Studying Media Events through Spatio-Temporal Statistical Analysis

Angelika Studeny, Robin Lamarche-Perrin, Jean-Marc Vincent

Studying Media Events through Spatio-temporal Statistical Analysis

Angelika Studeny
INRIA Grenoble Rhône-Alpes, Laboratoire d'Informatique de Grenoble, France
angelika.studeny@inria.fr

Robin Lamarche-Perrin
Max-Planck-Institute for Mathematics in the Sciences, Leipzig, Germany
Robin.Lamarche-Perrin@mis.mpg.de

Jean-Marc Vincent
Univ. Grenoble Alpes, Laboratoire d'Informatique de Grenoble, France
Jean-Marc.Vincent@imag.fr

Research Report

submitted September 2015
as delivrable L.3.2 in the context of the ANR project CORPUS GÉOMÉDIA

Abstract

This report is written in the context of the ANR Geomedia and summarises the developement of methods of spatio-temporel statistical analysis of media events (delivrable 3.2).

This documents present on-going work on statistical modelling and statistical inference of the ANR GEOMEDIA corpus, that is a collection of international RSS news feeds. Central to this project, RSS news feeds are viewed as a representation of the information flow in geopolitical space. As such they allow us to study media events of global extent and how they affect international relations. Here we propose hidden Markov models (HMM) as an adequate modelling framework to study the evolution of media events in time. This set of models respect the characteristic properties of the data, such as temporal dependencies and correlations between feeds. Its specific structure corresponds well to our conceptualisation of media attention and media events. We specify the general model structure that we use for modelling an ensemble of RSS news feeds. Finally, we apply the proposed models to a case study dedicated to the analysis of the media attention for the Ebola epidemic which spread through West Africa in 2014.

Résumé

Ce document présente les résultats d'un travail en cours sur la modélisation statistique et l'inférence appliqué au corpus de l'ANR GEOMEDIA qui est une collection des flux RSS internationaux. Au coeur du projet, les flux RSS sont considérés comme un marqueur représentatif des flux d'information dans l'espace géopolitique mondial. En tant que tel, ils nous permettent d'étudier des événements médiatiques globaux et leur impact sur les relations internationales. Dans ce contexte, on émet l'hypothèse que les modèles Markoviens cachés (HMM) constituent un cadre méthodologique adapté pour modéliser et étudier l'évolution des événements médiatiques dans le temps. Ces modèles respectent les propriétés des données, comme les corrélations temporelles et les redondances entre flux. Leur structure caractéristique correspond à notre conceptualisation de l'attention médiatique et des événements médiatiques. Nous spécifions la structure général d'un modèle HMM qui peut être appliqué a la modélisation simultané d'un ensemble des flux RSS. Finalement, on teste l'intérêt des modèles proposés à l'aide d'une étude de cas dédié à l'analyse de l'attention médiatique pour l'épidémie d'Ebola en Afrique de l'Ouest en 2014.

Contents

1 Introduction

Modern technology has led to a new form of journalism and opened the possibility for an almost continuous flow of information [Allan, 2006]. With few exceptions, newspapers nowadays publish online alongside their printed versions, by means of RSS feeds. Not astonishingly, this new and dynamic media landscape constitutes an interesting study object for various scientific disciplines [Mitchelstein and Boczkowski, 2009].

The GÉOMÉDIA research project unites media experts, geographers and computer scientists to study international relations through the analysis of media coverage of geopolitical events [Géomédia Doc Scientifique, 2011]. In particular, this project relies on the hypothesis that one can extract consistent information from the data flow of online newspapers to represent and understand the dynamics of international relations. As they are more or less covered by different media – depending on geographical, political and cultural proximities – the trace left by geopolitical events in the media space already provides interesting insights into international relations [Grasland et al., 2011]. In this context, the *media attention*, measured as the number of event-related publications in a given period, is taken as an indicator of the current state of the international system. Within this project, a database of newspaper articles aggregates the daily publications of approximately 300 RSS feeds across a selection of international and national newspapers. Each item in the resulting corpus can be tagged with event-specific keywords, beforehand assembled in dictionaries by expert knowledge, in order to estimate media attention for various events of interest.

In this report, we primarily aim at describing the evolution of media events by appropriate statistical models. Hence, these models need to be able to adequately take account of the properties of such data, *i.e.* multivariate time series of counts. In particular, we may expect temporal correlation in the number of event-related publications as well as correlation in media attention across RSS feeds. Furthermore, a modelling framework should allow for the inclusion of additional properties of RSS feeds which might influence their publication rate. Such "covariate information" incorporates feed-specific characteristics and explains heterogeneity between feeds. To the best of our knowledge, there has not been any attempt of advanced statistical modelling for such corpus data, which is able to take account of auto-correlation as well as covariate information.

RSS news feeds are not only a mirror of geopolitical events, but also a filter, in two senses. First, it is evident that an RSS feed will not be able to do justice to the complexity of the entire geopolitical space. Its data flow hence contains only a simplified representation of international relations. By tracing several different RSS feeds at the same time, one might hope to partly overcome this limitation and to capture more of the inherent complexity of the international system. Second, simultaneous occurrences of geopolitical events and the inability of newspapers to consistently report every such event both force editors on a daily basis to select events to cover in practice. Such editorial decisions depend on various

factors, which are usually not explicit and have therefore to be considered as a black box. The chosen modelling approach also tries to account for this inaccessible "hidden" part of the media landscape.

With this in mind, we want to propose an adequate statistical modelling framework for the temporal evolution of media events. This objective requires:

— To determine significant changes in the media attention towards a geopolitical event;
— To evaluate the influence of external and internal explanatory factors;
— To study the differences between news feeds in attention for an event.

We propose to use Hidden Markov Models (HMMs) to describe the evolution of a media event. HMMs are flexible discrete models of time series which are especially appropriate for data sets of small counts which are unlikely to fulfill the normality assumptions of classical time series models [Zucchini and MacDonald, 2009]. Their model structure accommodates our concept of RSS feeds as an indicator of the media attention towards an event: A latent (unobserved) state process contains temporal correlation as well as correlation in attention across RSS feeds and determines the stochastic behaviour of the observations, *i.e.* the number of event-related publications. From the data, we can deduce the sequence of states of global media attention, as well as locate changes in these states over time. Covariate information can be included in the model specifications for the observation process. We investigate the applicability of such models as well as provide a showcase that can be used for the analysis of other media events: The Ebola epidemic in West Africa in 2014 and its media reception is taken as an example of an event of global extent.

This report concentrates on modelling the evolution of media events rather than their geographical implications. However, by taking account of the geographic structure of the media space, our method is also capable of revealing aspects of the event-specific topology. There are two possible ways to integrate spatial information about a media event in our model: (1) analysing potential co-citations in the content of RSS items, hence embedding the event within some geographic locations (cited countries within the article) or (2) analysing the location of the reporting RSS feed itself, hence connecting the event to the place where the media information is actually produced (country of publication). We consider the latter option in this report.

The rest of the report is structured as follows. Assuming that HMMs have, in general, not been applied much in the context of social sciences, the potential reader might not be familiar with this kind of models. After a summary of the main concepts and the data characteristics of the corpus GÉOMÉDIA (Section 2) and the introduction of some notations (Section 3), we therefore provide an overview over the relevant theory in Section 4. In particular, we introduce the models considered in this report in Section 4.4. The Ebola epidemic is analysed as a case study in Section 5. We conclude with a discussion of this first modelling proposition, where possible extensions as well as alternative models are considered (Section 6).

2 Media Attention and Events in the Media

2.1 The Concept

Before discussing the details of a statistical modelling framework, the conceptualisation behind our modelling approach is laid out in the following section. Here, as in the wider context of the GÉOMÉDIA project, news are considered as flows of information regarding events in time. The media channel this flow of information as they allocate their attention depending on the newsworthiness atrributed to an event. This view includes certain hypotheses of how the media allocate their attention to specific events. In particular, we assume here that the newsworthiness of a specific geopolitical event is defined by the global attention which is allocated to it. This can be described by a global state of alertness of the media as an ensemble towards this event. For the moment, we assume that this event-related global state is shared by all the media. Thus a media event can be defined by the accumulation of information concerning the event. Hence an event becomes a media event because it gets into the focus of attention of the ensemble of the media and is characterized by the high correlation in attention across different media, in our case RSS feeds. To study the flow of information regarding a specific event, we are hence interested in the temporal evolution of the media attention. In time, the media attention towards an event changes and hence the global state of attention varies. To simplify the modelling task, we also do not consider competition for media attention between different events here. In terms of modelling the evolution of a media event, one of our interests is in detecting changes in this state process corresponding to an increase or decrease in media interest for the event. The global media attention modulates the response of the individual media along with other factors which can be related to the type or other characteristics of the media as well as their geopolitical relation towards the event. The central hypothesis of the GÉOMÉDIA project is that the media attention for geopolitical events can be *measured* via markers of the textual content of RSS news feeds published by international newspapers. The properties of these data are described in more detail in the next section. By analysing the temporal evolution of the appearance of an event in the news feeds, we can draw inference on the allocation of media attention in time. Adding geographical information enables us to deduce information about international relations and the geopolitical space.

2.2 The Corpus GÉOMÉDIA

— *Data collection.* With the intention to study these hypotheses, a database of RSS news feeds is currently collected at the CIST (Collège Internationale des Science du Territoire, in Paris), in the context of the GÉOMÉDIA project [Géomédia, 2015]. This database contains time-stamped sequences of news items (textual information) structured by feed identity. It is built in the following way: A web application has

been developed and has been running in a stable version since December 2013, while new RSS feeds have still been added up to May 2014. The application checks for updates of the feeds at an hourly basis and hence new publications are stored at almost real time. This database integrates 300 RSS news feeds from a selection of 160 international journals, spanning different categories ("international", "national", "general", "une", "breaking news", "unique") and 8 languages; the majority of publications are in English (55%) followed by articles in Spanish and French (17% and 14%, respectively). For each published news item, the title and a text content [1] are stored along with a time stamp (currently the date and time of reception by the CIST server) as well as some meta-data. A data processing pipeline allows the user to process the raw data and extract a textual corpus from this database. A number of selection criteria can be specified, such as the choice of language, the observation period, the RSS feeds category, and some other criteria. This is then followed by routines to clean the data which may contain double entries.

— *Tagging media events.* Event-related publications are then identified through an automated tagging procedure in order to obtain the textual marks to be analysed. To this end, a dictionary of event-characteristic key words has to be assembled by expert knowledge beforehand and in such a way that the risk of false-positive or false-negative identification is low. Based on such a set of key words, a 0/1-indicator (tag) is automatically associated to each RSS news item in the corpus if the item contains at least one key word where the search for keywords can be specified as only in the title or in both title and the textual content.

— *Data characteristics.* Such a series of tagged news items differs from standard textual corpus data as it comes with additional characteristics, in particular a temporal component. The RSS news feed on its own is essentially a time series of published items, as is the series of tagged items. As such, we are faced with temporally correlated data (**longitudinal correlation**). In addition, cycles, trend or changes in the publication rate over time may occur, independently from the specific media event under study. This needs to be modelled adequately. Secondly, we analyse an ensemble of RSS feeds and hence multivariate data. As a media event is by definition followed by the ensemble of the media, we expect inherent correlation between feeds (**contemporaneous correlation**). While the data collection works in continuous time, typical time scales for media events usually lead to discretisation and aggregation of the data in time. This usually also facilitates the data analysis and guarantees sufficient amount of data. After fixing a time scale, such as *e.g.* days or weeks, the data is transformed to counts of items and textual marks per time unit.

We detail below how these concepts and the particular data characteristics are taken into consideration by our proposed modelling framework.

1. In most cases, this is a summary of the full article given by the RSS feed. More rarely, it actually contains the full article or only a URL link.

3 Formal Description and Definitions

In this section, formal definitions of the constituting elements of the GÉOMÉDIA corpus are given. This allows us to easily translate them into mathematical terms for modelling purposes.

Definition 1 (RSS feed)**.** *An RSS feed G is a source of information which, in the course of time, publishes pieces of information in a structured manner (items) on the internet. Thus an RSS feed can be viewed as an (ordered) sequence of items $(\omega_t)_{t\geq 0}$. As a source of information, an RSS feed carries several characterising labels which constitute the profile of the feed:*

- *its name which presumably allows to uniquely identify it*
- *its language*
- *possibly its geographic location*
- *its temporal aspects (date of creation of the feed, frequency of publications)*
- *other meta-data*

Definition 2 (RSS item)**.** *An item ω_t is the constituting element of an RSS feed. It is published at a precise point in time, t. This date of publication is the moment from which onwards the item can be considered accessible on the web. Thus, ω_t is a time-stamped chain of words. If not specified otherwise, this chain of words consists of a title and some textual content.*

Definition 3 (Item tag)**.** *Let $\mathcal{G} = \{G^1, \ldots, G^K\}$ be a set of RSS feeds and $T = [t_0, t_1]$ with $t_0 \leq t_1$ be an interval of time. Furthermore, let $\omega_T^k = (\omega_t^k)_{t_0 \leq t \leq t_1}$ be the sequence of RSS items published by feed G^k, with $k \in \{1, \ldots, K\}$, in the interval T. Let $\mathcal{D}$ be a dictionary,* i.e. *a set of keywords. A* **tag** *is an indicator function*

$$\mathbf{tag} : \omega_T^k \longrightarrow \{0, 1\}, \quad \mathbf{tag}(\omega) \mapsto \begin{cases} 1 & \text{if } \omega \cap \mathcal{D} \neq \emptyset \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

In terms of the modelling task, the temporal dimension can be viewed as either continuous or discrete. Here, the modelling framework we choose necessitates discretising time. This implies a choice on the time scale which one expects to be characteristic for the phenomenon under study. In case of a global event of substantial duration, such as the Ebola epidemic, observations aggregated on either a daily of weekly basic are a good choice in order to capture the time scale of the event as well as to retain a sufficient amount of data points. But in general, the time scale chosen should relate to the order of time characteristic to the media event under study.

Hence, we discretise the interval $T = [t_0, t_1]$ by setting a fixed discretisation length Δ and considering the time steps $t_\ell = t_0 + \ell\Delta$ with $\ell = 1, 2, \ldots, L$ where $L = \left\lfloor \frac{t_2 - t_1}{\Delta} \right\rfloor$. ($\Delta$ will often be chosen such that the time steps t_ℓ correspond to days or weeks.)

We are interested in modelling the following aggregated quantities:

— the vector $\mathbf{N}^k = (N^k_\ell)_{\ell=1,2,\ldots,L}$ where $N^k_\ell = \left|\omega^k_{[t_{\ell-1},t_\ell[}\right|$ which is the number of items published by feed G^k during the time period $[t_{\ell-1}, t_\ell[$.

— the vector $\mathbf{X}^k = (X^k_\ell)_{\ell=1,2,\ldots,L}$ where $X^k_\ell = \left|\omega \in \omega^k_{[t_{\ell-1},t_\ell[} : \mathbf{tag}(\omega) = 1\right|$ which is the number of items published by feed G^k during the time period $[t_{\ell-1}, t_\ell[$ and carrying a tag.

4 Modelling Media Events

4.1 Preliminairies: Probabilistic Models for Count Data

Having defined the key quantities of interest, we would like to proceed with an adequate probabilistic model to describe their behaviour which we assume to be stochastic. Hence we need to make some distributional assumptions. Both, N_ℓ^k and X_ℓ^k, are sequences of counts. The following give two standard models for count data.

Definition 4 (Poisson Distribution)**.** *A random variable Z taking values in the set of positive integers $\{0, 1, 2, \ldots\}$ follows a Poisson distribution, $Pois(\lambda)$, iff*

$$\mathbb{P}(Z = n) = \frac{\lambda^n}{n!} e^{-\lambda}.$$

The parameter λ is called the rate *and it characterises completely the moments of the distribution*

$$\mathbb{E}[Z] = \mathbb{V}[Z] = \lambda.$$

Typically, a Poisson distribution models the probability of observing a certain event happening several times in a fixed interval (in time or in space) where the occurrences of the event take place independently of each other and with a constant rate. For fixed ℓ, the Poisson distribution provides a possible model for N_ℓ^k as well as the number of news items *about Ebola*, X_ℓ^k.

As items which carry a tag form a subset of items published in time unit, we can adopt yet another point of view and consider them in terms of their proportion of the total number of published items. In other words, we are interested in the probability that within the N_ℓ^k items published, there are exactly n about Ebola. In this case, we consider N_ℓ^k as a known nuissance parameter, i.e. it is not estimated from the data but fixed in advance. This leads to a binomial model.

Definition 5 (Binomial Distribution)**.** *A random variable Z in $\{0, 1, 2, \ldots, N\}$, where N is known, follows a binomial distribution, $B(N, \pi)$, iff*

$$\mathbb{P}(Z = n) = \binom{N}{n} \pi^n (1 - \pi)^{N-n}$$

with $\pi \in [0, 1]$. This distribution is characterised by the two parameters, N and π, where $\mathbb{E}[Z] = N\pi$ and $\mathbb{V}[Z] = N\pi(1 - \pi)$.

N is called the number of trials and is usually determined by the situation we are interested in; in our case $N = N_\ell^k$ for fixed ℓ. The parameter of interest is π. For example, in the case of news items published on Ebola, π is low when there is little media attention and it increases with the media interest regarding Ebola in the course of the crisis.

For modelling a set of observations, General Linear Models (GLM) have been introduced for data which follow non-normal distributions, such as the Poisson or the Binomial distribution [McCullagh and Nelder, 1989, Hastie and Tibshirani, 1990]. However, as in the standard linear modelling framework, these models rely on important assumptions — the independence of the observations, constant variance and a specific mean-variance-relationship. The latter means that the variance can be expressed as a function of the mean, as *e.g.* the identity of mean and variance in the Poisson model. Although there are some extensions to the GLM framework to deal with minor violations of these assumptions, it is clear that these are unlikely to hold for RSS feeds. *E.g.*, the rate of publication, represented by the parameters λ and π respectively for the two distributions, is variable as it is determined by the evolution of the media attention for an event. Hence, the variance which is a function of λ or π cannot be constant for all observations. More generally, as previously discussed, we are faced with time series data and hence important temporal correlation can be expected.

GLMs cannot include the temporal auto-correlation nor systematic changes in the mean-variance-relationship. Indeed diagnostic plots evaluating the fit of a preliminary model in a GLM framework showed that the data structure is not captured well. We propose a different modelling approach here in form of Hidden Markov Models. Both the Poisson distribution and the Bionomial distribution can be integrated in the framework of Hidden Markov Models and thus in a time series context. We discuss below which of the two probabilistic models makes more sense in the context of media analysis.

4.2 Introduction to Hidden Markov Models (HMMs)

Hidden Markov Models (HMM) are a well-established modelling tool for time series data [Zucchini and MacDonald, 2009]. This class of models owes its popularity to a highly flexible modelling framework, which has led to their successful application in a varity of disciplines such as speech recognition [Rabiner, 1989], finance [Rossi and Gallo, 2006], ecology [Schliehe-Diecks et al., 2012] and genetics [Eddy, 1996], amongst others. They offer the possibility to overcome the often too restrictive assumptions of classical time series models. In particular, they allow us to model discrete-valued time-series, such as counts, which do not comply to the normality assumption of the classical time series framework.

The general structure of an HMM is shown in Fig. 4.1. HMMs offer a modelling tool for dynamic systems which are observed through realisations of a stochastic process. They consist of two components, namely the observation process and the latent (not directly observable) state process. In our case, the observation process is the sequence of counts of event-specific publications as given by the tagged data on the time scale determined by Δ (*e.g.*, per day or per week). The latent process can be interpreted as a sequence of different levels of media attention towards the event (see section 2.1). Depending on the global state of interest in the event, a feed changes either the average publication rate λ about the event (Poisson model) or the probability π of publishing an article on the event (Binomial model). Later, we show how this can further be modulated by journal-implicit factors. Thus, the publication rate is time-varying and driven by the evolution of the state of media attention. The upper part of Fig. 4.1 shows the case of three different level of media attention towards some event. Consider the case of a Poisson model: Over time, media attention increased from an intermediate state (encoded by $S_1 = 2$) with low event-related publications to a state $S_2 = 3$ with high media interest. This corresponds to different publication rates λ_2, λ_3 with $\lambda_2 < \lambda_3$, hence an increase in the mean number of publications. It then drops to a base level $S_3 = 1$ with little or no publications ($\lambda_1 < \lambda_2$) and then alternates between these different states. At each point in time ℓ, the value of the state $S_\ell = i$ determines thus the publication rate λ_i. In turn, the number of publications X_ℓ is then considered a random realisation of a Poisson distribution with parameter λ_i.

In general, let $\mathbf{S} = (S_\ell)_{\ell=1,2,\dots}$ be the evolution of states of media attention for an event in time. As indicated by the name, this latent process of the HMM is assumed to evolve as a Markov chain. *I.e.*, the probability for a change in state at time step $t_{\ell+1}$ is assumed to depend only on the current state at time t_ℓ,

$$\mathbb{P}(S_{\ell+1} = j | S_1, \dots, S_\ell) = \mathbb{P}(S_{\ell+1} | S_\ell) \tag{4.1}$$

and can thus be written as $\gamma_{ij}(\ell) = \mathbb{P}(S_{\ell+1} = j | S_\ell = i)$. This probability may either be a function of time or be independent of the value of ℓ. In the latter case, one refers to the model as a homogeneous HMM. Depending on the current state S_ℓ at time t_ℓ, the observation X_ℓ (*i.e.*, the number of event specific publications) is realised according to a stochastic model. For an observation process according to a Poisson model, the counts are

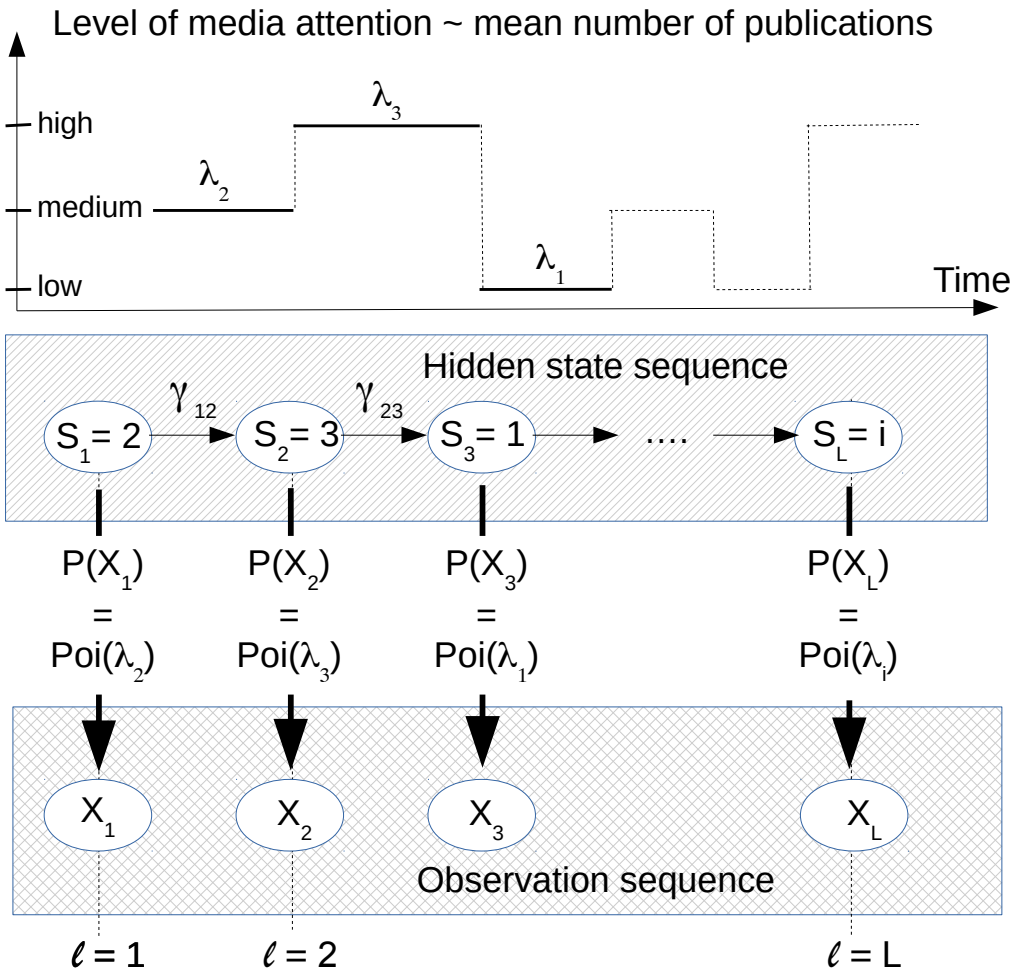

Figure 4.1 – Schematic of the modelling structure of a Poisson Hidden Markov Model with 3 states and its interpretation for RSS news feeds. The evolution of the hidden state sequence in time is determined by the transition probabilities γ_{ij}. The state at time step ℓ determines the parameter λ of a Poisson distribution which in turn generates the observation x_ℓ. In this example λ can assume one of three values with $\lambda_1 < \lambda_2 < \lambda_3$ corresponding to an increase in media attention and hence and increase in the expected number of publications. The total number of parameters in the model equals 9 (6 for the transition probabilities and 3 state dependent values of λ).

distributed according to

$$\mathbb{P}(X_\ell = n|S_\ell = i) = \mathbb{P}_{\lambda_i}[X_\ell = n] = \frac{(\lambda_i)^n}{n!}e^{-\lambda_i}. \tag{4.2}$$

Whereas for a Binomial model with probability π to publish an event-specific article, we have

$$\mathbb{P}(X_\ell = n|S_\ell = i) = \mathbb{P}_{\pi_i}[X_\ell = n] = \binom{N_\ell}{n}\pi_i^n(1-\pi_i)^{N_\ell - n} \tag{4.3}$$

where the actuality of the event is in state $S_\ell = i$ and where N_ℓ is the total number of publications by the feed at time t_ℓ. All temporal dependences are absorbed in the state process; given the state sequence, the observations are considered to be independent. This basic structure of an HMM can be extended in several ways to include covariate information and more complex dependence structures. The number of possible states needs to be specified before fitting the corresponding model. However, if several models with different number of states are fitted, model selection criteria, such as AIC or BIC, can be applied to determine the number of states which describe the data best.

4.3 A HMM for a Single RSS Feed

While our interest is in modelling a set of RSS feeds, we briefly present an HMM for a single news feed for illustration purposes of the potential applications within the HMM framework. Consider hence the RSS feed dedicated to "international news" of the Canadian journal *The Vancoucer Sun*, with items tagged for "Ebola" as explained above. This feed provides an interesting example, as the global publication rate changed significantly in April 2014. We restrict ourselves to the case of a basic 2-state model with a homogeneous Markov chain without any additional structure. More sophisticated models are considered in the next section. As transitions between states are assumed to be stationary in time, the transition probabilities (Eq. 4.1) can be summarised by a matrix

$$\mathbf{\Gamma} = \begin{pmatrix} \gamma_{11} & 1-\gamma_{11} \\ 1-\gamma_{22} & \gamma_{22} \end{pmatrix} \tag{4.4}$$

and the initial distribution $\boldsymbol{\delta} = (\mathbb{P}(S_0 = 1), \mathbb{P}(S_0 = 2))$ for the state at time t_0.

This simple 2-state HMM can already be applied to address different questions: First, we can fit such a model to the number N_ℓ^k of published RSS items thus locating the change in the overall publication rate by statistical means. In this case, we look at a Poisson model for the total number of daily items, *i.e.*, the two states correspond to different phases in the global behaviour of the feed, characterised by the average number of daily publications λ_1 and λ_2. Fitting the model consists in estimating the parameters γ_{11}, γ_{22}, λ_1 and λ_2. This is done by numerical optimisation of the likelihood where we adapt the R code given in [Zucchini and MacDonald, 2009]. The actual maximisation is carried out by

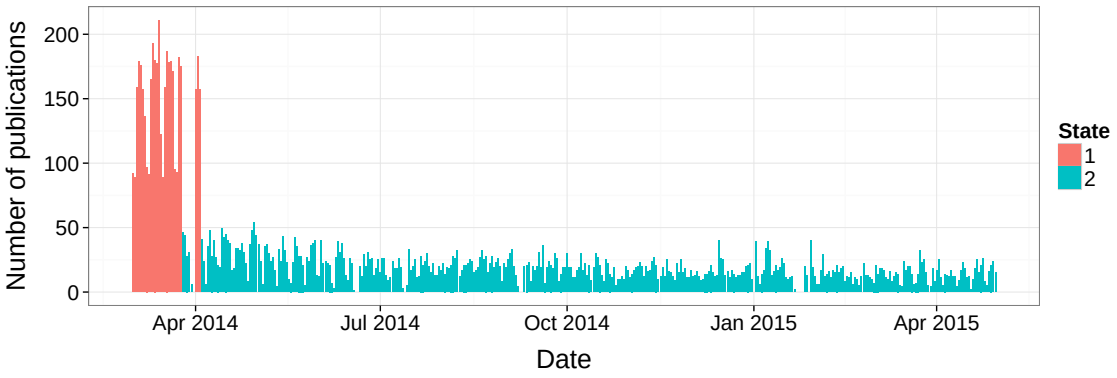

Figure 4.2 – A 2-state HMM with a Poisson distribution for the total number of RSS articles published by the *Vancouver Sun* between 1st of March 2014 and 30th of April 2015

the `nlm` package in R [R Core Team, 2014], and based on a Newton-type algorithm. The calculation of the likelihood can be implemented efficiently by recursion, considering its representation as a matrix product [1],

$$\mathcal{L} = \boldsymbol{\delta}\mathbf{P}_\lambda(x_1)\boldsymbol{\Gamma}\mathbf{P}_\lambda(x_2)\boldsymbol{\Gamma}\dots\mathbf{P}_\lambda(x_{L-1})\boldsymbol{\Gamma}\mathbf{P}_\lambda(x_L)\mathbf{1}' \tag{4.5}$$

where $\mathbf{P}(x_\ell)$ is a diagonal matrix with entries given by Eq. 4.2.

Based on the fitted model, the state sequence which most likely gave rise to the observations can be determined by locally (for each point in time) or globally (simultaneously) maximising the probability to observe a certain state or a certain state sequence, respectively. Fig. 4.2 shows the (locally) most probable state sequence for the fitted 2-state model for the total number of daily publications. Estimates for the publication rates are $\lambda_1 = 151$ and $\lambda_2 = 19$ and the corresponding states are quite separeted in time. *I.e.* We can distinguish two main phases for the *The Vancoucer Sun* where the average number of daily publications dropped from 151 before 25th March 2014 to 19 from 4th April 2014 onwards, with a short period of transition between the two states at the end of March. This signifies a change in the profile of the feed which can potentially affect its publication behaviour.

Next, the Ebola-related articles are modelled for the same feed. Since we dispose of the total number of items as well as the number of tagged items, two approaches are possible. A Poisson HMM models the counts of tagged items alone whereas a binomial model accounts for the proportion of tagged items to the total daily publications. While the former looks for changes in the number of articles reporting on Ebola, the result can be biased by the change in the overall publication rate. In this case, a change into state 2, *i.e.* an increase in publication rate for the Ebola event, is only picked up mid-October 2014 (Fig. 4.3(a)). On the other hand, in the binomial model the counts X_ℓ are considered as a proportion of the total number of publications N_ℓ. We expect to see $N_\ell\pi_i(\ell)$ articles on Ebola, where $\pi_i(\ell)$ is determined by the state $S_\ell = i$ with $i \in 1, 2$. The two states correspond to an *off*

1. The notation $\mathbf{1}'$ stands for a column vector of ones.

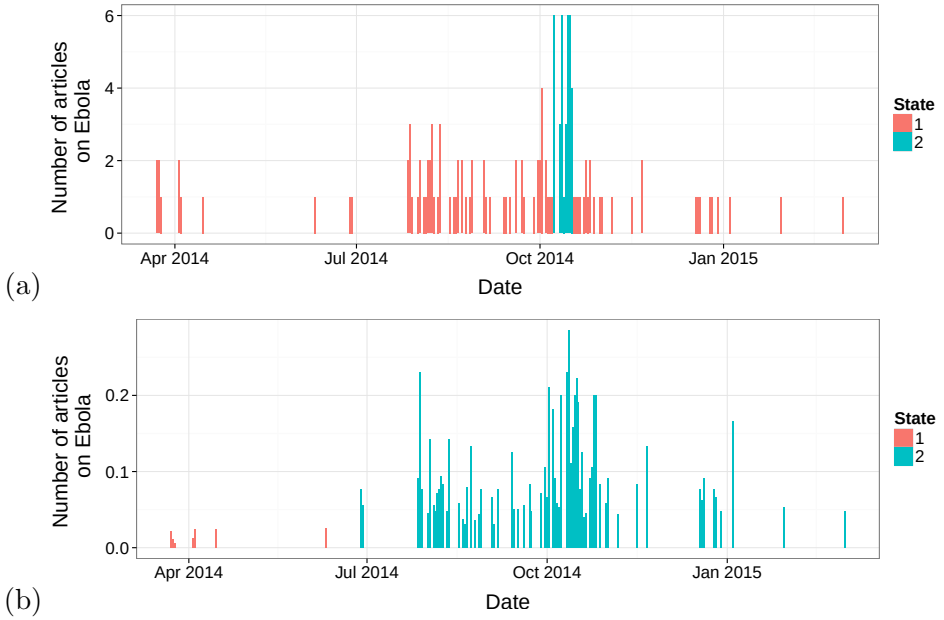

Figure 4.3 – 2-state HMMs for the number of RSS articles on Ebola published by the *Vancouver Sun* between 1st of March 2014 and 30th of April 2015 assuming (a) a Poisson model and (b) a Binomial model in the state dependent process. In contrast to Poisson model, the Binomial model considers the number of Ebola-related articles as a proportion of the total publications.

state, where there is no or only low coverage of the Ebola epidemic by the *Vancouver Sun*, and an *on* state, corresponding to a increased interest of the journal in the Ebola epidemic. Hence, we expect the probabilities of the binomial distribution in the second case, π_2, to be higher than π_1. However, even in phases of increased media interest, this probability remains small as the Ebola epidemic is still only one of many topics covered by the feed. Here, the estimated probabilities are $\pi_1 = .01$ and $\pi_2 = .09$. More importantly, for the binomial model we observe a switch to state 2 from the end of June 2014 onwards. This coincides with a declaration by Médecins Sans Frontières of the spread of the virus being "out of control" after it has reached several of the countries bordering Guinea. Again, four parameters had to be estimated: γ_{11}, γ_{22}, π_1 and π_2. Obviously, the number of parameters increases with the number of states and the model complexity. The estimated transition probabilities are $\hat{\gamma}_{11} = 0.992, \hat{\gamma}_{22} = 0.852$ in the case of the Poisson model and $\hat{\gamma}_{11} = 0.995, \hat{\gamma}_{22} = 0.997$ for the Binomial model. Hence, in the Binomial model the second state is more stable than for the Poisson HMM. Consequently, we expect to see less state variations and that the state sequence stays longer in state 2 than in the Poisson model.

While modelling a single RSS feed is not pursued further here, we point out that the models could be refined by considering a higher number of states and select the best fitting model by model selection criteria.

4.4 HMMs for Media Events Modelling

4.4.1 Modelling Options for a Set of RSS Feeds

After the example of a single news feed, we apply the HMM framework in the context of a media event as it is represented by a set of RSS news feeds.

There are different ways to combine the information provided by a set of feeds in an HMM:

— *Superposition of all feeds*: In this case, we build one HMM and compute one state sequence $S_1, \ldots, S_L$ to model the superposed data sequence $X_1, \ldots, X_L$ with $X_\ell = \sum_{k=1}^{K} X_\ell^k$. This is based on the point of view that a media event is by definition an event that is in the center of attention across all the media in the set. As the signal is reinforced by superposition and as the temporal fluctuations of single feeds are naturally smoothed, we can expect a stable model fit. However, the downside of such a model is that all heterogeneity between feeds is erased in the data pooling, and hence no inter-feed comparison is possible. Furthermore, superposition by a weighted sum might be more appropriate and more generally, other superposition operators than a sum are possible, too. Hence, this approach necessitates a decision about the actual form of the superposition. It is also not obvious how global changes that affect some feeds, such as in the example regarding the *Vancouver Sun* in the previous section, can be taken into account appropriately and posteriori inference on single feed is not possible.

— *Separate Models*: Alternatively, we build K HMMs, and compute one state sequence $S_1^k, \ldots, S_L^k$ for each, to independently model each data sequence $X_1^k, \ldots, X_L^k$, with $k \in \{1, \ldots, K\}$. This assumes that feeds behave as unrelated processes. A comparison between feeds is limited in this framework. In addition, one can expect numerical instabilities in the fitting procedures for the smaller, more sparse feeds. For these reasons, this modelling option is not pursued here.

— *Hybrid Model:* In a combination of the first two modelling options we may consider superpositions within subgroups of news feeds. We aggregate the K feeds into $C \leq K$ subgroups $(X_{k_c})_{k_c \subset \{1,\ldots K\}}$ and $k_{c_1} \cap k_{c_2} = \emptyset$ for $c_1 \neq c_2$. We build C HMMs and compute one state sequence $S_1^c, \ldots, S_L^c$ for each. Each of these state sequences model in turn the superpositions $X_\ell^c = \sum_{k \in k_c} X_\ell^k$ in subgroup c. In this case, we do not lose entirely the structure of the single feeds and have the possibility to look at semantically meaningful aggregates (see deliverable L.3.1). However the problem persists that the HMMs for the different subgroups cannot be reasonably compared. In such an approach, as with a complete superposition, we obtain a result for the entire subgroup which can not be readily decomposed at the single-feed level.

— *Product Model - Independent Realisations*: Finally, we can assume that the same hidden process drives the event histories across all feeds, and that conditional on the model, feeds behave independently. The parameters of the underlying Markov chain are the same for all feeds. The advantage of such an approach over the previous

one is the increase in stability of the model fitting procedure, the reduction in the number of parameters to be estimated and the possibility of inter-feed comparison. Two modelling options are possible in this case: The observed time series can be regarded as independent realisations of the same HMM. *I.e.*, we build one HMM and compute K state sequences $S_1^k, \ldots, S_L^k$, each one modelling a data sequence $X_1^k, \ldots, X_L^k$, with $k \in \{1, \ldots, K\}$.

— *Product Model - One Multivariate Realisation*: Alternatively, we can assume that only one state sequence drives the observations, corresponding to a global state of media attention toward the event (see Section 2.1). In this case, we build one HMM and compute one state sequence $S_1, \ldots, S_L$ to homogeneously model all the data sequences $X_1^k, \ldots, X_L^k$, with $k \in \{1, \ldots, K\}$ simultaneously. *I.e.* the time series $\mathbf{X} = (\mathbf{X_1}, \ldots \mathbf{X_L})$ with $\mathbf{X}_\ell = (X_\ell^1, \ldots, X_\ell^K)$ is considered as a multivariate observation. Heterogeneity between the coverage of the event by different feeds can be included in both models by covariate information (such as *e.g.* the size of a feed). However, the implicit correlation between feeds is only taken into account in the second approach. In this case, all information concerning heterogeneity between feeds has to be modelled in the observation process.

The different modelling options also correspond to restrictions on the parameter space. On the one end, fitting separate models for each feed allows for all parameters to vary freely and hence every model parameter is feed-specific. On the other end, the product model of a multivariate HMM with one state sequence implies that the same parameters are taken for all feeds. By including covariate information, we can allow some of the parameters to vary between feeds and thus take into account heterogeneity as well as correlation.

4.4.2 A Product Model for the Evolution of a Media Event

In this section, technical details regarding the model of a set of feeds are given. The model is then applied in the case study in Section 5. As explained above, our model choice is a multivariate HMM where one realisation of the state process is assumed to drive the time series of multivariate observations. We furthermore assume a Binomial model to account for differences in feed sizes and for potential changes in the overall publication rate N_ℓ^k for some of the feeds. In summary, the following modelling assumptions are made:

— The state process describes the global level of media attention.

— All feeds are driven by the same state sequence (**common state sequence**).

— Given the state sequence, observations are conditionally statistically independent in time (**longitudinal conditional independence**).

— Given a point in time and the current state of media attention, numbers of event related publications by different feeds are independent (**contemporaneous conditional independence**).

— The marginal distributions for the number of event-related publications are Binomials with state dependent probabilities π_i and feed-specific, time varying known parameter N_ℓ^k (**Binomial model**).

With the same notation as in 4.3 above, the likelihood for this model, given observations $\mathbf{x}_\ell = (x_\ell^1, \ldots, x_\ell^K)$ for a set of K feeds and time points $\ell = 1, \ldots, L$, is

$$\mathcal{L}_T = \boldsymbol{\delta}\mathbf{P}_\pi(\mathbf{x_1})\boldsymbol{\Gamma}\mathbf{P}_\pi(\mathbf{x_2})\boldsymbol{\Gamma}\ldots\mathbf{P}_\pi(\mathbf{x_{L-1}})\boldsymbol{\Gamma}\mathbf{P}_\pi(\mathbf{x_L})\mathbf{1}', \tag{4.6}$$

where $\mathbf{P}_\pi$ is a diagonal matrix with entries

$$P_\pi(x_\ell^1, \ldots, x_\ell^K) = \prod_{k=1}^{K} \binom{N_\ell^k}{x_\ell^k} \pi^{x_\ell^k}(1-\pi)^{N_\ell^k - x_\ell^k}. \tag{4.7}$$

Actually, the value of π depends on the state of the hidden process as described above, but the additional state subscript has been dropped in the notation to ease readability. As in the case of the model for a single feed, the likelihood can be computed efficiently making use of the recursive structure of eq. 4.7. It is optimised by a Newton-type algorithm, implemented in the `nlm` routine in R. The calculation requires reparametrisation of the transition probabilities γ_{ij} in order to dispose of unconstrained parameters for the optimisation. In addition, scaling steps are necessary to avoid numerical underflow of the probabilities in the product in eq. (4.7). The full R code for the model is an extension of [Zucchini and MacDonald, 2009] and can be found in Appendix B.

4.4.3 Extensions to the Observation Process Model

In this basic form, the model in Eq. 4.7 does not take into account potential differences between feeds. Moreover, because of the assumption of a unique state sequence, the sequence of state dependent parameters π in the observation process is also entirely determined, and the same for all feeds. In this section, we introduce possible extensions to the observation mode which enable us to include some heterogeneity.

The easiest, computationally least costly and most meaningful way is to include feed-specific covariate information in the state dependent parameter π. *E.g.* a covariate which is readily available is the size of the feed, as measured by the total number of publications per time unit (days or week), but others can be included in the same way. Covariates can be integrated by a GLM-like specification for the model parameter. In case of the size of a feed, the model for the probability π is specified by

$$\text{logit } \pi_i^k(\ell) = \alpha_{0,i} + \alpha_{1,i} N_\ell^k. \tag{Model I}$$

where the coefficients depend on the current state of the Markov chain. The logit link function is chosen in order to assure that the probability π is in $[0, 1]$. The aim of including covariate information in our case of a multivariate HMM with a common state sequence is to allow the state dependent probability π to vary between feeds and thus take account of heterogeneity between feeds. Particulary interesting in this context is the case of geographical covariate, *i.e.* feed-specific variables which are informative about the geographical setting. HMM are a modelling framework for discrete time series data and, as such, they do not belong to the toolbox of spatial statistical models per se. However,

including geographically meaningful terms in the observation model, allows us to infer about event-related geographical implications without an explicit spatial model. Here, a possible covariate is for example a factor which groups feeds by larger spatial regions. In this case, the definition of the regions needs some care to assure sufficient amount of data for a reliable estimation of the corresponding term. Another option would be the distance to the event; this implies a decision about the center of the event as well as assumes that there is a unique definition of the location of each feed. Here, we opt for the first modelling approach and group feeds by continent as European, African and American. This leads to the following model

$$\text{logit } \pi_i^k(\ell) = \alpha_{0,i} + \alpha_{1,i} N_\ell^k + \rho_{c(k),i} \qquad \text{(Model II)}$$

where $c(k) \in \{\text{Africa, America}\}$ [2] depending on the location of feed k. Again, model selection criteria, such as AIC or BIC, can be used to decide which of these models best describes the data, in other words which covariates should be kept in the model formulation.

2. This is the usual parametrisation for factor variables. The European group is represented by the model without the additional term and serves as a references for the two other groups

5 Case Study: The 2014 Ebola Epidemic

5.1 Evolution of the Ebola Epidemic in West Africa in 2014

The Ebola epidemic which originated in West Africa in 2014 and which is still ongoing presents an interesting case study for a statistical analysis of the mediatisation of a geopolitical event for several reasons. Not only is the point of origin of the epidemic in time and space known, but one can also localise the moment from which onwards it can be considered a media event. In addition, the geography of its propagation as well as corner stone events in the officical communication and recognition by international organisms, such as the World Health Organisation (WHO), seem to play an important role in its mediatisation. As such, it offers an ideal study example for the GÉOMÉDIA project.

Its global evolution is as follows [1]: A 2-year old boy who gets infected and dies from the virus on December 28th in Meliandou, Guinea, has been recognized in retrospective as the first case of the epidemic. Several other people are infected in this remote village in the South-East of Guinea, from where the epidemic spreads in the months to follow, across Guinea as well as the countries on its borders, Liberia and Sierra Leone, before first cases of airborne transmission are registered. As the epidemic spreads, a non-negligible part of its victims comes from medical staff, including some from First World countries. The identification of the virus as being Ebola takes place only in March 2014 after samples have been sent and analysed by the Pasteur Institute in Paris. This leads to the official recognition of an outbreak of Ebola by the WHO on March 23rd on its website. From this date onwards, the WHO publishes monthly statistics on registered cases and number of deaths. This date can also be seen as the starting point of the mediatisation of the epidemic. However, overall media coverage of the spread of the epidemic in the following months (from March to June) is moderate, in contrast to the steadily growing death toll of the virus (see Fig. 5.1). A visible change in the media representation takes place in July and August 2014 with a steeply growing number of publications. This trend continues through September and October where it peaks. This temporal lag in media attention could be related to the geography of the event and coincides with the first case of airborne transmission of the virus as well as the arrival of cases in the Western Hemisphere. It also falls together with the death toll passing over 1,000. The first WHO situation report is issued on August 29th.

While several outbreaks of Ebola have happend since its discovery in 1976, none had reached epidemic proportions before. While known to occur in Central Africa, this is the first time the virus appears in West Africa, one of the reasons why it was recognised with a substantial delay. The epidemic is now regarded as the most widespread in history and is still not defeated, though the number of new cases has been in decline since the beginning

1. For a complete timeline see for example [European Centre for Disease Prevention and Control, 2015]

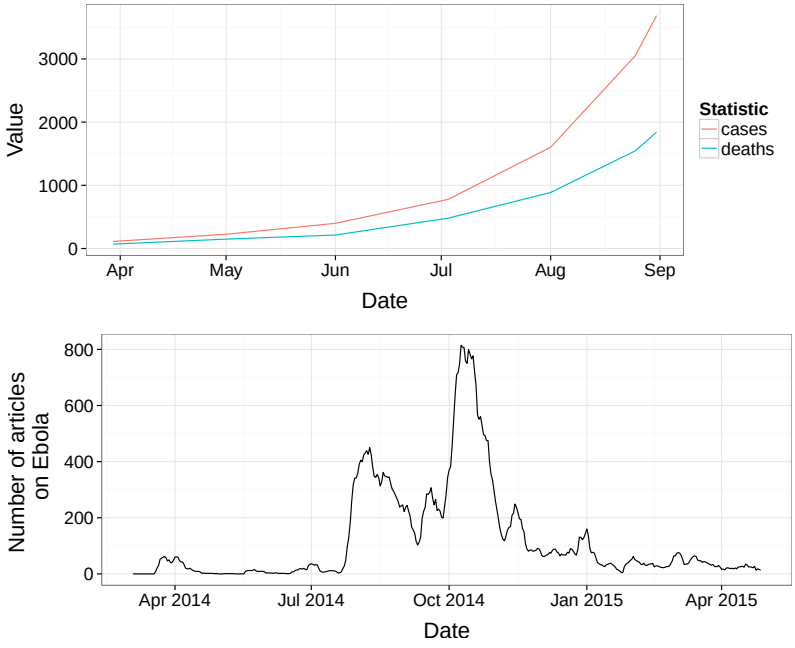

Figure 5.1 – (a) Number of Ebola cases and deaths across Guinea, Sierra Leone and Liberia between the 30th of March and the 31st of Aug 2014 (compiled from WHO statistics) and (b) rolling mean across 7 days of the proportion of Ebola related articles from all articles pooled for 39 international RSS news feeds from March 2014 until April 2015.

of 2015 and several countries have officially been declared Ebola-free. This Ebola epidemic has also been marked as the first time when the virus spread to reach major African cities as well as Western countries, although the number of victims in the latter rests very low in comparison to the number of deaths within the African population. As such, the Ebola epidemic presents an interesting case study in terms of its media representation as well as the analysis of international relations. In particular, it reflects the asymetrie in power between the countries affected by the epidemic and Western countries, where the latter are also implicated in providing important medical help and supplies. It has also implied the simultaneous and sometimes conflicting intervention of several international and non-governmental organisations.

5.2 The Data

For the study presented here, we choose to extract data collected between the 1st of March 2014 up until the 30th of April 2015 in the GÉOMÉDIA database. This covers the period from the Ebola outbreak up to a point well past the peak of media attention, while not including the very first months of the collection which can be considered as a burn-in period for the GÉOMÉDIA application with some technical instabilities and feeds still being added. We decided to retain only the international feeds, but did not exclude any language *a priori*. (Note however that languages are not represented equally by the GÉOMÉDIA corpus, see Section 2.2). The articles were tagged following the standard procedure included in the data processing pipeline and described in Section 2.2 above, based on the dictionary given in Appendix A.1. Amongst the 95 feeds, there are 30 feeds for which the tagging procedure did not return any items. We excluded these feeds subsequently, as such lack of data prevents statistcal modelling. The excluded feeds come in majority from South American journals (22 out of 30) and are written in Spanish or Portuguese. This could indicate that the chosen dictionary does not work well for these two languages and needs to be revised. In order to include geographical information in the analysis, feeds were grouped by continent. Since feed locations in Asia are very dispersed and thus unlikely to provide a geographically meaningful and homogenous covariate, we decided to not take them into account, but keep only European, African, and American feeds. The group of American feeds technically comprised all Americas, but contained mostly North American feeds (8 North American, 2 South American). This leave 39 feeds in total. The number of feeds in each group are 19 (Europe), 10 (Africa) and 10 (Americas), respectively. For the complete list of feeds which are included in the analysis, see Appendix A.2.

5.3 Modelling the Media Attention for the Ebola Epidemic

5.3.1 Model I: Including a Feed-specific and Time-varying Covariate

We model the daily number of publications on Ebola as a proportion of the total number of items as a multivariate Binomial HMM, as discussed in Section 4.4.2 above.

A first model in the most basic form, with one state sequence and the same success probabilities π_i for state i for all 65 feeds, could not be fitted successfully due to problems with the numerical optimisation routine: No suitable initial values could be found. This is not astonishing as such a model is unlikely to describe the data properly, assuming basically no heterogeneity between feeds. Consequently, we included feed-specific covariate information in the observation model to account for inter-feed differences. A readily available covariate, which not only varies across feeds, but also in time, is the total number of daily publications (Model I). This choice is backed up by a preliminary analysis from a generalised linear model (not taking into account the temporal dependence in the observations) which retained the total number of items as a significant explanatory variable. A HMM with this covariate was fitted for 1 to 6 states. The model with 1 state technically corresponds to independent realisations (no temporal correlation) of observations from a product Binomial with paramter π^k. It was included for completeness. Both AIC and BIC favour the model with 4 states (see Table 5.1). In the following, we present the results for this model. The point estimate of the transition matrix is

$$\hat{\mathbf{\Gamma}}_{(\mathrm{I})} = \begin{pmatrix} 0.899 & 0.097 & 0.004 & 0.000 \\ 0.192 & 0.756 & 0.052 & 0.000 \\ 0.000 & 0.091 & 0.872 & 0.037 \\ 0.000 & 0.000 & 0.156 & 0.844 \end{pmatrix}. \tag{5.1}$$

For each day within the observation period, the probability of being in each state can be calculated from the fitted model. This is shown in Fig. 5.2. For computational reasons, the total number of daily publications of each feed is standardised by the maximum number of publications observed over the entire data. Thus, each feed has a time-varying covariate "size" which is bound in [0, 1]. How the differences in size affect the probability for a publication on Ebola for the different states is shown in Fig. 5.3 and the estimates for the coefficients are given in Table 5.2. To quantify the precision of the estimates for the coefficients in the observation model, we derive Hessian-based, approximate Normal standard errors. For this, the inverse of the Hessian evaluated at the ML-estimates is calculated. However, this method encounters problems if parameter estimates lie on the boundary of the parameter space. In our case, this concerns those entries in $\hat{\mathbf{\Gamma}}$ which are essentially zero, but does not effect the parameters of the observation model. Hence, we report confidence intervals only for the estimates of the coefficients in the observation model in Table 5.2. As an alternative, we tried to apply a parametric bootstrap procedure based on resamples from the fitted HMM. However, this runs into difficulties with the initial values for the numerical optimisation as well as the estimation of the parameters, as some states might not be visited by the boostrap resample and label switching may occur. By looking at the confidence intervals, we see that the slope term for "size" is only significantly different from 0 for states 3 and 4, while the confidence intervals cover 0 for the other two states. Hence, size influences the publication behaviour of a feed only in states 3 and 4, where it negatively correlates in the first case and strongly positively correlates in the latter case. In particular, this means that during periods of highest media attention, feeds with many publications tend to also publish more on Ebola, while in phases with

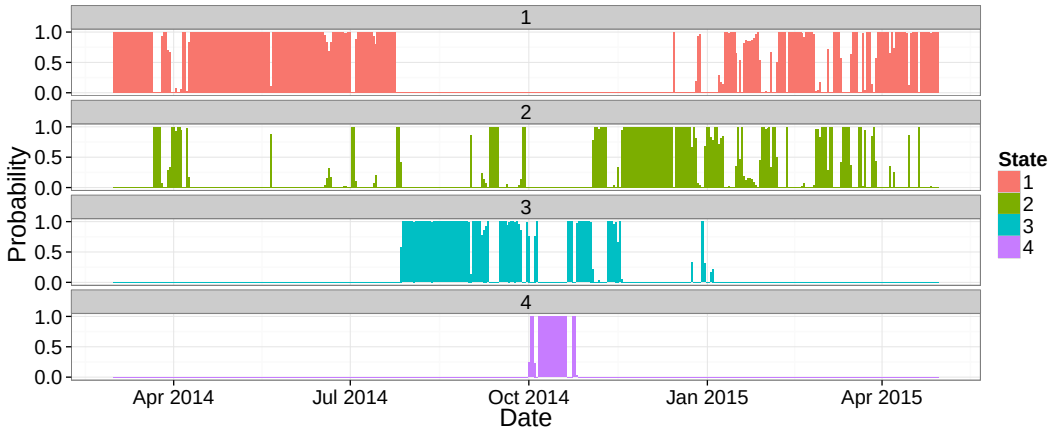

Figure 5.2 – Probability of the state sequence to be in each of the four states for each day in the observation period for Model I.

HMM	Model I			Model II		
	AIC	BIC	Size	AIC	BIC	Size
1 state	30561	30576	2	30383	30413	4
2 states	21680	21726	6	21390	21466	10
3 states	20242	20322	12	19920	20056	18
4 states	**19493**	**19644**	20	**19086**	**19298**	28
5 states	19513	19740	32	19120	19461	42
6 states	19537	19855	44	19996	20419	56

Table 5.1 – Model selection for the two considered multivariate HMMs. Model I contains a feed-specific covariate "size" only and Model II includes in addition a geographical grouping factor. The column Size of the model gives the number of parameters to be estimated. The 1-state model technically corresponds to independent realisations from a Binomial product model with a logistic model for the parameter π. An additional likelihood ratio test between the two 4-states models confirms the significantly better fit of Model II.

second highest level of media attention it is the inverse. By extending the current model in the next section, we will see that the strong effect by the "size" of a feed is implicitly related to geography. State 4 characterises the peak of media attention which is mostly generated by the European and North American press.

5.3.2 Model II: Adding Geography to the Observation Process

To analyse potential differences of attention for the Ebola epidemic depending on the spatial location of the media, we include a spatially informative variable as an explanatory factor in the state-dependent model. There are two principal possible ways to define informative covariates: (1) By including some factor, *i.e.* a discrete variable, which groups feeds according to geographical information or (2) by distance to the event, *i.e.* a continuous

state	parameter	value	95%-confidence
1	$\alpha_{0,1}$	-6.25	(-6.44, -6.08)
	$\alpha_{1,1}$	-0.07	(-1.04, 0.89)
2	$\alpha_{0,2}$	-4.13	(-4.25, -4.02)
	$\alpha_{1,2}$	-0.51	(-1.14 , 1.18)
3	$\alpha_{0,1}$	-2.83	(-2.90, -2.76)
	$\alpha_{1,3}$	-0.92	(-1.50, -0.33)
4	$\alpha_{0,4}$	-2.27	(-2.37,-2.17)
	$\alpha_{1,4}$	3.03	(2.20, 3.85)

Table 5.2 – Parameter estimates for the coefficients in the observation process model of the Binomial HMM with size of feed as feed-specific and time-varying covariate (Model I)

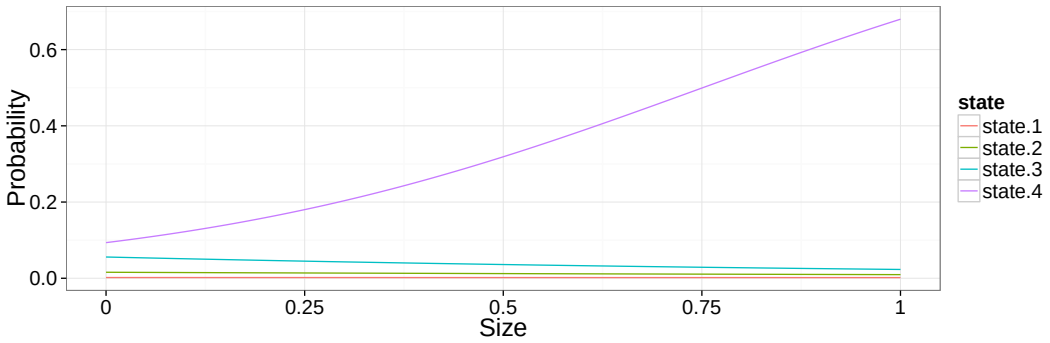

Figure 5.3 – Probability that an RSS item about Ebola is published as a function of the total number of articles published on the respective day, for the 4 states of the model. Size refers to the total number of articles standardised by the maximum number of daily publications across all feeds during the observation period.

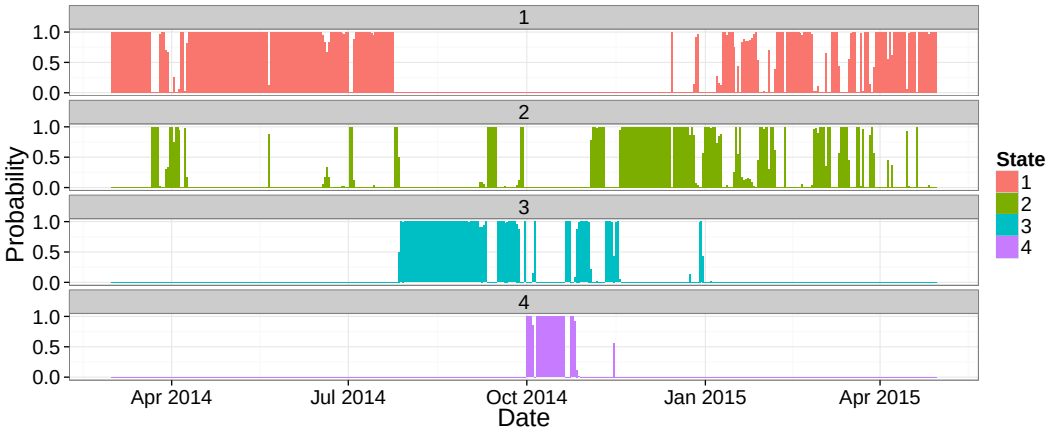

Figure 5.4 – Probability of the state sequence to be in each of the four states for each day in the observation period for Model II.

variable. In the second case, meaningful delimiters for such a distance have to be chosen. Amongst others, this demands a decision about the geographical center of the event and locating a feed in space. We follow option (1). Since we have to retain enough data points for a reliable model fit in each factor group, a compromise has to be made between spatial resolution and sufficient data. Here feeds are grouped by continent (see Section 5.2). We hence look at Model II, including daily total publications as feed-specific time-varying covariate and continent as a rough spatial covariate. Again, AIC as well as BIC selection criteria choose the model with 4 states (see Table 5.1). Comparing with the previous values for Model I, we see that in all cases the model including geographical information fits the data better. A likelihood ratio test was conducted for the two models with 4 states, with and without the geographic covariate added. The result shows that the better fit of Model II is indeed highly significant ($p < 0.001$).

The estimated transition probabiliites are close to the previous estimates and the profile plot of the sequence of state probabilities remains basically unchanged (see Fig. 5.4) :

$$\hat{\mathbf{\Gamma}}_{(\mathrm{II})} = \begin{pmatrix} 0.898 & 0.098 & 0.004 & 0.000 \\ 0.194 & 0.769 & 0.037 & 0.000 \\ 0.000 & 0.071 & 0.884 & 0.045 \\ 0.000 & 0.000 & 0.160 & 0.840 \end{pmatrix} \tag{5.2}$$

The grouping factor “continent” allows to alter the intercept in the equation of Model I (see Section 4.4.3) depending on the geographical location of the feed, for the African and American feeds individually. No factor term is added for the European feeds, *i.e.* European feeds are described by the equation of Model I and serve as a reference for the two other groups. The estimates for the coefficients are provided in Table 5.3. The dependence of π on the total daily publications is shown in Fig. 5.5. It follows the same shape as in Model I for both Europen and American feeds, but differs for African feeds. For the latter, the

state	parameter	value	95%-confidence
1	$\alpha_{0,1}$	-6.48	(-6.71, -6.26)
	$\alpha_{1,1}$	0.21	(-0.76 , 1.18)
	$\rho_{1,AU}$	0.61	(0.29 , 0.92)
	$\rho_{1,US}$	0.30	(0.03 , 0.56)
2	$\alpha_{0,2}$	-4.37	(-4.49, -4.25)
	$\alpha_{1,2}$	-0.21	(-0.84 , 0.42)
	$\rho_{2,AU}$	0.37	(0.21 , 0.54)
	$\rho_{2,US}$	0.49	(0.36 , 0.61)
3	$\alpha_{0,1}$	-3.18	(-3.27 , -3.09)
	$\alpha_{1,3}$	-0.06	(-0.64 , 0.52)
	$\rho_{3,AU}$	0.89	(0.80 , 0.99)
	$\rho_{3,US}$	0.38	(0.30 , 0.46)
4	$\alpha_{0,4}$	-2.36	(-2.46 , -2.25)
	$\alpha_{1,4}$	3.07	(2.30, 3.85)
	$\rho_{4,AU}$	-0.13	(-0.30 , 0.03)
	$\rho_{4,US}$	0.16	(0.06 , 0.26)

Table 5.3 – Parameter estimates for the coefficients in the observation process model of the Binomial HMM with geographical covariates (Model II)

probabilities π_3 and π_4 are very close. This is confirmed in a plot of π changing over time according to the sequence of the most probable states (Fig. 5.6). For the African feeds, state 3 and 4 lead to basically the same values for π while there is a significant difference between the two states for European and American feeds. Hence, the state of highest media attention (state 4) is in addition a state that increased publication probabilities drastically *for the Western countries only*. In state 3, on the other hand, African feeds have almost double the probability to publish on Ebola than European and American feeds on average.

5.3.3 Posteriori Model Interpretation and Conclusions

As it is the best fitting model, we now concentrate on Model II and its results. First, we look at the sequence of the most probable states over the observation period. A first change occurs on the 22nd of March 2014 and lasts three days. This corresponds to the official information of the WHO by the Guinean government about an outbreak of Ebola in the country, leading to the first WHO report on March 25th. A second increase in publications follows shortly at the end of March, lasting over the first days of April, and presumably documents the first statistics published by the WHO at the end of March. A further change from the base state into state 2, lasting more than one day, arises at the beginning of July and presumably reflects the press reaction to the announcement of Médécins Sans Frontières about the spread of Ebola being "out of control", coinciding with the updates of the WHO statistics at the end of June. From July 25th onward, the state sequence goes from a brief initial passage of state 2 to state 3. This date corresponds to the first case

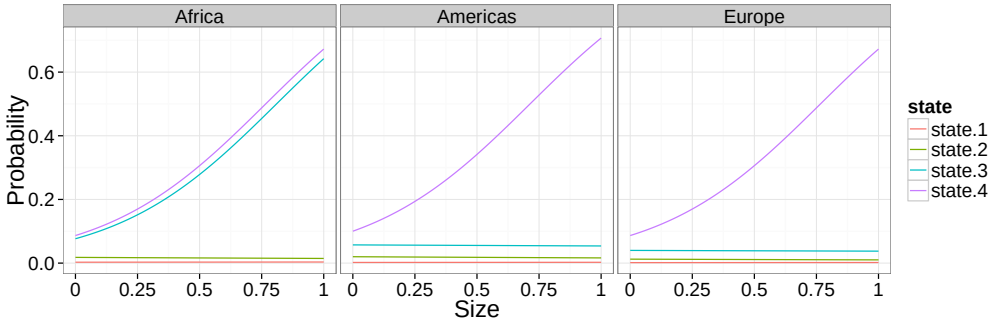

Figure 5.5 – Probability that an RSS item about Ebola is published as a function of the total number of articles published on the respective day, for the 4 states of the model and depending on the geographical location of the feed. Size refers to the total number of articles standardised the maximum number of daily publication across all feeds during the observation period.

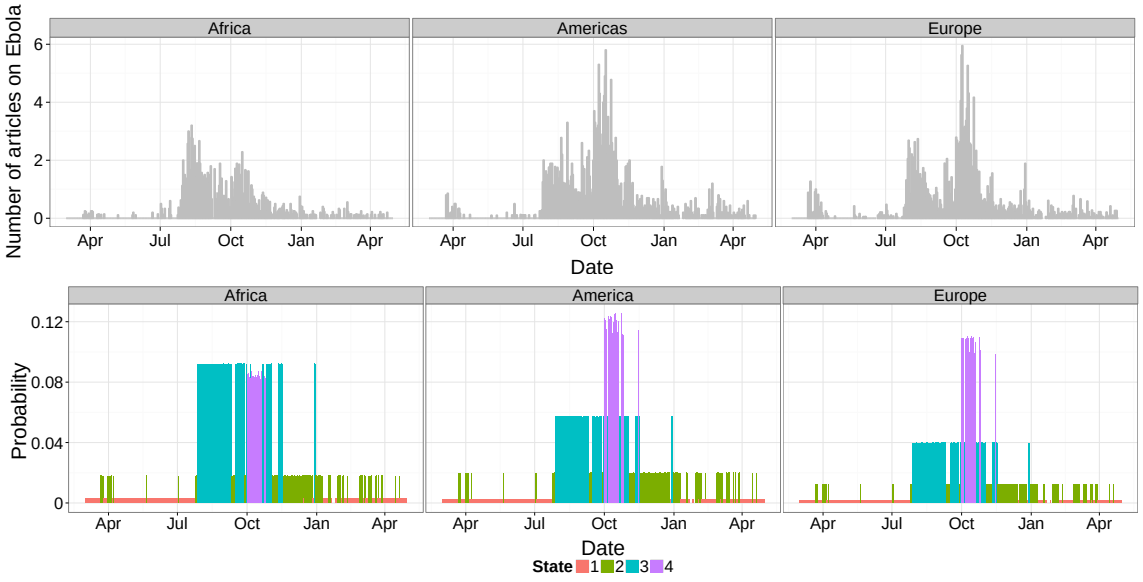

Figure 5.6 – The mean number number of articles published daily on Ebola by African, American and European RSS news feeds, respectively (upper panel) and the probabilities of publishing an article on Ebola as predicted by Model II (lower panel), for each of the three possible geographical locations and based on the common sequence of most probable states.

of airborne transmission to Nigeria and the first case in Sierra Leone's capital, Freetown, both examples of the unprecedented spread of the virus. The model then stays in state 3 with occasional short-term switches to state 2 over all of August and September, the two months when countries in the Western Hemisphere start to realise the dimension of the outbreak, with the number of death climbing over 1,000 and with the first cases of patients being evacuated to the United States and Europe. By October, the state sequence switches to state 4. The epidemic has fully reached the Western Hemisphere. This characterises the peak of media attention and lasts until the beginning of November. A short signal is observed at the end of December, presumably reflecting a case of Ebola being diagnosed in the UK (note that our data set is slightly unbalanced with more European newspapers).

Overall the state sequence seems to catch up on corner stone events in the evolution of the epidemic, and in particular: the first occasions of unprecedented spread (to African capitals and airborne transmission), the updates on WHO statistics and reports, and the cases in the Western Hemisphere. A highly interesting result is the fact that state 4, which corresponds to the period of the full implication of the Western Hemisphere, corresponds to significantly more publications in the European and the American press while there is basically no change in publication rates for the African newspapers. Apart from the period in state 4, the probability of a publication on Ebola is visibly higher for the African press. This reflects the spatial proximity and the fact that African countries are directly concerned by the spread of the virus. Overall, the publication pattern of the American and the European Press is highly synchronized (Fig. 5.6), which could stem from the fact that newspapers on both continents are alimented largely by the same press agencies but also reflect the political proximity between the US and Europe, in contrast to the African countries. Evidently, the models and the analysis for this case study could be refined. For the analyses demonstrated in this report, feeds are grouped geographically in *ad hoc* manner. While there is some justification for these three groups such as the fact that they are linked to three press agencies, the resulting geographical division can be perceived as largely artifical. An alternative grouping considers countries as "central", "semi-peripheric" and "peripheric", which reflects the media theories on the dominance North-South. However these classes are not unambiguous and there are several possible definitions of these groups [Grasland and Van Hamme, 2010]. As another possibility, we envisage to apply the aggregation techniques presented in deliverable 3.1. to identify alternative, spatially meaningful aggregates from the data. Likewise, for this analysis we chose to retain only feeds in the "international" category. In the future, we plan on including other categories. By including the types of feed as a grouping factor similar as we have done for the geographical locations, the modelling framework would allow us an analysis oriented towards a comparison between media rather than with regards to geography and to infer about differences in allocating media attention depending on the type of the feed.

6 Discussion

6.1 Summary

In this report, we presented a statistical modelling framework for the temporal evolution of a media event and its geographical implications. To this end, we analysed multiple series of news items published by international RSS feeds and provided by a collection of such data as part of the GÉOMÉDIA research project. The number of RSS publications on a specific event is taken as an indicator of its importance as a media event. An adequate model respects the particular data properties and allows us to infer about changes in the media attention for a geopolitical event. In particular, the derived model representation of the media system can be used for conclusions about the event's geopolitical implications.

We proposed to use hidden Markov models as a methodological framework to address this problem. As they are models for time series of discrete data, these models are not only appropriate for count data but also take into account the temporal properties of the information flow contained in the RSS news items. In addition, their structure appears to suit our conceptualisation of media: a latent state process – corresponding to the "real" event – drives the behaviour of the observation process – that is the "media" event, *i.e.* the number of event-specific publications. This corresponds to the fact that the degree of newsworthiness assigned to an event by the media is not directly measurable. The behaviour of the observation process is determined by a state dependent probability model. We presented the two canonical choices when dealing with count data, namely a Poisson or a Binomial distribution, and argued for choosing the latter as it takes into account changes in the overall behaviour of a feed's publication rate. Furthermore, we showed how covariate information can be included in the state dependent probabilities. This allowed us to model some heterogeneous behaviour between the RSS feeds. Via fitting an HMM, we can infer the most probable sequence of states giving rise to the data, and interpret it as the level of media attention towards the event of interest.

In a case study on Ebola, we demonstrated that changes in the identified sequence of states actually correspond to corner stone events in the spread and the global communication on the disease. The model of observation processes allows us to compare feeds, in particular by including geographical information. We illustrated differences in media attention between the African, the American and the European press. A higher probability for Ebola-related publication could be observed for African feeds at the beginning of the epidemic, and increasingly so over the month to follow, while the probability of publication was overall lower in the American and the European feeds over this time. This directly reflects the fact that the epidemic orginated in Africa and hence African countries are directly concerned from the beginning. The American feeds tend to report more on Ebola than the European feeds. This could be because the United States were the first to evacuate and treat cases evacuated outside of Africa. It could also reflect the geographic location of the

United Nations (UN) within the US and hence a higher attention to the announcements of UN organisations, such as the WHO. Third, this tendency could be rooted in a difference in mentality between the US where concerns about national security play an important role, and the European countries where this is less explicit. Finally, this can be seen as another example of the tendency of the Western media to report on events taking place outside their hemisphere, in particular in Third World Countries [Mansell and Nordenstreng, 2006]. The peak of media attention falls with the arrival of the epidemic in the Western Hemisphere and this confirms the long observed geographical "bias" in the media.

6.2 Perspectives

Clearly, the proposed models have a simple structure which could be refined. The slope coefficients are the same for all feeds. We could introduce further variability by extending the geographical grouping factors to the slope terms.

Even more general would be the introduction of random effects which means that feed-specific parameters are drawn from a common distribution to include additional variation between feeds. Instead of increasing the number of estimated parameters significantly by including additional feed-specific factors, only the characteristic parameters of this common distribution have to be additionally estimated. Random effects can thus capture otherwise unexplained heterogeneity. However they are usually much harder to interpret in comparison to explicit covariate information. The principal drawback of random effects is the increase in computational burden as each realisation of the corresponding random effects adds the evaluation of an integral to the calculation of the likelihood. In general, numerical derivation of the maximum likelihood estimates can only be maintained for a small number of random effects [Altman, 2007]. However, in the case of the multivariate product model (Eq. 4.7), the inclusion of random effects is not feasible since each factor in the product P_π in Eq. 4.6 would call for a realisation of the random effect from the distribution specified by the state. We have hence not considered random effects here, but state that in general, with a different model specification, random effects can well be included at some non-negligible additional computational cost. Overall they should only be considered if no other covariate information is available, but important structural heterogeneity in the data is assumed. In some cases, the computational burden can be lessened by considering discrete random effects [Maruotti and Rydén, 2009]. This approach replaces the continuous distribution for a random effect by a discrete one, resulting in a summation instead of an integral in the evaluation of the likelihood. In addition, discrete random effects are often more directly interpretable.

So far, we have considered extensions in the state dependent process to model heterogeneity within the ensemble of news feeds. Further model refinements can be obtain by enriching the specifications of the model for the state sequence. This could be achieved by including covariates that concern all feeds and leads to temporal variation in the transition probabilities. Such variables could for example be a general time trend or, more explicitly, the number of new cases or the number of deaths as reported by the WHO statistics.

With regards to the hidden state process, we also did not take into account competition between media events, but considered the media attention for one event as independent from others. This assumption clearly is a simplification and further developments to the proposed HMM framework should try to at least model two competing events.

Another limiting assumption concerning the state process is the finite number of states, which in addition is usually taken to be fairly small. Adding states rapidly increases the number of parameters in the transition matrix $\mathbf{\Gamma}$, whose estimation is rarely supported by the sample size. The assumption of a finite, small number of discrete states in the state process is a simplification of reality and may sometimes not be justified. In this case, state-space models with a continuous state process overcome this limitation [Durbin and Koopman, 2001]. It still is possible to consider an approximate version of such a model in the standard HMM framework [Langrock, 2011].

There are at least two main alternatives to Hidden Markov Models: Staying in the context of discrete time series models, other modelling tools which generalise the classical time series models, such as INAR and INGARCH, have been proposed [Jung and Tremayne, 2011]. However, the application and the fitting of these models and hence their interpretability seems highly context-specific. We choose the HMM framework here for its generality, its flexibility and the ease of its fitting routines. A different approach would be to remain in the continuous time frame. In this case, Markov Modulated Poisson Processes are the time continuous equivalent of Poisson HMMs. While mathematically attractive, they are harder to fit and limited to a Poisson Model. We argued above why we think a Binomial model is more appropriate. The simplification of a discrete time scale could also be adjusted by considering a continuous state process as discussed above. This has the advantage that models can still be fitted approximately in the HMM setting, thus drawing from the strength of their flexible and fairly simple framework rather than considering mathematically more complex solutions.

Acknowledgements

We would like to thank Roland Langrock at the University of St Andrews for valuable discussions about the correct model specification and help with the implementation in R.

List of Symbols

G	an RSS news feed
k	index of RSS news feeds
$\mathcal{G} = \{G^1, \ldots, G^K\}$	an ensemble of K RSS news feeds
ω_t^k	an RSS item, *i.e.* an article, published by RSS feed G_k, at time t
t_0, t_1, t_ℓ	points in time with $t_0 \leq t_1$
T	an interval in time
Δ	time step chosen for discretising a continuous time scale
$\mathcal{D}$	a dictionary, *i.e.* a set of keywords
tag	an indicator function which formally describes tagging of keywords
N_ℓ^k	a random variable giving the total number of publications of feed G_k at the ℓ-th unit of time
X_ℓ^k	a random variable giving the number of event related publications of feed G_k at the ℓ-th unit of time
$\mathbf{N}^k$	the vector $(N_\ell^k)_{\ell=1,2,\ldots,L}$
$\mathbf{X}^k$	the vector $(X_\ell^k)_{\ell=1,2,\ldots,L}$
S	random variable, state of a Markov chain
$\mathbb{P}$	a probability distribution of a discrete random variable
$\mathbb{E}, \mathbb{V}$	the expecation and variance of a random variable
λ	characteristic parameter of the Poisson distribution
N	positive integer, first characteristic parameter of the Binomial distribution
π	probability, second characterstic parameter of the Binomial distribution
γ_{ij}	transition probability of a Markov chain in state i to state j
$\mathbf{\Gamma}$	matrix of transition probabilities of a Markov chain
$\boldsymbol{\delta}$	probability distribution of the initial state of a Markov chain

Bibliography

[Allan, 2006] Allan, S. (2006). *Online News: Journalism And The Internet: Journalism and the Internet*. Open University Press. McGraw-Hill.

[Altman, 2007] Altman, R. M. (2007). Mixed hidden Markov models: An extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, 102(477):201–210.

[Durbin and Koopman, 2001] Durbin, J. and Koopman, S. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford.

[Eddy, 1996] Eddy, S. R. (1996). Hidden Markov models. *Current Opinion in Structural Biology*, 6(3):361–365.

[European Centre for Disease Prevention and Control, 2015] European Centre for Disease Prevention and Control (updated May 2015). Ebola outbreak in West Africa: Event background. http://ecdc.europa.eu/en/healthtopics/ebola_marburg_fevers/event-background/Pages/default.aspx.

[Géomédia, 2015] Géomédia (2015). Corpus géomédia: Carnet de recherche de l'ANR Corpus Géomédia. https://geomedia.hypotheses.org, OpenEdition – Hypotheses, Blogs académiques.

[Géomédia Doc Scientifique, 2011] Géomédia Doc Scientifique (2011). Observatoire des flux géomédiatique internationaux. Project proposal, ANR: Corpus, donneées et outils de la recherche en sciences humaines et sociales.

[Grasland et al., 2011] Grasland, C., Giraud, T., and Severo, M. (2011). Un capteur géomédiatique d'evénements internationaux. In *Fonder les sciences du territoire*, pages 184–190. Colloque international.

[Grasland and Van Hamme, 2010] Grasland, C. and Van Hamme, G. (2010). La relocalisation des activités industrielles : une approche centre-périphérie des dynamiques mondiales et européennes. *Espace géographique*, pages 1–19.

[Hastie and Tibshirani, 1990] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall/CRC.

[Jung and Tremayne, 2011] Jung, R. C. and Tremayne, A. (2011). Useful models for time series of counts or simply wrong ones? *AStA Advances in Statistical Analysis*, 95:59–91.

[Langrock, 2011] Langrock, R. (2011). Some applications of non-linear and non-Gaussian state-space modelling by means of hidden Markov models. *Journal of Applied Statistics*, 38(12):2955–2970.

[Mansell and Nordenstreng, 2006] Mansell, R. and Nordenstreng, K. (2006). Great media and communication debates: WSIS and the MacBride report. *Information Technologies and International Development*, 3(4):15–36.

[Maruotti and Rydén, 2009] Maruotti, A. and Rydén, T. (2009). A semi-parametric approach to hidden Markov models under longitudinal observations. *Statistics and Computing*, 19(4):381–393.

[McCullagh and Nelder, 1989] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton.

[Mitchelstein and Boczkowski, 2009] Mitchelstein, E. and Boczkowski, P. J. (2009). Between tradition and change. A review of recent research on online news production. *Journalism*, Vol. 10(5):562–586.

[R Core Team, 2014] R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

[Rabiner, 1989] Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *IEEE Proceedings*, 77:257–286.

[Rossi and Gallo, 2006] Rossi, A. and Gallo, G. M. (2006). Volatility estimation via hidden Markov models. *Journal of Emipirical Finance*, 13(2):203–230.

[Schliehe-Diecks et al., 2012] Schliehe-Diecks, S., Kappeler, P., and Langrock, R. (2012). On the application of mixed hidden Markov models to multiple behavioural time series. *Interface Focus*, 2:180–189.

[Zucchini and MacDonald, 2009] Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series. An Introduction Using R*. Chapman & Hall/CRC, Boca Raton.

A Supplementary Material

A.1 Dictionary of Tags for Ebola

keyword	type	language
ébola	virus name	Spanish
ebola	virus name	French
ebola	virus name	English
ebola	virus name	Polish
ebola	virus name	Italian
ébola	virus name	Portuguese
ebola	virus name	German
ebolafieber	virus name	German
ebolavirus	virus name	German
ebolavirus	virus name	English

A.2 List of Feeds Used in the Ebola Case Study

	language	country	name	feed type
European Newspapers				
1	English	United Kingdom	Daily Telegraph	international
2	English	United Kingdom	Financial Times	international
3	English	United Kingdom	The Guardian	international
4	English	United Kingdom	The Times	international
5	English	Malta	The Times of Malta	international
6	French	France	Dernière Heure	international
7	French	Belgium	Le Soir	international
8	French	France	France Antilles	international
9	French	France	Dernières Nouvelles d'Alsace	international
10	French	France	Le Figaro	international
11	French	France	Le Parisien	international
12	French	France	Libération	international
13	French	France	Le Monde	international
14	German	Germany	General Anzeiger	international
15	German	Germany	Die Welt	international
16	German	Germany	Frankfurter Allgemeine Zeitung	international
17	Italian	Italy	Corriere della Sera	international
18	Italian	Italy	repubblica	international
19	Polish	Poland	Rzeczpospolita	international
African Newspapers				
20	English	Kenia	The Daily Nation	international
21	English	Nigeria	This Day	international
22	English	Uganda	Daily Monitor	international
23	English	Uganda	New Vision	international
24	English	Uganda	Red Pepper	international
25	English	Zambia	IOL	international
26	English	Zimbabwe	Chronicle	international
27	English	Zimbabwe	The Herald	international
28	French	Algeria	El Watan	international
29	French	Algeria	l'Expression	international
American Newspapers				
30	English	USA	The Los Angeles Times	international
31	English	USA	The New York Times	international
32	English	USA	USA Today	international
33	English	USA	The Wallstreet Journal	international
34	English	USA	Washington Post	international
35	English	Canada	The Star	international
36	English	Canada	The Vancouver Sun	international
37	French	Canada	Le journal de Montréal	international
38	English	Antigua-Barbuda	Daily Observer	international
39	English	Brasil	Folha de S. Paulo	international

B R Code

The following code implements the likelihood of the multivariate Binomial HMM with one common state sequence and an obsevation process model containing scaled total daily publications and continent as a geographical grouping factor (Model II). Model I can be obtained easily, by deletion of the coefficient vector ρ.

```
##### parameter transformation to obtain unconstrained working parameters ######
pn2pw <- function(m, alpha,  rho,  gamma){
   talpha <- as.vector(t(alpha))
   trho <- as.vector(t(rho))
   tgamma <- NULL
  if(m>1){
     foo <- log(gamma/diag(gamma))
     tgamma <- as.vector(foo[!diag(m)])
  }
   parvect <- c(talpha,  trho, tgamma)
  parvect
}


##### back transformation to natural (constrained) parameters ######
pw2pn <- function(m, parvect){
  epar <- c(parvect[1:(4*m)], exp(parvect[-(1:(4*m))]))
   alpha <- matrix(epar[1:(2*m)],m,2, byrow=T)
rho <- matrix(epar[(2*m+1):(4*m)], m, 2, byrow=T)
gamma <- diag(m)
   if(m>1){
     gamma[!gamma] <- epar[-(1:(4*m))]
     gamma <- gamma/apply(gamma,1,sum)
   }
   delta <- solve(t(diag(m)-gamma+1),rep(1,m))
   list( alpha=alpha, rho=rho, gamma=gamma, delta=delta)
}


### negative log-likelihood of  the stationnary multivariate Binomial HMM ###
mllk <- function(parvect, x, size, cont_fact, m, ...){
pn <- pw2pn(m, parvect)

   n <- dim(x)[1]
  l <- dim(x)[2]
   p <- array(NA, dim=c(n,l,m))
   probs <- array(NA, dim=c(n,l,m))
   mx.size <- max(size, na.rm=TRUE)
```

```
size.scaled <- as.matrix(size/mx.size)

  for(j in 1:m){
    tt <- pn$alpha[j,1]+pn$alpha[j,2]*size.scaled+
        pn$rho[j, 1]*cont_fact[,,2]+pn$rho[j,2]*cont_fact[,,3]
    p[ , ,j] <- exp(tt)/(exp(tt)+1)
    probs[ , ,j] <- dbinom(as.matrix(x), as.matrix(size), p[ , ,j])
   }

  probs <- ifelse(!is.na(probs), probs, 1)
   allprobs<-apply(probs, c(1,3),prod)
   lscale <- 0
  foo <- pn$delta

   for(i in 1:n){
     foo <- foo%*%pn$gamma*allprobs[i,]
     sumfoo <- sum(foo)
     lscale <- lscale+log(sumfoo)
     foo <- foo/sumfoo
   }
   mllk <- -lscale
  mllk
}

########## Optimisation:  ML estimation the parameters ###############

mle <- function(x, size, cont_fact, m, alpha0,  rho0, gamma0){
  parvect0 <- pn2pw(m, alpha0,rho0,  gamma0)
  mod <- nlm(mllk, parvect0, x=x, size=size,cont_fact=cont_fact,
   m=m, print.level=2,iterlim=400, hessian=T)
  pn <- pw2pn(m,mod$estimate)
  mllk <-mod$minimum
  np <- length(parvect0)
  AIC <- 2*(mllk+np)
  n <- sum(!is.na(x))
  BIC <- 2*mllk+np*log(n)
  list(alpha=pn$alpha, rho=pn$rho, gamma=pn$gamma, delta=pn$delta,
   code=mod$code, mllk=mllk, AIC=AIC, BIC=BIC, hessian=mod$hessian)
}

## where
## x is an n x l matrix containing the number of items tagged positively for Ebola
## for a period of length n and an ensemble of l feeds;
## size is the n x l matrix of total number of items published on each day in the
## observation period by each feed in the study ensemble;
## cont_fact is an n x l x 2 array of 0 and 1 column vectors that indicate the
## membership of each feed (column) to the geographical grouping factors;
## alpha0, rho0, gamma0 are initial parameter values for the numerical optimisation
```

```
## The function mle returns the estimated parameter values at the maximum, a code
## indicating the successful convergence of the optimisation algorithm,  the value
##  of the loglikelihood at the maximum (mllk), the AIC and BIC and the Hessian
## evaluated at the max.



#### fitting Modell II with 4 states
m=4
alpha0 <- c(-4, 0.2,-2, 0.1,-1, -0.5,-1, -0.02)
rho0 <-  c( 0, 0.5,  -1,1,  0, -1, 0.5, -0.5)
gamma0<-matrix(0.1,m,m)
diag(gamma0)<-rep(1-0.1*(m-1),m)

mod.multi.4 <- mle(x.3, size.3, cont_fact, m, alpha0,rho0, gamma0)


##### Local decoding: Probabilities for each state given the observations ######

##### Recursive calculation of forward-backward probabilities
lalphabeta <- function(x,size, cont_fact, m, alpha,  rho, gamma, delta=NULL){
   f(is.null(delta))delta <- solve(t(diag(m)-gamma+1),rep(1,m))

n <- dim(x)[1]
  l <- dim(x)[2]
   lalpha <- matrix(NA, m,n)
lbeta <- matrix(NA, m,n)
   p <- array(NA, dim=c(n,l,m))
  probs <- array(NA, dim=c(n,l,m))
  mx.size <- max(size, na.rm=TRUE)
size.scaled <- as.matrix(size/mx.size)

   for(j in 1:m){
     tt <- alpha[j,1]+alpha[j,2]*size.scaled  +
     rho[j, 1]*cont_fact[,,2]+ rho[j,2]*cont_fact[,,3]
     p[ , ,j] <- exp(tt)/(exp(tt)+1)
     probs[ ,, j] <- dbinom(as.matrix(x), as.matrix(size), p[ , ,j])
  }

   probs <- ifelse(!is.na(probs), probs, 1)
   allprobs<-apply(probs, c(1,3),prod)
  foo <- delta*allprobs[1,]
   sumfoo <- sum(foo)
  lscale <- log(sumfoo)
   foo <- foo/sumfoo

lalpha[,1] <- log(foo)+lscale
for(i in 2:n){
```

```
    foo <- foo%*%gamma*allprobs[i,]
     sumfoo <- sum(foo)
    lscale <- lscale + log(sumfoo)
     foo <- foo/sumfoo
    lalpha[,i] <- log(foo)+lscale
   }

   lbeta[,n] <- rep(0,m)
  foo <- rep(1/m, m)
  lscale <- log(m)
   for(i in (n-1):1){
    foo <- gamma%*%(allprobs[i+1,]*foo)
     lbeta[,i] <- log(foo)+lscale
    sumfoo <- sum(foo)
    foo <- foo/sumfoo
     lscale <- lscale + log(sumfoo)
   }
   list(la=lalpha, lb=lbeta)
}

state_probs <-  function(x,size,cont_fact, m, alpha, rho, gamma, delta=NULL,...){
if(is.null(delta))delta <- solve(t(diag(m)-gamma+1),rep(1,m))
n <- dim(x)[1]
fb <- lalphabeta(x,size,cont_fact, m, alpha,  rho, gamma, delta)
la <- fb$la
lb <- fb$lb
c <- max(la[,n])
llk <- c+log(sum(exp(la[,n]-c)))
stateprobs <- matrix(NA, ncol=n, nrow=m)
for(i in 1:n) stateprobs[,i]<- exp(la[,i]+lb[,i]-llk)
stateprobs
}

state.probs <- state_probs(x.3, size.3, cont_fact, m=4, mod.multi.4$alpha,
mod.multi.4$rho, mod.multi.4$gamma)
```