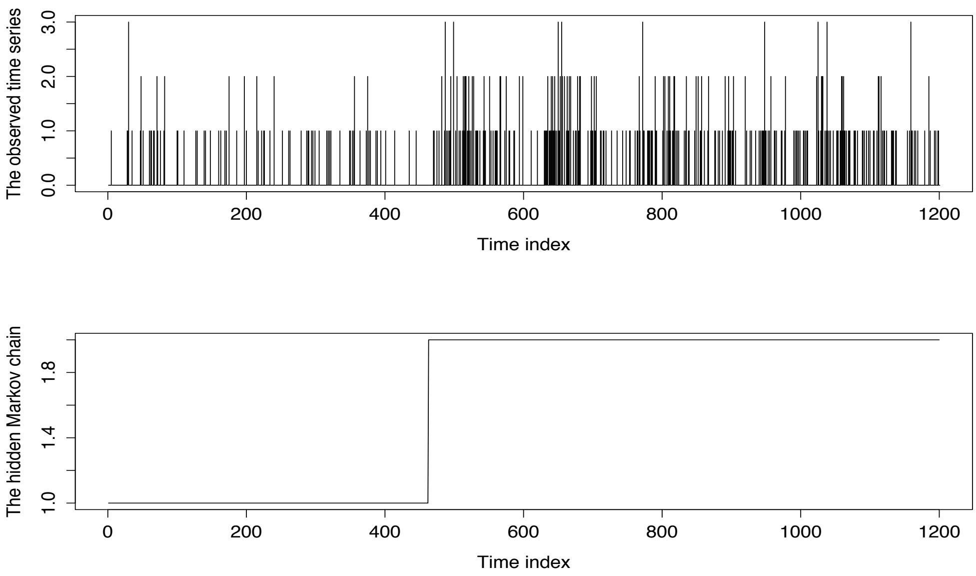

FIGURE 13. *A simulated time series and the corresponding regimes with the estimated INAR(6)-HMM model*

© Socit Franaise de Statistique et Socit Mathmatique de France (2017) ISSN: 2102-6238