SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Statistical models for count time series with excess zero

By

Anastasia Pitsari

A THESIS

Submitted to the Department of Statistics

of the Athens University of Economics and Business

in partial fulfilment of the requirements for

the degree of Master of Science in Statistics

Athens, Greece
August 2020


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

Στατιστικά μοντέλα για χρονοσειρές με περίσσεια μηδενικών

Αναστασία Πίτσαρη

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής

του Οικονομικού Πανεπιστημίου Αθηνών

ως μέρος των απαιτήσεων για την απόκτηση

Διπλώματος Μεταπτυχιακών Σπουδών στη Στατιστική

Αθήνα

Αύγουστος 2020


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

DEDICATION

to my mother Athina…


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ACKNOWLEDGEMENTS

At this point, I would like to thank you my supervisor Mr. Panagiotis Besbeas for his willingness to discuss with me about my premature ideas and his continuing guidance to complete this thesis. His advice and encouragement played an important role in achieving my goal. I would also like to thank all the faculty members of the Department of Statistics for their time and advices. Finally, I would like to thank my family and my friends, who through their love, support, patience and prompting, made it possible for me to fulfill most of my goals and dreams.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ABSTRACT

Anastasia Pitsari

Statistical models for count time series with excess zero

August 2020

This thesis deals with count time series with excess zero. The first category of models that would fit in this case is the zero inflated models to which the observation-driven and parameter-driven models belongs. We illustrate this approach but also the next ones using a real data set on injuries of hospital cleaners in a hospital. We start by applying a zero inflated Poisson model with autoregressive process. After that, we analyze another category of models known as Hidden Markov models (HMMs). A hidden Markov model is a statistical model in which the system being modelled is assumed to be a Markov process with unobservable states (hidden). The state process of an HMM is usually discrete and this is the first case we apply to injury data. However, sometimes it is useful in assuming a continuous state and these models are known as state space models (SSMs). SSMs have more difficult statistical techniques and demanding algorithms and in addition require extra calculations because there is no direct computation formula for the likelihood. Therefore, after the approach to zero inflated Poisson we apply a Poisson HMM and then we consider a continuous SSM model in which presented with two different implementations. Finally, we calculate the information criteria (AIC and BIC) and compare the models in order to arrive at the most appropriate for the injury data but also for the simulation data.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΠΕΡΙΛΗΨΗ

Αναστασία Πίτσαρη

Στατιστικά μοντέλα για χρονοσειρές με περίσσεια μηδενικών

Αύγουστος 2020

Στην παρούσα διπλωματική επεξεργαζόμαστε μοντέλα για χρονοσειρές που περιέχουν πολλά μηδενικά. Η πρώτη κατηγορία μοντέλων που αναλύουμε και ταιριάζει στα δεδομένα μας είναι τα μηδενικά-φουσκωμένα μοντέλα στα οποία ανήκουν τα μοντέλα παρατήρησης αλλά και τα μοντέλα παραμέτρων. Για να αναδείξουμε αυτά τα μοντέλα αλλά και τα επόμενα που αναλύουμε θα χρησιμοποιήσουμε διακριτά δεδομένα τραυματισμών που λαμβάνουν χώρα σε ένα νοσοκομείο και γίνονται από τους υπαλλήλους καθαρισμού. Ξεκινάμε την ανάλυση μας εφαρμόζοντας ένα Poisson μηδενικό-φουσκωμένο μοντέλο μαζί με αυτοπαλίνδρομο μοντέλο τάξεως ένα. Στην συνέχεια, προχωράμε στην ανάλυση μας με ακόμη μία κατηγορία μοντέλων τα οποία λέγονται κρυπτομαρκοβιανά μοντέλα. Τα κρυπτομαρκοβιανά μοντέλα είναι στατιστικά μοντέλα στα οποία η διαδικασία που θέλουμε να μοντελοποιήσουμε είναι μία Μαρκοβιανή αλυσίδα με μη παρατηρήσιμες καταστάσεις (κρυμμένες). Οι καταστάσεις των κρυπτομαρκοβιανών μοντέλων είναι συνήθως διακριτές τυχαίες μεταβλητές και αυτή θα είναι και η πρώτη περίπτωση που θα εφαρμόσουμε στα δεδομένα τραυματισμών. Ωστόσο, μερικές φορές είναι χρήσιμο να υποθέσουμε ότι οι καταστάσεις αυτές είναι συνεχείς και αυτά τα μοντέλα είναι γνωστά ως μοντέλα χώρου-καταστάσεων και έχουν ακριβώς την ίδια δομή με τα κρυπτομαρκοβιανά. Τα μοντέλα χώρου-καταστάσεων παρουσιάζουν πιο δύσκολες στατιστικές τεχνικές και πιο απαιτητικούς αλγορίθμους καθώς και επιπλέον υπολογισμούς διότι δεν υπάρχει άμεσος τρόπος υπολογισμού της πιθανοφάνειας. Επομένως, μετά την εφαρμογή του μηδενικά φουσκωμένου Poisson μοντέλου, εφαρμόζουμε ένα Poisson κρυπτομαρκοβιανό μοντέλο και στην συνέχεια ένα μοντέλο χώρου-καταστάσεων με δύο διαφορετικές

υλοποιήσεις. Τέλος, υπολογίζουμε τα κριτήρια πληροφορίας (AIC και BIC) και συγκρίνουμε τα μοντέλα προκειμένου να φτάσουμε στο καταλληλότερο που να περιγράφει τα δεδομένα τραυματισμών αλλά και τα δεδομένα του πειράματος της προσομοίωσης.

Contents

List of Figures


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

List of Tables


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 1

Introduction

Time series data involving counts are frequently encountered in many applications such as biomedical and general to health. In specific cases, the observed counts contain a high frequency of zeros (zero-inflation) and cannot be well accommodated by the widely used Poisson or Negative Binomial model. In general, there are two types of time series models: observation-driven models and parameter-driven models. In observation-driven models, the temporal correlation between adjacent observations is directly modeled through a function of past responses. In contrast, an unobserved latent process is employed in parameter-driven models to account for the serial correlation. In this thesis, we develop both observation-driven models and parameter-driven models for count time series with excess zeros. Also, another type of models that we will analyze, is Hidden Markov models which allow the probability distribution of each observation to depend on the unobserved (hidden) state of a Markov chain. We present the importance of hidden Markov models and we use a real data set on injuries at a hospital. Firstly, we will fit a hidden Markov model with discrete latent value and by extension we consider the use of this modeling methodology to fit general state-space models (SSMs). SSMs have almost the same features with HMMs but have the essential difference that the latent variable can be continuous. For this reason, we consider the use of two discretization approaches related to approximate an SSM likelihood. In addition, we quote the use of two packages of *R* which are called *"ZIM Package"* and *"HHMpa Package"*.

Chapter 2

Injury data series

In this thesis, we will use the occupational injury data from hospital cleaners to illustrate our methodologies. The injury data shown in *Table 2.1* and includes of monthly (4-week) counts of work-related injuries that are routinely reported at an aggregate population level from July 1988 to October 1995. [6]

3 9 0 3 2 2 0 3 0 2 0 0 2 4 2 0 0 1 0 3 0 2 4 3 1 2 1 0 0 3 9 2 2 2 4 0 4 0 6 3 8 6 3 3 3 0 1 0 0 0 0 2 3 0 1 1 0 1 0 0 2 0 1 0 0 0 2 3 0 0 0 0 3 0 0 0 2 0 1 0 0 3 1 0 0 0 0 0 0 0 0 0 2 3 2 0

Table 2.1: Weekly injuries of hospital cleaners

Also, *Table 2.2*, provides summary statistics of the data, including the mean and the variance of the counts.

T	Mean	Var	Min	Max
96	1.46875	3.8305921	0	9

Table 2.2: Summary Statistics of the injury data

As we can observe, the values range from 0 to 9 that is in a range of positive numbers. For this reason, we could use the Poisson distribution to describe the injury data, but it seems that the variance of the data is larger than the mean and this is not applicable to Poisson distribution. More specifically, we can note that the data is displayed overdispersion. [11] Negative binomial regression is a popular generalization of Poisson regression because it loosens the highly restrictive assumption that the variance is equal to the mean made by the Poisson model but and this model may not be the right

one to describe our data due to the large presence of zeros as we can see from *Table 2.1*. In addition, we should note that a participatory ergonomics intervention was introduced in the middle of the study in November 1992 and we can observe the reduction in the injury count as shown in the *Figure 2.1* and we observed a different pattern after the 57th week.

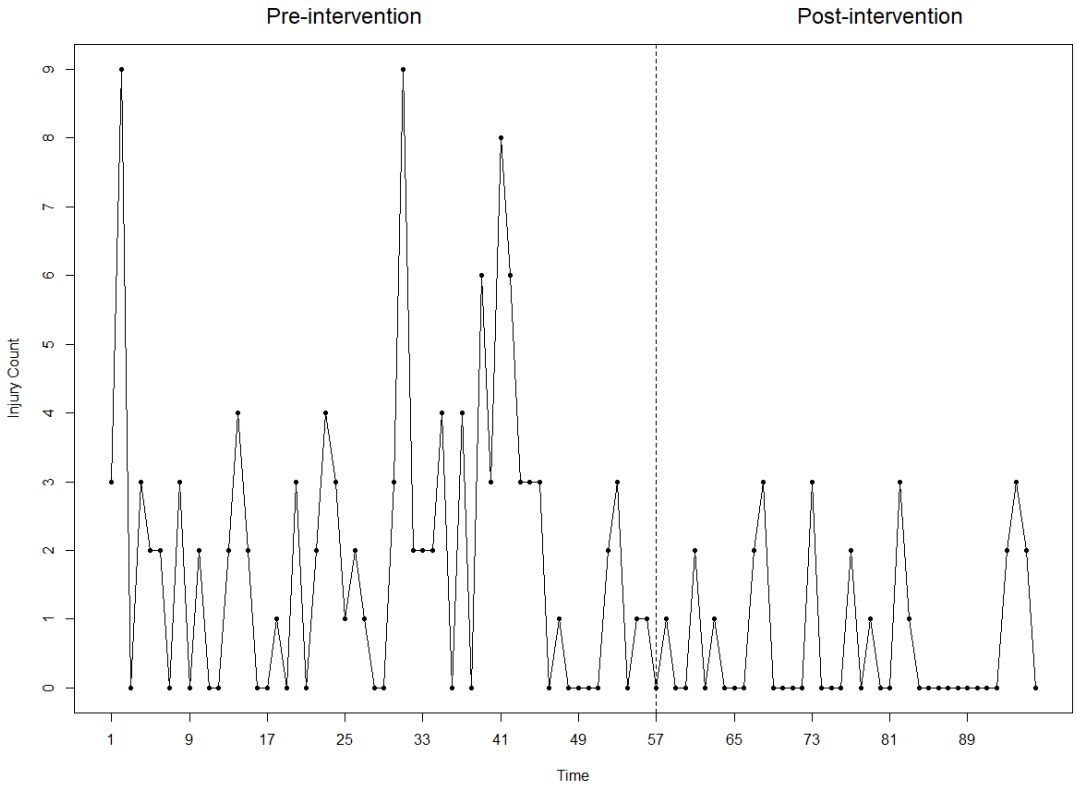

Figure 2.1: Weekly injuries of hospital cleaners before and after intervention

All the above leads us to the conclusion that mixture models and more specifically zero inflated models are the best choice for the description of the injury data.

2.1 Time series for injuries

Time series theory could be useful to present our data characteristics in a better way since we observe, in *Figure 2.1*, that they present trend and seasonality. [12] The data were collected sequentially over a period of $T = 96$ weeks. The trend shows the general tendency of the data to increase or decrease during a long period of time

and seasonality is the presence of variations that occur at specific regular intervals less than a year, such as weekly, monthly, or quarterly.

2.2 Autocorrelation function

The autocorrelation function (ACF) reveals how the correlation between any two values of the signal changes as their separation changes. In simple terms, it describes how well the present value of the series is related with its past values. Let $\{X_t\}$ be a random process, and t be any point in time (t may be an integer for a discrete-time process or a real number for a continuous-time process). Then X_t is the value (or realization) produced by a given run of the process at time t. Then the definition of the auto-correlation function between times t_1 and t_2 is

$$R_{xx}(t_1, t_2) = E[X_{t1}X_{t2}]$$

where E is the expected value operator.[13] In our example X_t will declare the weekly injuries of the hospital cleaners at time t. The ACF of X_t is shown in *Figure 2.2*. The y-axis should be the correlation coefficient and x-axis should be the lag.

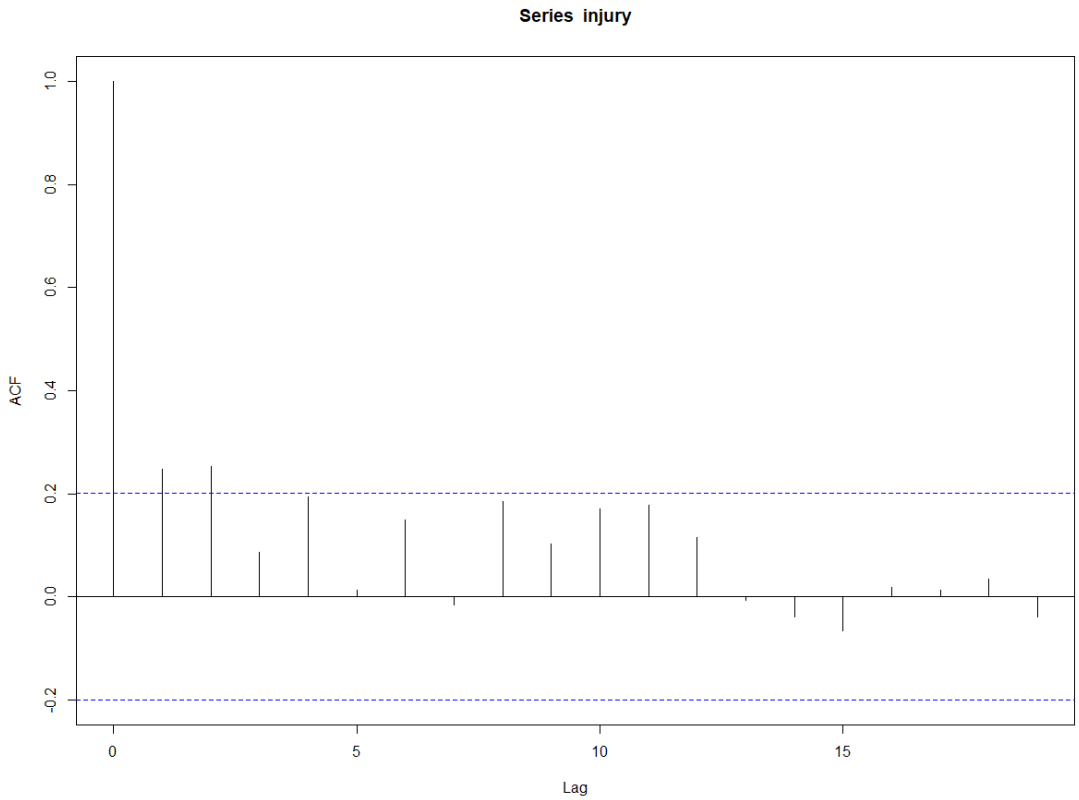

Figure 2.2: Injury data ACF

As we can observe in *Figure 2.2*, the autocorrelation function decay slowly and for this reason we suggest an autoregressive $AR(p)$ for $p = 1$ model which is a representation of a type of random process. Also, for *Figure 2.2*, it is obvious that the observations are serially dependent and the correlation for the most lags is positive.

2.3 Partial Autocorrelation function

The partial autocorrelation function (PACF) gives the partial correlation of a stationary time series with its own lagged values, regressed the values of the time series at all shorter lags. It contrasts with the autocorrelation function, which does not control for other lags.

Given a time series X_t, the partial autocorrelation of lag k, denoted $a(k)$, is the autocorrelation between X_t and X_{t+k} with the linear dependence of X_t on X_{t+1} through X_{t+k+1} removed; equivalently, it is the autocorrelation between X_t and X_{t+k} that is not accounted for by lags 1 through $k-1$, inclusive.

$$a(1) = corr(X_{t+1}, X_t), \qquad for\ k = 1,$$
$$a(k) = corr\left(X_{t+k} - \widehat{X_{t+k}}, X_t - \widehat{X_t}\right), \qquad for\ k \geq 2$$

Where X_{t+k} is the best linear predictor $X_{t+k} = \varphi_{k-1,1}X_{t+1} + \cdots + \varphi_{k-1,k-1}X_{t+k-1}$.
Similarly, $X_t = \varphi_{k-1,1}X_{t-1} + \cdots + \varphi_{k-1,k-1}X_{t-k+1}$.
The quantity $\varphi_{k,k}$ is called the partial autocorrelation of the process $\{X_t\}$ at lag k since it equals the partial correlation between the variables X_{t+k} and X_t adjusted for the intermediate variables $X_{t+1}, X_{t+2}, \dots, X_{t+k-1}$ (or the correlation between X_{t+k} and X_t is not accounted for $X_{t+1}, X_{t+2}, \dots, X_{t+k-1}$). [14]

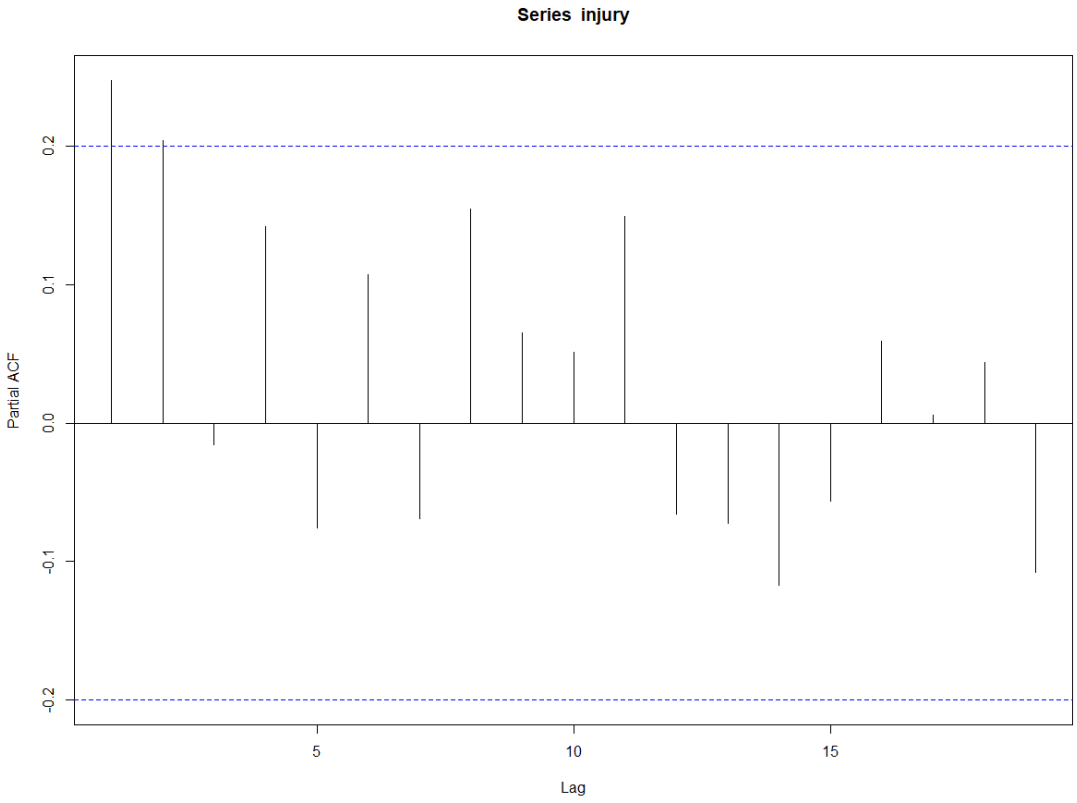

Figure 2.3: Injury data PACF

As we can observe, in *Figure 2.3* there is a significant correlation at lag 1 and followed from correlation that are not significant. This pattern indicates an autoregressive term of order 1: $AR(1)$.

Chapter 3

Observation-driven models

3.1 Definitions and properties

In statistics, a mixture model is a probabilistic model for representing the presence of subpopulations within an overall population, without requiring that an observed data set should identify the sub-population to which an individual observation belongs. Formally a mixture model corresponds to the mixture distribution that represents the probability distribution of observations in the overall population. Consider, the injury data and assume that the counts are identically distributed but there is no serial correlation in the series we can assume that injury data are independent counts. When we have independent unbounded counts, we can say that the Poisson distribution is suitable to describe our data and is presented that:

$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0,1, \ldots,$$

However, as we have shown in *Table 2.2* the sample mean ($\bar{x} = 1.4688$) and the sample variance ($s^2 = 3.8305$) are not equal as usual in the Poisson distribution and the injury data admit more variability than expected under the assumed distribution. For this reason, the Poisson distribution appears unlikely to be appropriate for this data. More specifically, as we have mentioned after the 57th week a participatory ergonomics took place and the injuries of the hospital cleaners were dramatically reduced. So, the injury data can be divided into groups, for example high accident rate and low accident rate.

Furthermore, as we can observe from *Figure 3.1*, another characteristic that appears in injury data, is the large number of zeros about 50% of all observations. This can lead to incorrect parameter estimation. One way of dealing with the overdispersion of data and the large number of zeros is to use a mixture model. [3]

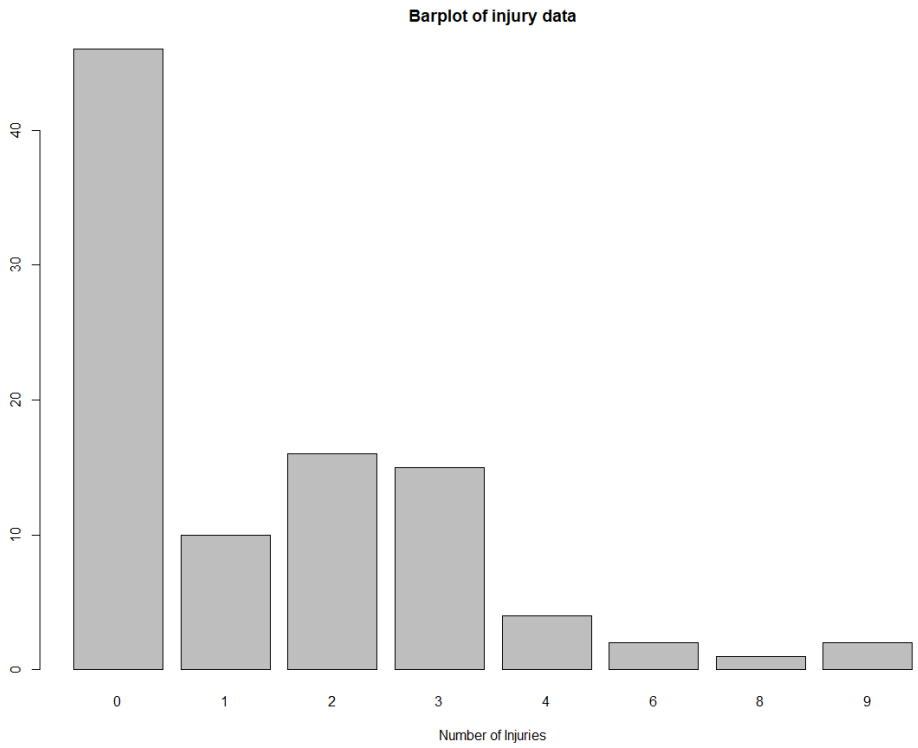

Figure 3.1: Bar plot of injury data

3.2 Zero-inflated Poisson

The first mixture model that we can use in injury data is the Zero-Inflated Poisson model which concerns a random event containing excess zero-count data in unit time. The Zero-Inflated Poisson (ZIP) model mixes two zero generating processes. The first process generates zeros. The second process is governed by a Poisson distribution that generates counts, some of which may be zero. The ZIP distribution with parameters π_i and μ_i, denoted by ZIP (π_i, μ_i), has the following probability mass function:

$$f(y_i|\pi_i, \mu_i) = \begin{cases} \pi_i + (1-\pi_i)exp(-\mu_i), & if \quad y_i = 0 \\ (1-\pi_i)exp(-\mu_i)\dfrac{{\mu_i}^{y_i}}{y_i!}, & if \quad y_i > 0 \end{cases}$$

where $0 \leq \pi_i \leq 1$ and $\mu_i \geq 0$.

The parameter π gives the extra probability thrust at the value 0.

The mean and variance of ZIP (π_i, μ_i) are

$$E(Y) = \mu_i(1 - \pi_i)$$

$$V(Y) = \mu_i(1 - \pi_i)(1 + \mu_i\pi_i)$$

However, we will not proceed with further estimation of the parameters because this model does not solve the problem of overdispersion (variance much larger than the mean). For this reason, we will introduce another type of models which no longer assume that the counts are independent but that they are dependent and contain a serial correlation.

3.3 Observation-driven ZIP models

Considering the injury data which belong to the category of count time series, there is a better and more effective model with which we can describe our data. This model belongs to the category of observation-driven models and more specifically because of the number of zeros in the group of observation-driven ZIP models. Also, with these models we can deal with the overdispersion and temporal correlation that is characteristic of count time series. With observation-driven models, temporal correlation between adjacent observations is directly characterized as a function of past responses. So, we can analyze and introduce an autoregressive model for Poisson time series with excess zeros based on an observation-driven approach. We maintain the same model structure as that introduced in *Section 3.2* to account for the Poisson mixture and moreover, we employ lagged responses as covariates to deal with the temporal correlation.

Let y_t declare the response series, composed of discrete count data. Define the information set

$$F_{t-1} = \sigma y_{t-1}, y_{t-2}, \dots, x_t$$

so as to impersonate all that is known at time t about the dependent variable (response) and any relevant covariate processes. In this way, the vector x_t impersonate an assortment of past and possibly present time-dependent covariates that are observed at

time $t-1$. Given the information F_{t-1}, we assume the count series is conditionally distributed as $ZIP(\lambda_t, \omega_t)$ with probability mass function defined as follows:

$$f_{Y_t}(y_t|F_{t-1};\omega_t,\lambda_t) = \omega_t I_{(y_t=0)} + (1-\omega_t)exp(-\lambda_t)\frac{{\lambda_t}^{y_t}}{y_t!} \qquad (3.4)$$

or equivalently

$$f_{Y_t}(y_t|F_{t-1};\omega_t,\lambda_t) = \begin{cases} \omega_t + (1-\omega_t)exp(-\lambda_t), & if\ \ y_t = 0 \\ (1-\omega_t)exp(-\lambda_t)\frac{{\lambda_t}^{y_t}}{y_t!}, & if\ \ y_t > 0 \end{cases} \qquad (3.5)$$

As before, λ_t introduce the intensity parameter of the baseline Poisson distribution and μ_t the zero-inflated parameter.

Also, we can observe that the ZIP distribution described by (3.4) or (3.5) can be defined as a two-component mixture of Poisson distribution with a dichotomous variable u_t indicating whether the observed variable y_t comes from the degenerate distribution $(u_t = 1)$ or the ordinary Poisson distribution $(u_t = 0)$. [8]

So, we have the following hierarchical model:

$$u_t|F_{t-1} \sim Bernoulli(\omega_t) \qquad (3.6)$$

$$Y_t|u_t, F_{t-1} \sim Poisson((1-u_t)\lambda_t) \qquad (3.7)$$

For any non-negative integer m, the cumulative distribution function (c.d.f.) of $Y_t|F_{t-1}$ is given by:

$$F_{Y_t}(m|F_{t-1}) = Pr(\,Y_t \le m|F_{t-1}) = \sum_{y_t=0}^{m} f_{Y_t}(y_t|F_{t-1})$$

$$= \omega_t + (1-\omega_t)exp(-\lambda_t)\sum_{y_t=0}^{m}\frac{{\lambda_t}^{y_t}}{y_t!}$$

According to the equations (3.6) and (3.7), the mean of $Y_t|F_{t-1}$ can be expressed as

$$E(Y_t|F_{t-1}) = \lambda_t(1-\omega_t)$$

Proof.

$$E(Y_t|F_{t-1}) = E\{E(Y_t|u_t, F_{t-1})\} = E\{(1-u_t)\lambda_t|F_{t-1}\} = \lambda_t(1-\omega_t)$$

and the variance of $Y_t|F_{t-1}$ can be written as:

$$Var(Y_t|F_{t-1}) = \lambda_t(1-\omega_t)(1+\lambda_t\omega_t)$$

Proof.

$$\begin{aligned} Var(Y_t|F_{t-1}) &= E\{Var(Y_t|u_t, F_{t-1})\} + Var\{E(Y_t|u_t, F_{t-1})\} \\ &= E\{(1-u_t)\lambda_t|F_{t-1}\} + Var\{(1-u_t)\lambda_t|F_{t-1}\} \\ &= \lambda_t(1-\omega_t) + {\lambda_t}^2\omega_t(1-\omega_t) = \lambda_t(1-\omega_t)(1+\lambda_t\omega_t) \end{aligned}$$

At this point, we can also observe the difference that exist in relation to the *Section 3.2*. Both parameters are modeled via log-linear and logistic link functions and are presented below:

$$\eta_t = log(\lambda_t) = {\boldsymbol{x}_{t-1}}^T\boldsymbol{\beta} \qquad (3.8)$$

and

$$\xi_t = logit(\omega_t) = {\boldsymbol{z}_{t-1}}^T\boldsymbol{\gamma} \qquad (3.9)$$

where $\boldsymbol{x}_{t-1} = (x_{t-1,1}, \ldots, x_{t-1,p})^T$ and $\boldsymbol{z}_{t-1} = (z_{t-1,1}, \ldots, z_{t-1,q})^T$ are sets of time-dependent explanatory variables for the corresponding vectors of regression coefficients β and γ. Also, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^T$ (3.8) and $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_q]^T$ (3.9) are the regression coefficients for the log-linear and logistic part, respectively. [8]

3.4 Partial likelihood

The estimation of model parameters is usually performed by the method of maximum likelihood (ML) and the case of mixture distribution is not different. The likelihood function plays an important role in the process of estimating the unknown parameters. Also, likelihood function is defined as the joint probability density function of the data, viewed as a function of the parameters. Generally, for a random sample $y_1, \dots, y_n$ the likelihood of a mixture model and more specifically the likelihood of an observation-driven zero inflated model we can express it through partial likelihood. Concretely, partial likelihood estimation uses the product of conditional densities as the density of the joint conditional distribution.

The partial data likelihood of the observed series is:

$$PL(\boldsymbol{\theta}) = \prod_{j=1}^{n} f_{Y_t}(y_t|F_{t-1})$$

where $\boldsymbol{\theta} = [\boldsymbol{\beta^T}, \boldsymbol{\gamma^T}]^{\boldsymbol{T}}$ is the $(p+q)$ −dimensional vector of unknown parameters. The partial likelihood does not demand the derivation of the joint distribution of the response and the covariates. Also, is largely simplified relative to the full likelihood. This approach expedites conditional inference for a large class of transitional processes where the response depends on its past values.

For the observation-driven ZIP model is convenient to take the logarithm of the likelihood, known as the log partial likelihood function which is given below:

$$logPL(\boldsymbol{\theta}) = \sum_{t=1}^{N} log f_{Y_t}(y_t|F_{t-1})$$

$$\sum_{t=1}^{N} log\{\, \omega_t I_{(y_t=0)} + (1-\omega_t) exp(-\lambda_t) \frac{{\lambda_t}^{y_t}}{y_t!}\}$$

In general, the maximization of a mixture likelihood is possible to perform analytically and we must resort to numerical methods and the vector $\hat{\theta}$ gained by maximizing the partial likelihood is called the maximum partial likelihood estimator (MPLE). [8]

3.5 Parameter estimation for observation-driven ZIP model

To get the MPLE, we must maximize $logPL(\boldsymbol{\theta})$. More specifically, we must solve the equation $\boldsymbol{S}_N(\boldsymbol{\theta}) = \mathbf{0}$, where the equation $\boldsymbol{S}_N(\boldsymbol{\theta})$ is defined as follows:

$$\boldsymbol{S}_N(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} logPL(\boldsymbol{\theta}) = \sum_{t=1}^{N} \boldsymbol{C}_{t-1} \boldsymbol{v}_t(\boldsymbol{\theta})$$

with $\boldsymbol{C}_{t-1}$ and $\boldsymbol{v}_t(\boldsymbol{\theta})$ given by:

$$\boldsymbol{C}_{t-1} = \begin{bmatrix} \boldsymbol{x}_{t-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{z}_{t-1} \end{bmatrix}$$

and

$$\boldsymbol{v}_t(\boldsymbol{\theta}) = \begin{bmatrix} v_{1,t}(\boldsymbol{\theta}) \\ v_{2,t}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} y_t - \dfrac{\lambda_t(1-\omega_t I_{(y_t=0)})}{\omega_t + (1-\omega_t)exp(-\lambda_t)} \\ \omega_t\left(\dfrac{I_{(y_t=0)}}{\omega_t + (1-\omega_t)exp(-\lambda_t)} - 1\right) \end{bmatrix}$$

Due to the nonlinear nature of the problem, there is no closed-form solution to the partial score equation. For this reason, we will use iterative algorithms for parameters estimation. [8]

3.5 Simulation Study

In this chapter, we want to present a simulation study and more specifically, we investigate the finite sample behavior of the MPLE. As we have mentioned, we assume that the time series data is generated by the following model:

$$\eta_t = log(\lambda_t) = \beta_0 + \beta_1 I_{(y_{t-1}>0)}$$

and

$$\xi_t = logit(\omega_t) = \gamma_0 + \gamma_1 I_{(y_{t-1}>0)}$$

Then, we consider $\boldsymbol{\theta} = (\beta_0,\ \beta_1,\ \gamma_0,\ \gamma_1)^T = (1.2, 0.6, 0.4, -0.8)^T$ is the true parameter vector.

Table 3.1 summarizes the finite sample results of the MPLE under three different sample sizes $N = (100, 200, 500)$.

Simulation Results				
$\boldsymbol{N}$	$\boldsymbol{\theta}$	$\widehat{\boldsymbol{\theta}}$	Bias	SE
100	$\boldsymbol{\beta_0}$	*0.92*	*0.28*	*0.17*
	$\boldsymbol{\beta_1}$	*0.71*	*-0.11*	*0.19*
	$\boldsymbol{\gamma_0}$	*0.37*	*0.03*	*0.35*
	$\boldsymbol{\gamma_1}$	*-0.38*	*0.42*	*0.43*
200	$\boldsymbol{\beta_0}$	*1.23*	*-0.03*	*0.10*
	$\boldsymbol{\beta_1}$	*0.58*	*0.02*	*0.11*
	$\boldsymbol{\gamma_0}$	*0.57*	*-0.17*	*0.22*
	$\boldsymbol{\gamma_1}$	*-0.74*	*0.06*	*0.29*
500	$\boldsymbol{\beta_0}$	*1.08*	*0.12*	*0.06*
	$\boldsymbol{\beta_1}$	*0.75*	*-0.15*	*0.07*
	$\boldsymbol{\gamma_0}$	*0.55*	*-0.15*	*0.14*
	$\boldsymbol{\gamma_1}$	*-0.92*	*-0.12*	*0.18*

Table 3.1: Finite sample results of the MPLE simulated independently from ZIP model

As we can observe from the results in the *Table 3.1* the bias of the estimated parameters decreases as the sample increases. Note here, that the bias of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated. Also, bias shows to us how accurate is an estimator and in this sense means that it is neither an overestimate nor an underestimate. If an overestimate or underestimate does happen, the mean of the difference is called a "bias". Furthermore, in the next column of the *Table 3.1* we can see the standard error values. The standard error of the estimate is a measure of the accuracy of predictions. From the results we see that the standard error decreases as the samples increases. In addition, the values

between the initial values of the parameters in relation to the estimated, we can be seen to be very close.

3.6 Observation-driven model for injury data

Back to injury data, we can adjust the model accordingly and, in this way, estimate our parameters. Then we have the model in which we get the first equation when we have zero accidents to the hospital cleaners and the other equation when we have several accidents greater than zero.

$$f_{Y_t}(y_t|F_{t-1};\omega_t,\lambda_t) = \omega_t I_{(y_t=0)} + (1-\omega_t)exp(-\lambda_t)\frac{{\lambda_t}^{y_t}}{y_t!}$$

Also, we represent both parameters and we use an autoregressive model of order $p = 1$, abbreviated $AR(1)$. Therefore, we have the following equations:

$$\eta_t = log(\lambda_t) = \beta_0 + \beta_1 I_{(y_{t-1}>0)} + \beta_2 x_t$$

and

$$\xi_t = logit(\omega_t) = \gamma_0 + \gamma_1 x_t$$

Here, $x_t = \frac{t}{1000}$ represents the deterministic linear trend, which is always forced in the model since characterizing the trend is the primary objective of the study.

ZIP Model (AIC = 306.21)			
$\boldsymbol{\theta}$	Estimate	SE	P-Value
$\boldsymbol{\beta_0(Intercept)}$	*1.01*	*0.22*	*<0.05*
$\boldsymbol{\beta_1(AR1)}$	*0.38*	*0.19*	*0.05*
$\boldsymbol{\beta_2(Trend)}$	*-7.45*	*4.16*	*0.07*
$\boldsymbol{\gamma_0(Intercept)}$	*-1.16*	*0.50*	*0.02*
$\boldsymbol{\gamma_1(Trend)}$	*17.68*	*9.41*	*0.06*

Table 3.2: Final ZIP autoregression on observation-driven model for the injury data

Table 3.2 displays the regression output for the observation-driven model. As we can see the p-values of the coefficients are less than 0.05 and this tells us that they are statistically significant as the standard errors seem to be relatively small except for the trend parameter. The Akaike criterion has a similar value to the parameter-driven model which we analyze in *Chapter 4*.

Chapter 4

Parameter-driven models

4.1 Definition

In this chapter, we focus on parameter-driven models for zero-inflated time series. Firstly, we can mention that the parameter-driven model is an extension of the traditional linear model. Also, in time series, there are some cases with a high frequency of zeros and in this situation, we introduce a class of parameter-driven models for count time series in which the parameter estimation can be performed with a Monte Carlo EM (MCEM) algorithm.

4.2 Dynamic ZIP Model

In this section, we propose a dynamic ZIP model to accommodate zero-inflation in count time series. Specifically, we assume there is a stationary $AR(p)$ process $\{z_t\}$ such that

$$z_t = \varphi_1 z_{t-1} + \cdots + \varphi_p z_{t-p} + \varepsilon_t$$

where ε_t is a white noise process with mean 0 and variance σ^2. Conditioning on the current state z_t, we assume that the observation y_t has a ZIP distribution with probability mass function

$$f_{Y_t}(y_t|z_t;\lambda_t,\omega) = \begin{cases} \omega + (1-\omega)exp(-\lambda_t)\,, & if\ y_t = 0 \\ (1-\omega)exp(-\lambda_t)\dfrac{{\lambda_t}^{y_t}}{y_t!}, & if\ y_t > 0 \end{cases}$$

We use the following log-linear model to characterize the intensity parameter λ_t:

$$log(\lambda_t) = log(w_t) + {x_t}^T\boldsymbol{\beta} + z_t,$$

where x_t is a set of explanatory variables and $\boldsymbol{\beta}$ is the vector of regression coefficients. In addition, $log(w_t)$ is referred to as the offset variable and let $\boldsymbol{\theta} = (\omega, \boldsymbol{\beta}^T \boldsymbol{\varphi}^T, \sigma)^T$ denote the vector of unknown parameters. For simplicity, the zero-inflation parameter ω is considered as constant but if we want, we can represent it with a separate logistic model. [6]

The dynamic ZIP model can be written in the following state-space form:

$$s_t \,|s_{t-1} \;\sim\; N_p(\boldsymbol{\Phi} s_{t-1}, \boldsymbol{\Sigma}) \tag{3.1}$$

$$u_t \sim Bernouli(\omega) \tag{3.2}$$

$$y_t | s_t, u_t \;\sim\; Poisson((1-u_t)\lambda_t), \tag{3.3}$$

where $s_t = \left(z_t, \dots, z_{t-p+1}\right)^T$ is a p-dimensional state vector and u_t is an unobservable membership indicator. The initial state s_0 is assumed to be normally distributed with mean $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$. Here $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$ are $p \times p$ matrices defined as follows:

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Note that the covariance matrix Σ is not positive definite. *Figure 4.1* illustrates the dynamic ZIP model that is defined by (3.1) - (3.3).

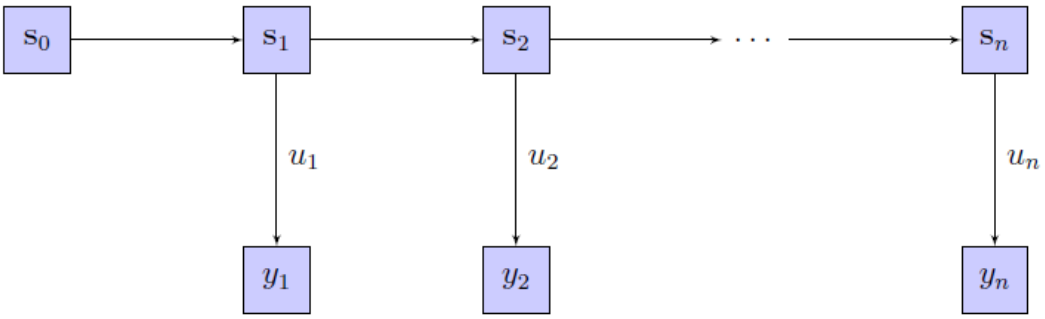

Figure 4.1: Graphical illustration of the state evolution and data generation in the dynamic ZIP model.

4.3 Likelihood

Based on the state-space representation (analytically in *Chapter 7*) and with the same methodology as in the previous Chapter to find the estimated parameters of the parameter-driven ZIP model, we will decompose the complete-data likelihood as follows:

$$L_C(\theta) = f(s_0) \prod_{t=1}^{n} f(s_t|s_{t-1}) \prod_{t=1}^{n} f(u_t) \prod_{t=1}^{n} f(y_t|s_t, u_t)$$

Also, the complete-data log-likelihood is given by:

$$\begin{aligned} l_C(\boldsymbol{\theta}) = \frac{n}{2} log\sigma^2 & \\ & - \frac{1}{2\sigma^2} \sum_{t=1}^{n} (z_t - \varphi^T s_{t-1})^2 \\ & + \sum_{t=1}^{n} \{u_t log\omega + (1-u_t) log(1-\omega)\} \\ & + \sum_{t=1}^{n} (1-u_t)\{y_t x_t{}^T \boldsymbol{\beta} - w_t exp(x_t{}^T \boldsymbol{\beta} + z_t)\}. \end{aligned}$$

For the implementation of the EM algorithm, we need to compute the expectation of $l_C(\boldsymbol{\theta})$ given the observed data but there is no analytical form for the conditional expectation due to the non-normality of the data. For this reason, for the approximation

of the conditional expectation we use Monte Carlo extensions that called Kalman methods. [6]

4.4 Simulation Study

In this section, we assume that the time series data are simulated from a parameter-estimation ZIP model from which we suppose z_t is an $AR(1)$ autoregressive process such that

$$z_t = \varphi z_{t-1} + \varepsilon_t,$$

where ε_t is a white noise process with mean 0 and standard deviation $\sigma = 0.5$. Furthermore, we will set $\varphi = 0.8$ to ensure that the $AR(1)$ autoregressive process is stationary.

Conditioning on the current state z_t, we assume the observation y_t has a ZIP distribution with probability mass function by:

$$f_{Y_t}(y_t|z_t;\lambda_t,\omega) = \begin{cases} \omega + (1-\omega)exp(-\lambda_t)\,, & if\ y_t = 0 \\ (1-\omega)exp(-\lambda_t)\dfrac{{\lambda_t}^{y_t}}{y_t!}, & if\ y_t > 0 \end{cases}$$

where we have $log(\lambda_t) = \beta_0 + z_t$ with $\beta_0 = 0.2$. Thus, summarizing all the above we will have the follow true parameters in the generating model.

- Parameter-driven ZIP Model

$\omega = 0.6$, $\beta_0 = 0.2$, $\varphi = 0.8$ and $\sigma = 0.5$.

Also, the sample sizes (i.e., length of the series) is set to be 100, 200 and 500.

Simulation Results				
	$\boldsymbol{\omega}$	$\boldsymbol{\beta_0}$	$\boldsymbol{\varphi}$	$\boldsymbol{\sigma}$
True	*0.60*	*0.20*	*0.80*	*0.50*
ZIP-100	*0.65*	*0.75(0.25)*	*0.65(0.26)*	*0.64*
ZIP-200	*0.53*	*0.15(0.21)*	*0.78(0.08)*	*0.95*
ZIP-500	*0.61*	*0.18(0.14)*	*0.86(0.04)*	*0.91*

Table 4.1: True and estimated parameters (with standard errors) for the simulated examples.

In *Table 4.1* we present the estimates of the parameters from the simulation experiment under three different sample sizes $N = (100, 200, 500)$. In the first line we can see the real value of the parameters and in the next ones we obtain their estimates together with the standard errors. Also, the MCEM algorithm is used to fit models and to reduce time requirements we set $N = 500$ and $R = 500$ and we stopped the MCEM algorithm after 100 iterations. From the results, we can conclude that with the increase of the sample size the parameter estimates are closer to the real values and the standard errors are reduced.

4.5 Parameter-driven model for injury data

We revisit the application pertaining to occupational injury data. As we have mentioned, the application concerns the assessment of a participatory ergonomics intervention in reducing the incidence of workplace injuries among a group of hospital cleaners. The data consists of monthly (4-week) counts of work-related injuries that were routinely reported at an aggregate population level from July 1988 to October 1995. During the study period, many zero counts are observed due to the heterogeneity in risk and the dynamic worker population. Since the injury count series contain excess zeros relative to a Poisson distribution, we modelled the data using a Zero Inflated Poisson (ZIP) mixed autoregression.

In our analyses, we will assume that there is autocorrelation which is explained by the simple $AR(1)$ structure. Specifically, we employ the linear predictor

$$log(\lambda_t) = \beta_0 + \beta_1 x_t + z_t, \qquad t = 1, \dots, 96,$$

where $x_t = I_{t>57}$ is a dummy variable indicating whether the time index t is greater than the intervention time (57 months). Also, β_1 reverberate the reduction in injury risk due to intervention. We consider the dynamic model structure ZIP + AR (1) and *Table 4.2* shows results for this.

ZIP + $\boldsymbol{AR(1)}$ (AIC = 309.09)			
	Estimate	SE	P-Value
$\boldsymbol{\omega}$	0.29	-	-
$\boldsymbol{\beta_0}$	0.89	0.22	<0.05
$\boldsymbol{\beta_1}$	-1.00	0.31	<0.05
$\boldsymbol{\varphi}$	0.41	0.26	0.11
$\boldsymbol{\sigma}$	0.44	0.05	-

Table 4.2: Parameter estimates (standard errors) for dynamic model fit to the injury count series

Table 4.2 displays the regression output for the parameter-driven model. As we can observe the coefficient β_1 is less than zero and this is something to be expected because it shows us that after the ergonomics intervention the number of injuries from hospital cleaners is decreasing. That is, the model indicates a significant reduction of work-related injuries after the introduction of the intervention. Also, we can see that the p-values of the coefficients are less than 0.05 and this tells us that they are statistically significant as the standard errors seem to be relatively small. The Akaike criterion has a similar value to the observation-driven model.

4.6 The R Package 'ZIM'

For all calculations that we have presented so far have been done with the help of the ZIM package. ZIM package (Zero-Inflated Models) has developed to analyze count time series with excess zeros. [20] We can use it in both observation-driven models and parameter-driven model as follows:

- Observation-driven models

For this category of models and for the case of the zero-inflated Poisson which are presented in *Chapter 3* we can use the ZIM package with the following functions:

zim: The function *zim* is a user-friendly function to fit zero-inflated observation-driven models. Its usage is very similar to that of the well-known function *glm.*

- Parameter-driven models

Compared to observation-driven models, parameter estimation in parameter-driven models is much more challenging. The following is the function that can be used to fit the parameter-driven ZIP model that has been proposed in *Chapter 4.*

dzim: The function *dzim* is a user-friendly function to fit zero-inflated parameter-driven models. The default order for the autoregressive process is assumed to be one.

Chapter 5

Markov Chains

In this section, we introduce Markov chains, a second building-block of hidden Markov models. Our treatment is restricted to those few aspects of discrete-time Markov chains that we need. Thus, although we shall make passing reference to properties such as irreducibility and aperiodicity, we shall not dwell on such technical issues. A Markov process is a stochastic process that satisfies the Markov property and more specifically, conditional on the present state of the system, its future and past states are independent. In simpler terms, it is a process for which predictions can be made regarding future outcomes based solely on its present state. [15]

5.1 Definition

A Markov chain is a sequence of discrete random variables

$$C_1, C_2, C_{3,\dots}$$

with the Markov property, namely that the probability of moving to the next state depends only on the present state and not on the previous states:

$$P(C_{t+1}|C_1, C_2, \dots, C_{t-1}, C_t) = P(C_{t+1}|C_t)$$

In the other words, given the present of the process, C_t , its future, C_{t+1}, is independent of its past, $C_{t-1}, C_{t-2}, \dots, C_1$.

5.2 Types of Markov chains

The system's state space and time parameter index need to be specified. *Table 5.1* gives an overview of the different instances of Markov processes for different levels of state space generality and for discrete time v. continuous time. As we can see from *Table 5.1*

Markov chains can be divided into discrete time and continuous time. A discrete time Markov chain is a sequence of random variables $C_1, C_2, C_{3,\dots}$ with the Markov property. A continuous-time Markov chain C_t, for $t \geq 0$ is defined by a finite or countable state space *S*, a transition rate matrix Γ with dimensions equal to that of the state space and initial probability distribution defined on the state space. [16]

	Countable state space	Continuous state space
Discrete time	Discrete time Markov chain on a countable or finite state space	Markov chain on a measurable state space
Continuous time	Continuous-time Markov process	Any continuous stochastic process with the Markov property

Table 5.1: Overview of the different instances of Markov processes

5.3 Transition probability matrix

The probabilities associated with various state changes are called transition probabilities. The process is characterized by a state space, a transition matrix describing the probabilities of particular transitions, and an initial state (or initial distribution) across the state space. [1]

Important quantities associated with a Markov chain are the conditional probabilities called transition probabilities:

$$Pr(C_{s+t} = j | C_s = i).$$

If these probabilities do not depend on *s*, the Markov chain is called homogeneous, otherwise nonhomogeneous. Unless there is an explicit indication to the contrary, we shall assume that the Markov chain under discussion is homogeneous, in which case the transition probabilities will be denoted by:

$$\gamma_{ij}(t) = Pr(C_{s+t} = j | C_s = i)$$

Therefore, we can define as transition probability matrix $\boldsymbol{\Gamma}$ for Markov chain $\{C_t\}$ at time t the matrix with (i,j) element $\gamma_{ij}(t)$. This means each row of the matrix is a probability vector, and the sum of its entries is 1. So, we present a mathematical presentation of the $\Gamma_{m\times m}$ transition probability matrix:

$$\Gamma_{m\times m} = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ \vdots & \ddots & \vdots \\ \gamma_{m1} & \cdots & \gamma_{mm} \end{pmatrix}$$

where m denotes the number of states of the Markov chain and the row sums for each state $i,\ i = 1,2 \ldots, m$ are equal to 1 which we can write mathematically: [1]

$$\sum_{j=1}^{m} \gamma_{ij} = 1$$

5.4 Stationary Markov chain

A stationary distribution δ is a (row) vector, whose entries are non-negative and sum to 1, is unchanged by the operation of transition matrix $\boldsymbol{\Gamma}$ on it and so is defined by:

$$\delta\boldsymbol{\Gamma} = \delta$$

and

$$\delta\mathbf{1}' = \delta$$

where $\mathbf{1}'$ is a row vector of ones, δ is a probability distribution with all requirements (of a pdf) satisfied. [1]

Chapter 6

Hidden Markov Models

Consider again the observed injury data series displayed in *Figure 1.1* on p. 4 we have mentioned that the observations are unbounded counts and the Poisson distribution is a natural choice to describe them. On the other hand, as we noticed in *Table 2.2*, the sample variance of the observations is substantially greater than the sample mean, indicating overdispersion relative to the Poisson distribution. For this reason, we could use a mixture model to describe our injury data and in fact we would avoid the overdispersion in this way. But with this approach we will deal with the problem for the serial dependence which there is in our data. The sample autocorrelation function, displayed in *Figure 6.1*, gives a clear indication that the observations are serially dependent. The main way we could apply, is to relax the assumption that the parameter process is serially independent. In this situation, we can use the idea that we have a Markov chain and we analyze the resulting model for the observations which is called a Poisson-hidden Markov model and it belongs to the category of hidden Markov models (HMMs). [18]

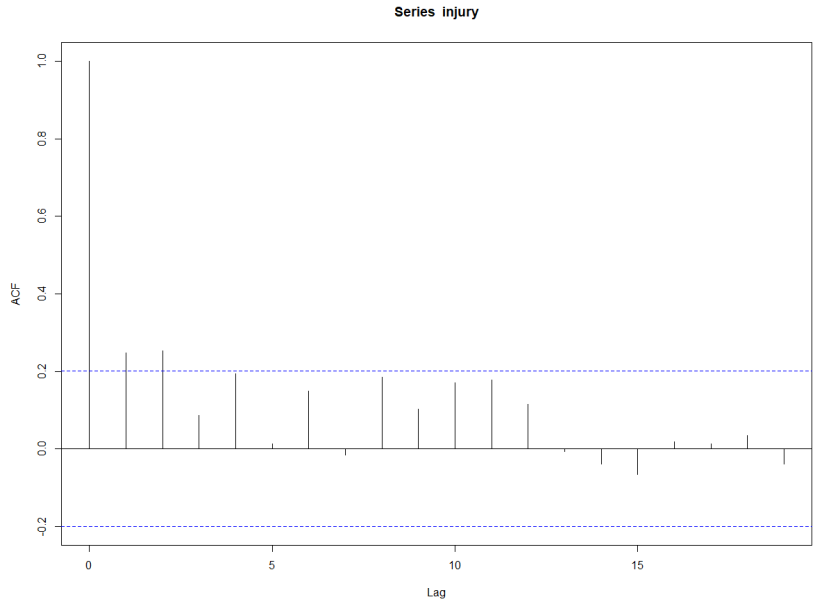

Figure 6.1: Injury data ACF

6.1.1 Definition

Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process – call it C_t – with unobservable ("*hidden*") states. HMM assumes that there is another process X_t whose behavior "depends" on C_t. The goal is to learn about C_t by observing X_t. More specifically, a hidden Markov model $\{X_t : t \in N\}$ is a particular kind of dependent mixture. We can assume that $X_1, X_2, \dots, X_t$ be observations and $C_1, C_2, \dots, C_t$ be the corresponding latent states representing the state of the process at time 1 to time t respectively. One can summarize the simplest model of this kind by:

$$Pr(C_t|C_{t-1}, \dots, C_1) = Pr(C_t|C_{t-1}), \quad t = 2,3, \dots$$

$$Pr(X_t|X_{t-1}, C_t) = Pr(X_t|C_t), \quad t \in N$$

If the Markov chain $\{C_t\}$ has m states, we call $\{X_t\}$ an m-state HMM. All the above, are summarized in *Figure 6.2* which is illustrated a graphical representation of an HMM.

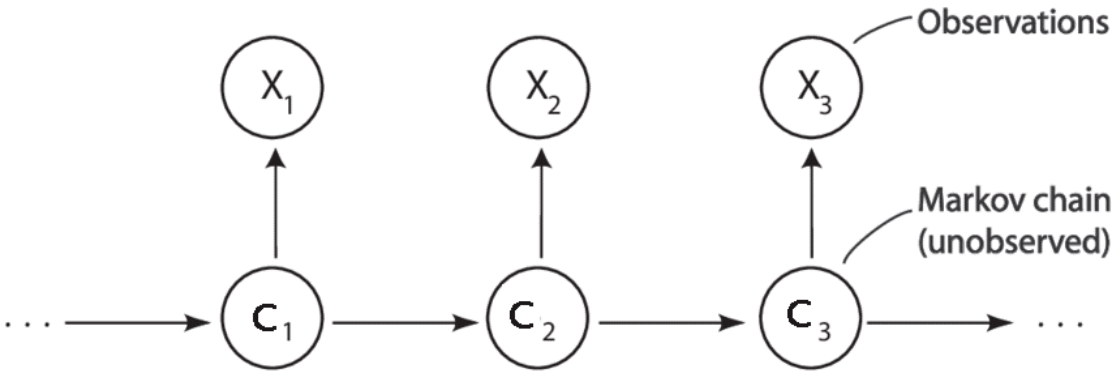

Figure 6.2: Directed graph of basic HMM.

Furthermore, a joint probability of an HMM is easy to be calculated of a set of discrete or continuous observed variables $X^{(T)} = \{X_1, X_2, \dots, X_t\}$, and discrete hidden variables $C^{(T)} = \{C_1, C_2, \dots, C_t\}$, for T observations that factors the joint distribution as follows:

$$P\left(X^{(T)}, C^{(T)}\right) = C_1 \prod_{t=1}^{T-1} P(C_{t+1}|C_t) \prod_{t=1}^{T} P(X_t|C_t) \qquad (1)$$

6.2 The Likelihood

Following our analysis of the Hidden Markov models, the aim is to represent a definite formula for the likelihood L_T of T consecutive observations $x_1, x_2, \dots, x_T$ assumed to be generated by an m-state HMM. The likelihood shows us how probable is to observe specific data under a certain model. Hence, we suppose there is an observation sequence $x_1, x_2, \dots, x_T$ generated by such a model.

6.2.1 Definition

The likelihood of an HMM is given as

$$L_T = \delta P(x_1) \Gamma P(x_2) \Gamma P(x_3) \dots \Gamma P(x_T) 1'. \qquad (2)$$

Where, $P(x)$ is defined as the diagonal matrix with ith diagonal element $p_i(x)$, for $i = 1, \dots, m$ which is the state-dependent probability (density) function. [1] Furthermore, this HMM has initial distribution δ and transition probability matrix Γ. If δ is the stationary distribution of the Markov chain, then *Equation 2* can be equivalently expressed as:

$$L_T = \delta \Gamma P(x_1) \Gamma P(x_2) \Gamma P(x_3) \dots \Gamma P(x_T) 1'. \qquad (3)$$

Proof.

We present only the case of discrete observations. First note that from *Equation 1*,

$$P\left(X^{(T)} = x^{(T)}\right) = \sum_{c_1,\dots,c_T=1}^{m} P\left(X^{(T)}, C^{(T)}\right)$$

$$= \sum_{c_1,\dots,c_T=1}^{m} P(C_1) \prod_{t=1}^{T-1} P(C_{t+1}|C_t) \prod_{t=1}^{T} P(X_t|C_t)$$

As a result,

$$L_T = P\left(X^{(T)} = x^{(T)}\right) = \sum_{c_1,\dots,c_T=1}^{m} P(C_1)\prod_{t=1}^{T-1} P(C_{t+1}|C_t)\prod_{t=1}^{T} P(X_t|C_t)$$

$$= \sum_{c_1,\dots,c_T=1}^{m} \delta_{c_1} p_{c_1}(x_1)\gamma_{c_1,c_2}\, p_{c_2}(x_2)\gamma_{c_2,c_3} p_{c_3}(x_3)\dots\gamma_{c_{T-1},c_T} p_{c_T}(x_T)$$

$$= \delta P(x_1)\Gamma P(x_2)\Gamma P(x_3)\dots\Gamma P(x_T)1'$$

6.3 Poisson HMM

The first model we will choose to describe our data and belongs to the category of HMMs is a Poisson-HMM, where $X_t|C_t$ has the Poisson distribution. So, we will try to fit an m-state Poisson hidden Markov model with $m = 1,2,3,4$ in order to analyze the injury data. For this process we will use optimization methods and for this reason we need to make sure that parameters are transformed properly so as constraints will not violated. In general, there are two groups of constraints: those that apply to the parameters of the state-dependent distributions and those that apply to the parameters of the Markov chain. [1]

Transformation 1: Estimated Poisson parameters λ_i that is, the means λ_i are always non-negative

$$\kappa_i = log(\lambda_i)$$

where $i = 1, \dots, m$

Transformation 2: For initial distribution, δ , in the case where δ is stationary provided that the mean Poisson λ_i is selected with probability δ_i, where $i = 1, \dots, m$ we will transform using the logit function which is defined in interval (0,1).

$$logit(\omega_i) = \delta_i$$

$$\omega_i = log\left(\frac{\delta_i}{1 - \sum_{j=1}^{m}\delta_j}\right)$$

Transformation 3:

Define the matrix

$$T = \begin{pmatrix} - & \tau_{12} & \tau_{13} \\ \tau_{21} & - & \tau_{23} \\ \tau_{31} & \tau_{32} & - \end{pmatrix}$$

And let g be a strictly increasing function e.g. $g(x) = exp(x)$

We then set

$$\gamma_{ij} = \frac{g(\tau_{ij})}{\sum_{j=1}^{m} g(\tau_{ij})}$$

So, generally for m-state

$$\gamma_{ij} = \frac{exp(\tau_{ij})}{1 + \sum_{i \neq j}^{m} exp(\tau_{ij})}$$

$$\tau_{ij} = log\left(\frac{\gamma_{ij}}{1 - \sum_{i \neq j}^{m} \gamma_{ij}}\right)$$

for $i \neq j$

Parameters ω_i, κ_i and τ_{ij} are called working parameters. [1]

After making the above transformations we are able to fit the models, using the *R* optimizer *optim*. More specifically, this can be done by numerically maximizing the corresponding log-likelihoods. In *Table 6.1* we present the results and the comparison among 1-state, 2-state, 3-state and 4-state of the Poisson-HMM. The choice of m will be based on AIC and BIC information criteria.

m-state	***-logL***	***n***	***AIC***	***BIC***
m=1	186.8437	1	375.6873	378.2517
m=2	157.4736	4	322.9472	**333.2046**
m=3	151.0059	9	**320.0118**	343.0910
m=4	151.6011	16	335.2023	376.2319

Table 6.1: Comparison of Poisson HMMs by AIC and BIC

As, we can see the 1-state and 4-state HMM have both the highest value in both criteria. On the other hand, the 3-state HMM has the smallest $AIC = 320.0118$ but the 2-state model has the smallest $BIC = 333.2046$. In this situation, we will choose the AIC criterion since it is more appropriate in case of prediction. [19]

Results of a 3-state HMM:

$$\Gamma = \begin{pmatrix} 0.42 & 0.49 & 0.09 \\ 0.48 & 0.51 & 0.01 \\ 0.09 & 0.15 & 0.76 \end{pmatrix}$$

$$\delta = (0.40, 0.45, 0.15)$$

$$\lambda = (0.00, 1.85, 4.65)$$

$$Loglikelihood = -151.0059$$

$$AIC = 320.0118$$

$$BIC = 343.0910$$

6.4 Package 'HMMpa'

The above results are calculated with the R optimizer *optim*. On the other hand, there is a package called HMMpa and it gives us the ability to perform the same calculations. This package provides functions that we can use to estimate all the necessary parameters and it help us to save time programming.

More specifically, first a hidden Markov model is trained to estimate the number of m of hidden physical activity states and the model specific parameters delta (stationary process), gamma (transition probability matrix) and the parameters of the data distribution. Then, an algorithm decodes the trained HMM to classify each time series count into the m hidden physical activity states. Finally, the estimated distribution mean values (PA-levels) corresponding to the hidden physical activity states are extracted and the time series counts are assigned by the total magnitudes of their corresponding PA-levels to given physical activity ranges by the traditional cut-off point method. [9]

We illustrate some examples of HMMpa orders and their corresponding results:

Results of a 3-state HMM:

$$\Gamma = \begin{pmatrix} 0.42 & 0.50 & 0.06 \\ 0.46 & 0.53 & 0.01 \\ 0.09 & 0.21 & 0.68 \end{pmatrix}$$

$$\delta = (0.00, 8.45x10^{-65}, 1.00)$$

$$\lambda = (0.00, 1.86, 4.95)$$

$$Loglikelihood = -148.8881$$

$$AIC = 331.7763$$

$$BIC = 375.3702$$

Chapter 7

State space Models

7.1 Definition

State space model (SSM) refers to a class of probabilistic graphical model (Koller and Friedman, 2009) that describes the probabilistic dependence between the latent state variable and the observed measurement. The state or the measurement can be either continuous or discrete. In the previous chapter, we have analyzed the HMMs in which the state process is discrete. However, if we go back to the injury data and assume that we have two or three states (small or large number of injuries), then the difference between them can be quite large in contrast to having more states and the difference from one to the other being less noticeable. So, we will use state space models that share the same dependence structure as HMMs, and we want to succeed the better estimations.

Let X_t be the number of injuries to the hospital cleaners at time t and C_t be the respective latent states for $t = 1,\ldots,96$. In this category of models, in contrast to HMMs the states are continuous and more specifically are characterized by two processes: a continuous-valued Markov state process, $\{C_t\}$ and an observation process $\{x_t\}$. They can be described by the following continuous distribution:

$$x_t = \varepsilon_t \beta exp\left(\frac{C_t}{2}\right), \tag{7.1}$$

$$C_t = \varphi C_{t-1} + \sigma\eta_t, \tag{7.2}$$

where $|\varphi| < 1$, $\beta, \sigma > 0$, $\eta_t \sim N(0,1)$ known as process noise and $\varepsilon_t \sim N(0,1)$ known as observation error are independent sequences of independent standard normal random variables. [1] Here the state process, $\{C_t\}$ is an autoregressive process of lag 1. *Equation*

(7.1) describes the relationship between the true abundance and the observations x_t at time t. *Equation* (7.2) describes the evolution of the unobserved state variables $\{C_t\}$.[2] Also, if we assume that

$$C_t \sim Normal\left(0, \frac{\sigma^2}{1-\varphi^2}\right) \qquad (7.3)$$

Proof.

$$Cov(C_t, C_t) = Var(C_t) = Cov(\varphi C_{t-1} + \sigma\eta_t, \varphi C_{t-1} + \sigma\eta_t) = \varphi^2 Var(C_t) + \sigma^2$$

Therefore, $Var(C_t) = \frac{\sigma^2}{1-\varphi^2}$

Then, if we want to calculate the likelihood of a state space model, we have to integrate all possible values of the state process at each time an observation is made because cannot be computed explicitly in general. We adopt a discretization approach of the state space into m states that avoid the complexity of the multiple integrals and the model can be approximated by an HMM. [2] [5]

We will start with a finite range of C_t-values by split into m equally intervals $B_i = (b_{i-1}, b_i), i = 1, \ldots, m$ of equal length $h = \frac{b_m - b_0}{m}$.

7.2 Numerical Integration of Likelihood

The following proof provide a result for the calculation of the likelihood as we have calculated in Chapter 6.2 but instead of a sum of products, we have a multiple integral with respect to C_i states $i = 1, \ldots, m$.

Proof.

$$L_T = \int \dots \int f(\boldsymbol{x}; \boldsymbol{c})\, d\boldsymbol{c}$$

$$= \int \dots \int f(x_1, \dots, x_T, c_1, \dots, c_T)\, dc_T \dots c_1$$

$$= \int \dots \int f(x_1, \dots, x_T | c_1, \dots, c_T) f(c_1, \dots, c_T)\, dc_T \dots c_1$$

$$= \int \dots \int f(c_1) f(x_1|c_1) \prod_{t=2}^{T} f(c_t|c_{t-1}) f(x_t|c_t)\, dc_T \dots c_1$$

$$\approx \int_{b_0}^{b_m} \dots \int_{b_0}^{b_m} f(c_1) f(x_1|c_1) \prod_{t=2}^{T} f(c_t|c_{t-1}) f(x_t|c_t) dc_T \dots c_1$$

where f is used as a general symbol for a density.

Then, we will make the following substitutions to make computing calculation effective and straightforward:

- The first approach is

$$f(c_t|c_{t-1}) \approx f\left(c_t \in B_{i_t} \middle| c_{t-1} = {b_{i_t}}^*\right)$$

$$f(X_t|c_t) \approx f\left(X_t \middle| {b_{i_t}}^*\right)$$

where ${b_{i_t}}^*$ a representative point in B_{i_t} e.g. the midpoint. [2] [5]

In more detail, the innermost integral in the multiple integrals has been approximated as follows:

$$\int_{b_0}^{b_m} f(c_T|c_{T-1}) f(x_T|c_T) dc_T \approx \sum_{i_T=1}^{m} f\left(c_T \in B_{i_T} \middle| c_{T-1} = {b_{i_{T-1}}}^*\right) f\left(x_T \middle| {b_{i_T}}^*\right)$$

and using the following property:

$$\int_a^b f_1(x) f_2(x) dx \approx f_1(r) \int_a^b f_2(x) dx$$

where r is a representative point in (a, b), for functions $f_1(x)$ and $f_2(x)$.

Therefore, the likelihood replaces it with:

$$L_T \approx \sum_{i_1=1}^{m} \dots \sum_{i_T=1}^{m} f\left(c_1 \in B_{i_1}\right) f\left(X_1 \middle| {b_{i_1}}^*\right) \prod_{t=2}^{T} f\left(c_T \in B_{i_t} \middle| {b_{i_{t-1}}}^*\right) f\left(X_t \middle| {b_{i_t}}^*\right) =$$

$$\sum_{i_1=1}^{m} \dots \sum_{i_T=1}^{m} f\left(c_1 \in (b_{i_1-1}, b_{i_1})\right) f\left(X_1 \middle| {b_{i_1}}^*\right) \prod_{t=2}^{T} f\left(c_t \in (b_{i_t-1}, b_{i_t}) \middle| {b_{i_{t-1}}}^*\right) f\left(X_t \middle| {b_{i_t}}^*\right) =$$

$$\sum_{i_1=1}^{m} \dots \sum_{i_T=1}^{m} \left(P\left(c_1 \le b_{i_1}\right) - P\left(b_{i_1-1} \le c_1\right)\right) P\left(x_1 \middle| {b_{i_1}}^*\right)$$

$$\prod_{t=2}^{T} \left(P\left(c_t \le b_{i_t} \middle| b_{i_{t-1}}\right) - P\left(b_{i_t-1} \le c_t \middle| {b_{i_{t-1}}}^*\right)\right) f\left(X_t \middle| {b_{i_t}}^*\right) \qquad (7.2.1)$$

- The second approach is:

$$f(c_t|c_{t-1}) \approx f\left({b_{i_t}}^* \middle| {b_{i_{t-1}}}^*\right)$$

$$f(x_t|c_t) \approx f\left(x_t \middle| {b_{i_t}}^*\right)$$

where ${b_{i_t}}^*$ a representative point e.g. a mid-point of the B_{i_t} subinterval. [4]

Therefore, the corresponding integral can be written:

$$\int_{b_0}^{b_m} f(c_T|c_{T-1}) f(x_T|c_T) dc_T \approx h \sum_{i_T=1}^{m} f\left({b_{i_T}}^* \middle| {b_{i_{T-1}}}^*\right) f\left(x_T \middle| {b_{i_T}}^*\right)$$

and the likelihood has been approximated as follows:

$$L_T \approx h \sum_{i_1=1}^{m} \dots \sum_{i_T=1}^{m} f\left({b_{i_1}}^*\right) f\left(x_1 \middle| {b_{i_1}}^*\right) \prod_{t=2}^{T} f\left({b_{i_t}}^* \middle| {b_{i_{t-1}}}^*\right) f\left(x_t \middle| {b_{i_t}}^*\right)$$

Both approaches we can rewrite with the form of a matrix product:

$$L_T = \delta \Gamma P(x_1) \Gamma P(x_2) \Gamma P(x_3) \dots \Gamma P(x_T) 1'$$

As we have mentioned in Chapter 6 the likelihood of an HMM must contain three elements in order to be able to express it. The transition probability matrix Γ, the initial distribution δ and the diagonal matrix $P(x)$ which is defined as the diagonal matrix with ith diagonal element $p_i(x)$, *for* $i = 1, \dots, m$. [4] [5]

7.3 The transition probability matrix

We define the following transition probability matrix which contains an m-state homogenous Markov chain

$$\gamma_{ij} = P(C_t = j | C_{t-1} = i)$$

for $i = 1, \ldots, m$ and $j = 1, \ldots, m$ corresponds to the probability of moving from state i (at time $t-1$) to state j (at time t)

This matrix with transition probabilities from one state to another we can define it:

- For the first approach is

$$\gamma_{ij} = \Phi\left(\frac{b_j - \varphi b_i^*}{\sigma}\right) - \Phi\left(\frac{b_{j-1} - \varphi b_i^*}{\sigma}\right) = P(C_t \in B_j | C_{t-1} = b_i)$$

- For the second approach is

$$\gamma_{ij} = P(C_t \in (b_{j-1}, b_j) | C_{t-1} \in (b_{j-1}, b_j))$$
$$\gamma_{ij} = P(C_t = b_i^* | C_{t-1} = b_i^*)$$

where b_i^* a representative point of the interval $B_i = (b_{i-1}, b_i)$ and Φ denotes the cumulative distribution function of the standard normal distribution. Therefore, as we observe there corresponds to a piece of likelihood (Equation 7.2.1).

7.4 The initial distribution

Also, we define initial state distribution δ with elements $\delta_i = f(C_1 \in B_i)$, for $i = 1, \ldots, m$ and more specifically we can write as follows:

- For the first approach is

$$\delta = (f(c_1 \in (b_{1_1-1}, b_{i1})), \ldots, f(c_1 \in (b_{m_1-1}, b_{m_1})))$$

$$\delta = \left(P(c_1 \leq b_{1_1}) - P(b_{1_1-1} \leq c_1), \ldots, P(c_1 \leq b_{m_1}) - P(b_{m_1-1} \leq c_1)\right)$$

where f is the probability mass function and from Equation (7.3) $C_1 \sim Normal\ \left(0, \frac{\sigma^2}{1-\varphi^2}\right)$.

- For the second approach is

$$\delta = (hf(b_1^*), \ldots, hf(b_1^*))$$
$$\delta = (\frac{b_m - b_0}{m} f(b_1^*), \ldots, \frac{b_m - b_0}{m} f(b_1^*))$$

7.5 The diagonal matrix

Lastly, for both approaches we define a diagonal matrix $P(x_t)$ with *i*-th diagonal entry the normal density with mean 0 and variance $\beta^2 exp(b_i^*)$ and all diagonal elements as follows:

$$f_i(x_t) = f(X_t = x | C_t = b_i^*)$$

In conclusion, the approximating likelihood can then be written exactly as and constitutes the matrix form of the likelihood for HMMs:

$$L_T = \delta P(x_1) \Gamma P(x_2) \Gamma P(x_3) \ldots \Gamma P(x_T) 1'.$$

where $1'$ is a column vector of ones.

7.6 Fitting an SSM to injury data

7.6.1 Transition probability matrix calculation

As we have mentioned, the first thing we are interested in HMMs is to define the transition probability matrix. We have the same goal in SSMs models and more specifically, as we said C_t are random latent variables and transition probability matrix

Γ contains the transition probabilities of moving from one state at time $t-1$ to another at time t and we can represent it with the following equation:

$$\gamma_{ij} = P(C_t \in B_j | C_{t-1} \in B_i)$$

where B_i, B_j are the subintervals at i and j state, respectively. [4]

7.6.2 First Transition probability matrix approach

In this approach, we choose to replace C_{t-1} with a point estimation and transition probability matrix is represented as follow:

$$\gamma_{ij} = P(C_t \in (b_{j-1}, b_j) | C_{t-1} = b_i^*)$$

From probability theory we, for a continuous random variable with cumulative distribution $F(x)$, it is represented that $P(x \in [a,b)) = F(b) - F(a)$. Therefore, the transition probability matrix of approach is given by:

$$\begin{aligned} P\left(C_t \in \left(b_{j-1}, b_j\right) \middle| C_{t-1} = b_i^*\right) &= P\left(b_{j-1} < C_t \leq b_j \middle| C_{t-1} = b_i^*\right) \\ &= P\left(C_t \leq b_j \middle| C_{t-1} = b_i^*\right) - P\left(b_{j-1} < C_t \middle| C_{t-1} = b_i^*\right) \\ &= F_{C_t|C_{t-1}}(b_j; \mu_i, \sigma) - F_{C_t|C_{t-1}}(b_{j-1}; \mu_i, \sigma) \end{aligned}$$

where F is the cdf of $C_t|C_{t-1}$, i.e. $N(\varphi C_{t-1}, \sigma^2)$ and $\mu_\iota = \varphi b_i^*$.

The loglikelihood is calculated using the function *optim* and depends on the values m, b_0 and b_m. After several tests, we noticed that as we increase the range $[b_0, b_m]$ and the number of m-states the likelihood value decreases significantly. For this reason, we choose to select $m = 50$ and $b_m = 6$ in order to avoid two major problems that may occur. The first thing we need to note is that as m increases then the probabilities γ_{ij} get smaller and this can lead at least one estimated transition probability to become zero. So, it would be good not to choose a large number of states. Also, regarding the

range $[b_0, b_m]$ a good option of subintervals is not to be big enough because this will make the calculations more difficult and slower.

The parameter estimations we will see in *Table 7.1* for the injury data.

7.6.3 Parameter Estimation

We provide the parameter estimations for the selected $m = 50$ and $b_m = 6$

	SSM (1)	
Parameter	*Estimate*	*SE*
$\boldsymbol{\beta}$	*0.2374496*	*0.03116393*
$\boldsymbol{\varphi}$	*0.6580702*	*0.00651268*
$\boldsymbol{\sigma}$	*4.6090791*	*0.01766361*

Table 7.1. Parameter estimates and standard errors

Also, we provide and the following results:

$$Loglikelihood = -119.4765$$

$$AIC = 244.9529$$

$$BIC = 252.646$$

7.6.4 Second Transition probability matrix approach

As we have mentioned, the transition probability matrix Γ we can represent as follows and for the second approach we will follow the procedure below:

$$\gamma_{ij} = P(C_t \in (b_{j-1}, b_j) | C_{t-1} \in (b_{j-1}, b_j))$$

where $i = 1, \ldots, m$ and $j = 1, \ldots, m$

- In order to make the calculation easier and we estimate the matrix we can replace each subinterval by a representative point such as the midpoint and the transition probability matrix Γ will be written as shown below:

$$\gamma_{ij} = P(C_t = b_i^* | C_{t-1} = b_i^*)$$

- Also, with this approach the computation is faster and more specifically, in the case we choose subintervals with a small range.

Also, for this approach, the loglikelihood is calculated using the function *optim* and depends on the values m, b_0 and b_m. After several tests, we noticed that as we increase the range $[b_0, b_m]$ and the number of m-states the likelihood value decreases significantly. For this reason, we choose to select $m = 30$ and $b_m = 5.5$ .The results of the parameter estimations we will see in *Table 7.2* for the injury data.

7.6.5 Parameter Estimation

We provide the parameter estimations for the selected $m = 30$ and $b_m = 5.5$

	SSM (2)	
Parameter	*Estimate*	*SE*
$\boldsymbol{\beta}$	*-1.2477848*	*0.11969391*
$\boldsymbol{\varphi}$	*0.6970656*	*0.03906823*
$\boldsymbol{\sigma}$	*4.3129098*	*0.20347159*

Table 7.2: Parameter estimates and standard errors

Also, we provide and the following results:

$$Loglikelihood = -133.9979$$

$$AIC = 273.9959$$

$$BIC = 281.6889$$

The computation of the standard errors resulted from the Hessian matrix as follows:

$$H = -\left(\frac{\partial^2 l}{\partial\theta_i \partial\theta_j}\right)$$

where, $\theta = (\beta, \varphi, \sigma)$ are the estimated parameters of the two methods presented above and l is the maximum log-likelihood estimates. [21]

7.7 Simulation Study for State Space and Parameter-Driven models

After the theoretical analysis of the HMMs models with continuous state space we will proceed to a simulation experiment which is done between SSMs models and parameter-driven models. The initial parameters were set at $\varphi = 0.8$, $\sigma = 0.5$ and $\beta = 0.2$.

Simulation Results			
		Loglik	***AIC***
100	***ZIP***	*108.52*	*225.05*
	SSM	***104.80***	***215.61***
200	***ZIP***	*250.30*	*508.60*
	SSM	***238.87***	***483.75***
500	***ZIP***	*706.23*	*1420.46*
	SSM	***503.32***	***1012.64***

Table 7.3: Results from Loglikelihood and Akaike criterion for simulated data

As we can observe, from *Table 7.3* we present the results from Loglikelihood and Akaike criterion for simulated data under three different sample sizes $N = (100, 200, 500)$. It is obvious that in all cases concerning the sample size the loglikelihood and Akaike criterion are lower in the SSMs models in relation to the parameter-driven models. This finding can lead us to the fact that SSMs models describe better our data and are more appropriate in cases of data with many zeros.

Chapter 8

Model Selection

The main objective of this work is to select the appropriate model that correctly describes our data. For this reason, we dealt with and analyzed two model cases (from Section 3 and 4), which concern the observation-driven and parameter-driven models. Then we analyzed three options (from Section 5 and 6) where in Section 6 we selected the 3-state Poisson HMM and in Section 7 we implemented two approaches of a state space model with X_t, indicates the injuries of the hospital cleaners and follows a Poisson distribution with $\lambda = \beta exp(C_t)$ and the continuous states C_t correspond to an $AR(1)$ model. Table 8.1 present these five models and contains the log-likelihood, the number of estimated parameters and the Akaike and Bayesian information criterion (AIC and BIC).

- Observation-driven model: Zero inflated Poisson with $AR(1)$.
- Parameter-driven model: Zero inflated Poisson with $AR(1)$.
- 3-State HMM: 3-state Poisson-HMM model
- SSM1: State space model with $b_m = 5.5$ ($b_0 = -5.5$) and $m = 30$.

Computed Γ_{30x30} transition probability matrix using interval approximation of C_t and a point approximation for C_{t-1}.

- SSM2: State space model with $b_m = 5.5$ ($b_0 = -5.5$) and $m = 30$.

Computed Γ_{30x30} transition probability matrix with mid-point approximation of C_t and C_{t-1}.

Models				
Model	*n*	*Log-l*	*AIC*	*BIC*
Observation-driven	*5*	*-148.10*	*306.21*	*318.97*
Parameter-driven	*5*	*-147.54*	*305.08*	*317.90*
3-state HMM	*9*	*-151.00*	*320.01*	*343.09*
SSM1	***3***	***-119.47***	***244.95***	***252.64***
SSM2	*3*	*-133.99*	*273.99*	*281.68*

Table 8.1: Injury data: Comparison of Observation-driven and Parameter-driven model with hidden Markov model and two versions of a state-space model by AIC and BIC, where n denotes the number of estimated parameters

From the *Table 8.1* we can observe that the first approach of SSM1 has the smallest AIC and BIC and for this reason we could say that it is the most appropriate. Nevertheless, and the results from the second approach of SSMs (SSM2) are very close and they seem to have similar performance. In addition, regarding the three remaining models (Observation-driven, Parameter-driven and 3-state HMM), we notice that their results are very close and we can say that they also seem to make a good adjustment to the injury data. However, state space models provide additional information about the hidden states and the fit of these models requires a small number of estimated parameters ($n = 3$).

Chapter 9

Conclusions

Count time series with excess zeros are often encountered in many health applications and not only there. In this thesis, we first analyzed the injury data and tried to find the most appropriate model. We started with observation-driven and parameter driven models but also with Hidden Markov models with discrete state space. As shown by the values of the AIC and BIC these three models have similar behavior and it should be noted that among them, we probably choose the parameter-driven model due to the smaller number of estimated parameters ($n = 5$). Nevertheless, in the continuation of the analysis we proceed to the application of state space models where we aim to include as much heterogeneity among time periods as possible. We implemented a discretization method and we used two approaches which seem to work properly and one of them is the best option for describing our data since it presents the lowest value in AIC (*244.95)* and BIC (*252.64)* and has the smallest number of estimated parameters ($n = 3$). After the real examples, we chose to perform a simulation experiment for each method separately and for different sample sizes $N = (100, 200, 500)$. And for this case we can observe that from the AIC it seems that the Space State models make a better fit in our simulated data in different sample sizes. In conclusion, we can say that state space models worked better as we expected, and it make sense to perform all the statistical techniques and demanding algorithms in order to implement them.

References

[1] Zucchini, Walter, Iain L. MacDonald, and Roland Langrock. Hidden Markov models for time series: an introduction using R. Chapman and Hall/CRC, 2017.

[2] Cowen, L. L., Besbeas, P., Morgan, B. J., & Schwarz, C. J. (2017). Hidden Markov models for extended batch data. Biometrics, 73(4), 1321- 1331.

[3] Kyriaki Nik. Asimaki, Master Thesis Independent mixture and hidden Markov models for supermarket sales data. Athens University of Economics and Business, June 2018

[4] Leskou Eugenia, Master Thesis Non-linear state-space models for count time series. Athens University of Economics and Business, February 2020

[5] Langrock, Roland, Iain L. MacDonald, and Walter Zucchini. "Some nonstandard stochastic volatility models and their estimation using structured hidden Markov models." Journal of Empirical Finance 19.1 (2012): 147-161

[6] Ming Yang, Joseph E Cavanaugh and Gideon KD Zamba. "State-space models for count time series with excess zeros" 15(1), 70-90

[7] Besbeas, P., Morgan, B. J. (2019). "A general framework for modeling population abundance data"

[8] Fan Tang, Joseph E Cavanaugh. "State-space models for Binomial time series with excess zeros".

[9] Witowski, Vitali. "HMMpa-package: Analyzing Accelerometer Data Using Hidden Markov Models." (2018).

[10] ScienceDirect: Autocorrelation Function
https://www.sciencedirect.com/topics/chemistry/autocorrelation-function

[11] Statistical Analysis of Finite Mixture Distributions, Titterington, D.M.
and Titterington, P.S.D.M. and Smith, A.F.M. and Makov, U.E.,1985

[12] allBusiness: Seasonality
http://www.allbusiness.com/barrons dictionary/dictionary-seasonality-4946957-1.html

[13] Gubner, John A. (2006). Probability and Random Processes for Electrical and Computer Engineers. Cambridge University Press.

[14] Box, G. E. P.; Jenkins, G. M.; Reinsel, G. C. (2008). Time Series Analysis, Forecasting and Control (4th ed.). Hoboken, NJ: Wiley.

[15] B. Sericola, Markov Chains: Theory, Algorithms and Applications. London: ISTE Ltd and John Wiley & Sons Inc, 2013.

[16] Gagniuc, Paul A. (2017). Markov Chains: From Theory to Implementation and Experimentation. USA, NJ: John Wiley & Sons.

[17] ScienceDirect: Hidden Markov Models
https://www.sciencedirect.com/topics/computer-science/hidden-markov-models

[18] Sanjay Dorairaj: Hidden Markov Models Simpli_ed
https://medium.com/@postsanjay/hidden-markov-models-simplified-c3f58728caab

[19] Machine Learning Mastery: Probabilistic Model Selection with AIC, BIC, and MDL
https://machinelearningmastery.com/probabilistic-model-selection-measures/

[20] Ming Yang, Gideon Zamba, Joseph Cavanaugh: ZIM-package: Zero-Inflated Models for Count Time Series with Excess Zeros (2018)

[21] Aaron A. King and Edward L. Ionides: Standard errors via the observed Fisher information