Faculty of
Mathematics
and Informatics

VILNIUS UNIVERSITY
FACULTY OF MATHEMATICS AND INFORMATICS
MODELLING AND DATA ANALYSIS
MASTER'S STUDY PROGRAMME

MODELLING INTEGER-VALUED AUTOREGRESSIVE PROCESSES WITH APPLICATION TO COVID-19 DATA

Master's thesis

Author: Jūratė Dulkevičiūtė
VU email address: jurate.dulkeviciute@mif.stud.vu.lt
Supervisor: prof. habil. dr. Remigijus Leipus

Vilnius

2022

Sveikareikšmių autoregresijos procesų modeliavimas ir taikymas COVID-19 duomenims

Santrauka

Šiame darbe yra pristatomi sveikareikšmiai autoregresijos modeliai kartu su keliais modelio papildymais: sezoniniu INAR ir dvimačiu INAR. Modelių su Puasono, neigiamu binominiu ir apibendrintu Puasono skirstiniais vertinimui naudojami trys skirtingi metodai: momentų metodas, sąlyginis mažiausių kvadratų metodas ir sąlyginis didžiausio tikėtinumo metodas. Aprašyti vertinimo būdai yra patikrinami naudojant simuliuotus duomenis ir pritaikomi COVID-19 mirčių ir susirgimų Lietuvoje, Estijoje, Kroatijoje ir Izraelyje laiko eilutėms. Taip pat darbe pateikiamos trumpalaikės prognozės.

Raktiniai žodžiai: INAR, sezoninis INAR, dvimatis INAR, COVID-19, Puasono skirstinys, neigiamas binominis skirstinys, apibendrintas Puasono skirstinys.

Modelling integer-valued autoregressive processes with application to COVID-19 data

Abstract

This work introduces an integer-valued autoregressive model and its extensions: seasonal INAR, bivariate INAR. Three estimation methods: method of moments, conditional least squares and conditional maximum likelihood, are presented for models with Poisson, Negative Binomial and Generalized Poisson distributions. The described methods are tested on simulated data and applied on COVID-19 cases and deaths for Lithuania, Croatia, Estonia and Israel. Lastly, a short term forecast is presented.

Key words: INAR, Seasonal INAR, Bivariate INAR, COVID-19, Poisson distribution, Negative Binomial distribution, General Poisson distribution.

Content

1 Introduction

Coronavirus disease 2019 (COVID-19) is infectious pneumonia caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [3]. This disease for the first time was reported in December 2019 in Wuhan city, the capital of Hubei province in China [23]. Since then it rapidly spread across China and throughout the world causing many national lockdowns. As of 9th January 2022, over 305.5 million cases and more than 5.4 million deaths have been reported in 192 countries and regions [1]. The World Health Organization (WHO) has announced the COVID-19 outburst as a Public Health Emergency of International Concern and a pandemic.

In 2020, prevention and restrictions were the only options that were able to help control increasing numbers of cases and deaths. WHO estimated mortality rate which is around 3.4%, by comparison, seasonal flu kills less than 1%. COVID-19 became a huge challenge for people around the world. The pandemic provoked devastating consequences in the global financial, merchandise and job markets. Some estimations show economic growth will contract by 4.7% in Europe and Asia, even 7.2% in Latin America. Because of travel restrictions tourism sector shrunk dramatically, growth 2020/2019 in Asia was -86%, Europe -67%, according to the World Tourism Organization [5]. Exponential growing cases became a challenge for the health system, all hospital beds were occupied, doctors working overtime, lack of disposable masks or gloves. In order to control the situation is it important to understand the dynamics of the virus, detect the most vulnerable groups and take actions.

In the past few years, we can find plenty of models applied for epidemic data. The most commonly used models are Susceptible-Infectious-Removed (SIR) and Susceptible-Exposed-Infectious-Removed (SEIR). These models are constructed by the system of ordinary differential equations that simulate the spread of the virus, which is extremely useful for making decisions [14]. From the theoretical point of view, it is known the growth of infectious diseases spread is exponential. Naturally, a lot of researchers use a generalized-growth model in order to capture exponential or sub-exponential growth [21], [10]. Another intuitive model applied for this type of data is the Poisson autoregressive model, which assumes distribution of new cases at some time t, conditional on the information $t-1$, is Poisson distribution [7].

An epidemic data, when looking at numbers of deaths or infected persons, are recorded as a series of counts. However, the typical models (e.g. ARMA) are improper for such data because they assume continuous distributions. For this reason, a lot of studies focuses on binomial thinning operator based integer-valued autoregressive model (INAR) [8]. The majority of articles focus on Poisson distribution, which means the mean and variance of innovations are equal, however, this is rarely the case. Real-life data usually are overdispersed, for this case distributions like Negative Binomial or Generalized Poisson should be proper to use. INAR type models do have some extensions, like adding seasonality [9] or constructing models for data with bivariate structure [18].

In this thesis, count data models are applied for the number of deaths and number of infections caused by a coronavirus in Lithuania, Estonia, Croatia and Israel. We present properties for INAR, seasonal INAR and bivariate INAR models, derive how to estimate models' coefficients with the method of moments (MM), conditional least squares (CLS) and conditional maximum likelihood (CML) while using Poisson, Negative Binomial and Generalized Poisson distributions for innovations. Lastly, a short term forecast is performed.

2 Literature review

2.1 SARS-CoV-2

Coronavirus 2 (SARS-CoV-2) is a highly transmissible and pathogenic coronavirus, which increased a huge threat to human health and public safety. In general, coronaviruses are a diverse group of viruses that infect a wide range of species, as well as humans, and can cause mild to serious respiratory infections [13]. Fever, dry cough, weakness, and, in serious cases, dyspnea are common symptoms of COVID-19. Many infections are asymptomatic, particularly in children and young adults, while older people and/or people with co-morbidities are more likely to develop serious illness, respiratory failure, and death. In a study of 72,314 cases in China, 81% were classified as moderate, 14% as extreme cases requiring intensive care unit (ICU) ventilation, and 5% were classified as serious (meaning the patients had respiratory failure, septic shock, and/or multiple organ dysfunction or failure) [23].

Since the virus is highly transmissible Centres for Disease Control and Prevention (CDC) [4] defines three principal ways COVID-19 can spread:

1. Inhalation of air carrying very small fine droplets and aerosol particles that contain COVID-19 virus, the risk of transmission is highest within three to six feet (approximate one to two meters) of an infectious source.

2. Deposition of virus carried in exhaled droplets and particles onto exposed mucous membranes (i.e., “splashes and sprays”, such as being coughed on).

3. Touching surfaces that have been contaminated by the virus when touching their eyes, nose or mouth without cleaning their hands.

Knowing how the virus spread, some precautions were proposed and applied: sanitizing public surfaces, such as door handles, public transportation stops, buses, trains; social distancing; wearing masks in public places and of course mandatory isolation in cases of contact with an infected person.

Epidemics, as well as COVID-19 pandemic, can be decomposed into three phases [14]:

1. The first phase of the epidemic is indicated by a linear increase in the number of reported cases, with the number of daily cases being nearly constant from day to day.

2. The second phase of the epidemic is an exponentially growing phase, in which the number of infected increases exponentially but the transmission rate remains constant.

3. Due to massive public interventions and social distancing efforts, the third phase of the epidemic correlates to a time-dependent exponentially decreasing transmission rate.

2.2 Modelling epidemics

Time series analysis is widely used to analyse and forecast different contagious diseases, such as Ebola, influenza pandemic, smallpox and many more. In this section we present an overview of models specifically used for dealing with epidemic data.

2.2.1 Generalized-growth model

Theoretically, the growth of infectious diseases spread is exponential due to the sufficiency of resources. Using this assumption we can describe cumulative number of cases by equation $C(t) = C(0)e^{rt}$, where r is the growth rate, t - time, $C(0)$ - number of cases at the start of outbreak. However not always exponential growth is the best option, in order to relax this assumption about exponential growth - simple generalized model can be used: $dC(t)/dt = rC(t)^p$, here p - deceleration growth parameter. The generalized-growth model capture sub-exponential pattern (when $0 < p < 1$), lower p value leads to slower growth, and exponential epidemic growth (when $p = 1$). Identifying growth dynamics and patterns of disease helps to understand the specifics of contagious disease transmission. Article [21] analyse epidemic growth patterns of different infectious disease, such as measles, smallpox, Ebola by estimating parameters p and r jointly.

2.2.2 Integer value autoregression

Since infected population and casualties are counted variables, the popular model to use is integer-valued autoregressive (INAR). The multiplication operator in AR-type processes is replaced with the thinning operator, which ensures the integer discreteness of the process. In terms of innovations, mostly are used Poisson distribution and Negative Binomial distribution.

INAR with an oscillating Weighted Cosine Geometric innovation term was applied for a few small developing states: Singapore, Cape Verde, Bahrain, Maldives and Mauritius [15]. Analysis of this paper focuses on COVID-19 cases in countries mentioned before. The authors performed the Ljung-Box test which confirmed data is serially correlated, moreover, it was observed significant over-dispersion and that the time pattern of COVID-19 data is oscillating. For this issue, it was selected to use the Weighted Cosine Geometric distribution, which has been proven to mimic such patterns nicely in discrete data. Additionally, in the model were included five covariates: population, GSI (Government Stringency Index, which is computed from nine indicators corresponding to sanitary measures, such as school closures, cancellation of public events), monthly temperature, air quality and a transmission mode. The results showed GSI and transmission rate are highly significant, showing that local factors are extremely important and health authorities should control it.

INAR models can be extended by including seasonality, this type of model allows the autoregression parameter to vary with season. Seasonal INAR can be convenient to use for modelling diseases, that have a higher incidence rate at a certain time of the year. Article [17] analyses arrivals to hospital with influenza symptoms in Spain. Five years data shows seasonality every 12 months, looking like a sinusoidal wave. Authors build and applied INAR(1) and INAR(2) with innovations that follow Poisson distribution with different means $\lambda_t = \lambda_{t+\tau}$, here τ represents observed seasonality.

Another article [20] analyses monthly polio incidences in the United States of America. The author offers to use seasonal geometric distribution and negative binomial thinning operator, SGINAR(12), instead of the popular Poisson distribution and binomial thinning operator. This model performed better, compared with INAR(1), INAR$(1)_{12}$ and NGINAR(1) (geometric INAR model with no seasonality), based on AIC and BIC criteria. The article showed that SGINAR was a quite good model for overdispersed seasonal data, however, models did not capture all fluctuations very well.

Overall, INAR-type models might help to capture seasonality, also it is flexible, therefore researchers can use a different distribution, based on with what data they are working.

2.2.3 SIR type models

One of the most widely used models for the simulation of the spread of viruses is Susceptible-Infectious-Removed (SIR) and Susceptible-Exposed-Infectious-Removed (SEIR) models. Initially, the SEIR model was used to simulate the spread of flu, but it is easy to apply in other cases, more specific when individuals experience a long incubation duration, such that the individual is infected but not yet infectious. This model represents the course of the disease: Susceptible - Exposed - Infectious - Recovered, if recovery does not confer lifelong immunity SEIRS (Susceptible - Exposed - Infectious - Recovered - Susceptible) model may be used, such diseases may be rotavirus or malaria.

Article [12] introduces SEIR model constructed from seven differential equations, it includes susceptible, exposed, infectious without intervention, infectious with intervention, recovered, quarantined and hospitalized classes and it involves external input from the foreign countries. The main hypothesis of this model is that all individuals will move from one class to another with some probabilities and/or parameters. Some of the parameters can be fixed, such as temporary immunity rate while others have to be estimated or calculated based on historical data. Article analyses parameters by dividing Hubei province, China, data into two stages: outbreak and inhibition. It was proved that the dynamics of the proposed model is different with different sets of parameters.

Authors of the another article [14] analyse cases in South Korea, Italy, France and Germany. The purpose is to predict the cumulative number of reported and unreported cases. As mentioned earlier this type of model requires some fixed parameters, however, true parameters are unknown. Authors assumed some fraction f of total cases are unreported, the chosen f were

0.6 and 0.1, meaning 40% or 90% of symptomatic infectious cases are not reported. Another assumption is that the typical infectious time for symptomatic infected persons is 7 days. During the analysis it was observed that in the European countries the pandemic situation wass getting worse, cases are increasing, while the situation in South Korea is stable because important measures were implemented early. By comparing the situation in European countries and South Korea, it can be seen that government intervention is crucial, measures should start as early as possible, and should be strong.

2.2.4 Poisson autoregressive model

Poisson autoregressive model also is popular for modelling infectious diseases. In the article [7], log-linear version of Poisson autoregression is used for understanding the spread of the COVID-19 virus. It is assumed that the distribution of new cases at some time t, conditional on the information t-1, is Poisson distribution. Moreover, short-term and long-term dependence were included in the model. Log-linear intensity specification allows for negative dependency, opposite that linear. Data used in the article covers the time period from January 20 to March 8, 2020, and represents infection in China, Iran, South Korea and Italy. China during this period has a complete cycle: increasing trend, a peak and decreasing trend, meanwhile, other countries looks like are still increasing. After calculations, significance for both short and long term dependencies was confirmed. Such type of model helps to identify the stage of the contagion cycle.

3 Integer value autoregression models

In this chapter we look at how inter-valued autoregressive models are constructed and estimated. We use three methods for parameters estimation as well as use three different distributions.

3.1 Multiplication Problem and Thinning Operator

Straight forward application of simple autoregressive models to count time series is not feasible. AR(1) recursion $X_t = \alpha \cdot X_{t-1} + \varepsilon_t$ cannot be applied, even if innovations ε_t are non-negative integers. Multiplication $\alpha\cdot$ does not preserve the discrete range, this issue is called multiplication problem. A few AR(1)-like models for count processes were introduced by McKenzie in the 1985 paper 'Some simple models for discrete variate time series'. The author proposed new mechanisms for reducing X_{t-1}, one of them - binomial thinning.

Definition 1. Let ξ_j be counting sequence of independent and identically distributed Bernoulli random variables with mean $\alpha \in [0,1]$ and X a non-negative integer-valued random variable with a range $\{0, 1, ..., n\}$, independent of the counting sequence. The binomial thinning operator $\alpha\circ$ is defined by

$$\alpha \circ X \coloneqq \begin{cases} \sum_{j=1}^{X} \xi_j(\alpha) & X \geq 0, \\ 0 & X = 0. \end{cases}$$

Interpretation of binomial thinning is very intuitive. First, lets suppose the X_{t-1} represents population at time $t-1$. At the next time step t, the population may have shrunk due to individuals dying. If we can assume that each individual survives independently of each other with the probability α of a individual surviving from time $t-1$ to t, then the population at time t can be given as $X_t = \alpha \circ X_{t-1}$, where $\alpha \circ X_{t-1}$ is the number of survivors from $t-1$. See also Weiß [22].

Some important properties of binomial thinning operator:

1. $0 \circ X = 0$;
2. $1 \circ X = X$;
3. $\alpha_1 \circ \alpha_2 \circ X \overset{d}{=} \alpha_2 \circ \alpha_1 \circ X$;
4. $\alpha_1 \circ (X + Y) \overset{d}{=} \alpha_1 \circ X + \alpha_1 \circ Y$;
5. $\alpha_1 \circ (\alpha_2 \circ X) \overset{d}{=} (\alpha_1\alpha_2) \circ X$;
6. $\alpha_1 \circ X + \alpha_2 \circ X \overset{d}{\neq} (\alpha_1 + \alpha_2) \circ X$;

7. Let $X \coloneqq \alpha \circ Y$. The mean value and the variance of X are given
$$\mathbb{E}(X) = \alpha E(Y) \text{ and } \mathbb{V}ar(X) = \alpha^2 \mathbb{V}ar(Y) + \alpha(1-\alpha)\mathbb{E}(Y).$$

Multiple modifications to the binomial thinning operator have been developed in order to make integer-valued thinning models more flexible for practical uses. For instance, some of them allow the dependence between the indicators of counting series (ξ_j), this type of operator is called generalized thinning. Another one - signed binomial thinning operator can handle over-dispersed and non-stationary integer-valued time series, moreover, it can be used for negative count data, which is an advantage, since binomial thinning can be applied only for non-negative time series.

3.2 INAR

In this section we define the integer-values autoregressive process of order one, denote INAR(1), this model was first introduced by McKenzie [16] and Al-Osh & Alzaid [8] in 1990's.

Definition 2. Let $(\varepsilon_t)_{\mathbb{N}}$ be the innovations consisting of i.i.d random variables with $\mathbb{E}(\varepsilon_t) = \mu_\varepsilon$ and variance $\mathbb{V}ar(\varepsilon_t) = \sigma_\varepsilon^2$. Let $\alpha \in (0,1)$. A process $(X_t)_{\mathbb{N}}$ of observations, which follows the recursion
$$X_t = \alpha \circ X_{t-1} + \varepsilon_t \tag{1}$$
is said to be an INAR(1) process if all thinning operations are performed independently the each other and of $(\varepsilon_t)_{\mathbb{N}}$ and if the thinning operations at each time t as well as ε_t are independent of $(X_s)_{s<t}$

We should stress, that INAR(1) process described in Definition 2 and based on binomial thinning operator is a stationary process. Below we present and prove some of the most important properties of INAR(1) process.

Properties of the INAR(1) based on binomial thinning operator:

1. $\mathbb{E}(X_t) = \dfrac{\mu_\varepsilon}{1-\alpha}$;

2. $\mathbb{V}ar(X_t) = \dfrac{\alpha\mu_\varepsilon + \sigma_\varepsilon^2}{1-\alpha^2}$;

3. $\mathbb{E}(X_t|X_{t-1}) = \alpha X_{t-1} + \mu_\varepsilon$;

4. $\mathbb{V}ar(X_t|X_{t-1}) = \alpha(1-\alpha)X_{t-1} + \sigma_\varepsilon^2$;

5. $\mathbb{C}ov(X_t, X_{t+i}) = \alpha^i \mathbb{V}ar(X_t)$;

6. $\mathbb{C}orr(X_t, X_{t+i}) = \alpha^i$.

Proof:

1. We have

$$\begin{aligned}
\mathbb{E}(X_t) &= \mathbb{E}(\alpha \circ X_{t-1} + \varepsilon_t) = \mathbb{E}(\alpha \circ (\alpha \circ X_{t-1} + \varepsilon_{t-1}) + \varepsilon_t) \\
&= \mathbb{E}(\alpha^2 \circ X_{t-1} + \alpha \circ \varepsilon_{t-1} + \varepsilon_t) = \ldots = \mathbb{E}(\sum_{i=0}^{\infty} \alpha^i \circ \varepsilon_{t-i}) \\
&= \sum_{i=0}^{\infty} \alpha^i \mathbb{E}(\varepsilon_{t-i}) = \sum_{i=0}^{\infty} \alpha^i \mu_\varepsilon = \frac{\mu_\varepsilon}{1-\alpha}.
\end{aligned}$$

Here, for the first equality we use the definition of INAR(1) model (1). Using recursion we get infinite sum of $\alpha \circ \varepsilon$, since the mean of innovations is μ_ε, by using an infinite geometric series formula we get the result.

2. We have

$$\begin{aligned}
\mathbb{V}ar(X_t) &= \mathbb{V}ar(\sum_{i=0}^{\infty} \alpha^i \circ \varepsilon_{t-i}) = \sum_{i=0}^{\infty} \mathbb{V}ar(\alpha^i \circ \varepsilon_{t-i}) \\
&= \sum_{i=0}^{\infty} (\alpha^{2i} \mathbb{V}ar(\varepsilon_{t-i}) + \alpha^i (1-\alpha^i) \mathbb{E}(\varepsilon_{t-i})) \\
&= \sum_{i=0}^{\infty} (\alpha^{2i} \sigma_\varepsilon^2 + \alpha^i (1-\alpha^i) \mu_\varepsilon) \\
&= \frac{\sigma_\varepsilon^2}{1-\alpha^2} + \frac{\mu_\varepsilon}{1-\alpha} - \frac{\mu_\varepsilon}{1-\alpha^2} = \frac{\sigma_\varepsilon^2 + \mu_\varepsilon + \alpha\mu_\varepsilon - \mu_\varepsilon}{1-\alpha^2} \\
&= \frac{\sigma_\varepsilon^2 + \alpha\mu_\varepsilon}{1-\alpha^2}.
\end{aligned}$$

Here, for the first equality we again use recursion. The second equality is true, because we know ε_{t-i} are i.i.d. For the third we use binomial thinning seventh property.

3. We have

$$\begin{aligned}
\mathbb{E}(X_t | X_{t-1}) &= \mathbb{E}(\alpha \circ X_{t-1} + \varepsilon_t | X_{t-1}) = \mathbb{E}(\alpha \circ X_{t-1} | X_{t-1}) + \mathbb{E}(\varepsilon_t | X_{t-1}) \\
&= \alpha \mathbb{E}(X_{t-1} | X_{t-1}) + \mathbb{E}(\varepsilon_t) = \alpha X_{t-1} + \mu_\varepsilon.
\end{aligned}$$

Here we use the independence between ε_t and X_{t-1}.

4. We have

$$\begin{aligned}
\mathbb{V}ar(X_t | X_{t-1}) &= \mathbb{V}ar(\alpha \circ X_{t-1} + \varepsilon_t | X_{t-1}) = \mathbb{V}ar(\alpha \circ X_{t-1} | X_{t-1}) + \mathbb{V}ar(\varepsilon_t | X_{t-1}) \\
&= \mathbb{V}ar(\alpha \circ X_{t-1} | X_{t-1}) + \sigma_\varepsilon^2 = \alpha(1-\alpha) X_{t-1} + \sigma_\varepsilon^2.
\end{aligned}$$

Here we use the seventh binomial thinning property and use the independence between ε_t and X_{t-1}.

5. We have

$$\begin{aligned}\mathbb{C}ov(X_t, X_{t+h}) &= \mathbb{C}ov(X_t, \alpha^h \circ X_t + \sum_{i=0}^{h-1} \alpha^k \circ \varepsilon_{t+h-i}) \\ &= \mathbb{C}ov(X_t, \alpha^h \circ X_t) + \mathbb{C}ov(X_t, \sum_{i=0}^{h-1} \alpha^k \circ \varepsilon_{t+h-i}) \\ &= \alpha^h \mathbb{C}ov(X_t, X_t) = \alpha^h \mathbb{V}ar(X_t).\end{aligned}$$

Here we use the fact that ε_t are i.i.d in t and $t + h - k > t$ for $k < h$.

6. We have

$$\mathbb{C}orr(X_t, X_{t+i}) = \frac{\mathbb{C}ov(X_t, X_{t+j})}{\sqrt{\mathbb{V}ar(X_t)\mathbb{V}ar(X_{t-1})}} = \frac{\alpha^i \sigma_X^2}{\sqrt{\sigma_X^4}} = \alpha^i.$$

Having mean and variance of INAR(1) process, we can obtain mean and variance for innovations:

$$\mu_\varepsilon = \mu(1-\alpha), \quad \sigma_\varepsilon^2 = \sigma^2(1-\alpha^2) - \alpha\mu_\varepsilon. \tag{2}$$

The most common distribution used for INAR processes is Poisson, however, it is rarely a case that data has the same mean and variance. Calculation of the dispersion index and zero index helps to identify if the Poisson distribution is appropriate. Equidispersion property is described the following:

$$I := I(\mu, \sigma^2) := \frac{\sigma^2}{\mu} \quad \in (0; \infty). \tag{3}$$

$I = 1$ shows that mean and variance are equal, which satisfies the Poisson distribution definition. In case $I > 1$ we have overdispersed distribution, for example, a better choice would be Negative Binomial distribution, and $I < 1$ shows underdispersion.

Zero index is another characterization that can help to choose what distribution to use, it is obtained using the following:

$$I_{zero} := I_{zero}(\mu, p_0) := 1 + \frac{\ln p_0}{\mu} \quad \in (-\infty; 1). \tag{4}$$

Here p_0 is the probability of observing a zero $p_0 := P(X = 0) = \exp(-E(X))$. $I_{zero} > 0$ indicates zero inflation - the excess of zeros compared to Poisson distribution and in case $I_{zero} < 0$

we are observing zero deflation in the data.

One of the key applications of the model for the observed INAR(1) process is to estimate future outcomes of the process. Having observed $X_1, X_2, ..., X_t$, we wish to forecast X_{t+h} for some $h \geq 1$. The conditional mean is the most popular type of point forecast for real-valued processes, as it is known to be optimal in terms of mean squared error. The h-step-ahead conditional mean is given by

$$\begin{aligned}
\mathbb{E}(X_{t+h}|\mathcal{F}_t) &= \mathbb{E}(\alpha \circ X_{t+h-1} + \varepsilon_{t+h}|\mathcal{F}_t) \\
&= \mathbb{E}(\alpha \circ (\alpha \circ X_{t+h-2} + \varepsilon_{t+h-1}) + \varepsilon_{t+h}|\mathcal{F}_t) \\
&\overset{d}{=} \mathbb{E}(\alpha^2 \circ X_{t+h-2} + \alpha \circ \varepsilon_{t+h-1} + \varepsilon_{t+h}|\mathcal{F}_t) \\
&\overset{d}{=} ... \\
&\overset{d}{=} \mathbb{E}(\alpha^h \circ X_t + \sum_{j=0}^{h-1} \alpha^j \circ \varepsilon_{t+h-j}|\mathcal{F}_t) \\
&\overset{d}{=} \mathbb{E}(\alpha^h \circ X_t|\mathcal{F}_t) + \mathbb{E}(\sum_{j=0}^{h-1} \alpha^j \circ \varepsilon_{t+h-j}|\mathcal{F}_t) \\
&\overset{d}{=} \alpha^h X_t + \mu_\varepsilon \frac{1-\alpha^h}{1-\alpha}.
\end{aligned} \tag{5}$$

We can note, that due to the Markov property, the conditional mean depends only on X_t and not on previous observations. The biggest disadvantage of this type forecast is that it almost always returns a non-integer result, while X_{t+h} will certainly be an integer value from $\mathbb{N}_0$. Another way for forecasting is to use the conditional median instead of conditional expectation, however, in most cases, the conditional median can be difficult to compute.

3.2.1 Estimation of parameters

INAR(1) model is defined by the thinning parameter α and parameters specifying the marginal distribution of innovations. Having data $X_1, ..., X_t$ the goal is to estimate these parameters. In this section we consider the estimation of the unknown parameters based on three methods: MM, CLS and CML and analyse three different distributions: Poisson distribution which is the most popular for the count data, the Negative Binomial and Generalized Poisson.

3.2.1.1 Poisson distribution

Assume we have $\varepsilon_t \sim Pois(\lambda)$ with probability mass function

$$\mathbb{P}(\varepsilon_t = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, ... \tag{6}$$

In Poisson distribution case it is known that mean and variance is equal: $\mathbb{E}(\varepsilon_t) = \mathbb{V}ar(\varepsilon_t) = \lambda$.

Method of moments

The most simple method for estimating parameters is method of moments. The goal is to choose appropriate moment relations so that real model parameters could be found by solving system of equations. The MM estimates are found by replacing true parameters into corresponding sample moments.

For Poisson INAR(1) we have two unknown parameters α and λ. MM estimator for α is calculated as $\hat{\alpha}^{MM} := \hat{\rho}(1) := \hat{\gamma}(1)/\hat{\gamma}(0)$, here $\hat{\gamma}(k) = \frac{1}{n}\sum_{t=k+1}^{n}(X_t - \bar{X})(X_{t-k} - \bar{X})$ for $k \in \mathbb{N}_0$. For the mean of innovation we have $\hat{\mu}^{MM} := \bar{X}$ and by applying (2) $\hat{\lambda}^{MM} := \bar{X}(1 - \hat{\alpha}^{MM})$.

Conditional least squares

Another approach for estimating parameters is to accumulate the squared deviations among X_t and $\mathbb{E}(X_t|\mathcal{F}_{t-1})$ and to select parameters so that the conditional sum of squares is minimized [8]. The CLS estimator $\hat{\boldsymbol{\theta}}^{CLS} = (\hat{\alpha}^{CLS}, \hat{\lambda}^{CLS})^T$ of $\boldsymbol{\theta} = (\alpha, \lambda)^T$ is given by

$$\begin{aligned} Q(\alpha, \lambda) := \underset{\boldsymbol{\theta}}{\operatorname{argmin}}(\sum_{t=2}^{n}[X_t - \mathbb{E}(X_t|\mathcal{F}_{t-1})])^2 &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}(\sum_{t=2}^{n}[X_t - \mathbb{E}(\alpha \circ X_{t-1} + \varepsilon_t|\mathcal{F}_{t-1})])^2 \\ = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}(\sum_{t=2}^{n}[X_t - \alpha X_{t-1} - \lambda])^2. \end{aligned}$$

Taking the partial derivatives of $Q(\alpha, \lambda)$ and equating them to 0 we get the following system of equations:

$$\begin{cases} \frac{\partial Q}{\partial \alpha} &= -2\sum_{t=2}^{n}(X_t - \alpha X_{t-1} - \lambda)X_{t-1} = 0, \\ \frac{\partial Q}{\partial \lambda} &= -2\sum_{t=2}^{n}(X_t - \alpha X_{t-1} - \lambda) = 0. \end{cases} \tag{7}$$

From the (7) we get the following equations:

$$\sum_{t=2}^{n} X_t X_{t-1} - \alpha \sum_{t=2}^{n} X_{t-1}^2 - \lambda \sum_{t=2}^{n} X_{t-1} = 0 \tag{8}$$

and

$$\lambda = \frac{\sum_{t=2}^{n} X_t - \alpha \sum_{t=2}^{n} X_{t-1}}{n-1}. \tag{9}$$

Now we subside (9) in (8) and multiply by $(n-1)$:

$$\begin{aligned} 0 &= (n-1)\sum_{t=2}^{n} X_t X_{t-1} - \alpha(n-1)\sum_{t=2}^{n} X_{t-1}^2 - \sum_{t=2}^{n} X_t \sum_{t=2}^{n} X_{t-1} + \alpha(\sum_{t=2}^{n} X_{t-1})^2 \\ &= (n-1)\sum_{t=2}^{n} X_t X_{t-1} - \sum_{t=2}^{n} X_t \sum_{t=2}^{n} X_{t-1} + \alpha((\sum_{t=2}^{n} X_{t-1})^2 - (n-1)\sum_{t=2}^{n} X_{t-1}^2). \end{aligned} \tag{10}$$

From (10) we can express α, which leads to formulas for $\hat{\alpha}^{CLS}$ and $\hat{\lambda}^{CLS}$:

$$\hat{\alpha}^{CLS} := \frac{(n-1)\sum_{t=2}^{n} X_t X_{t-1} - \sum_{t=2}^{n} X_t \sum_{t=2}^{n} X_{t-1}}{(n-1)\sum_{t=2}^{n} X_{t-1}^2 - (\sum_{t=2}^{n} X_{t-1})^2}, \tag{11}$$

$$\hat{\lambda}^{CLS} := \frac{1}{n-1}(\sum_{t=2}^{n} X_t - \hat{\alpha}^{CLS} \sum_{t=2}^{n} X_{t-1}). \tag{12}$$

The CLS estimates in (11) and (12) are strongly consistent and asymptotically normally distributed (see Weiß [22]).

Conditional maximum likelihood

The maximum likelihood approach is similar to MM, it tries to select parameter values so that the sample which is observed becomes the most plausible.

By fixing the first observation X_1, we avoid estimating $p_{X_1}(\boldsymbol{\theta})$ and get the conditional log-likelihood function which is used to get estimates of $\boldsymbol{\theta} = (\alpha, \lambda)^T$:

$$\hat{\boldsymbol{\theta}}^{CML} = \underset{\boldsymbol{\theta}}{\text{argmax}} \sum_{t=2}^{n} \text{log} p_{x_t|x_{t-1}}(\boldsymbol{\theta}),$$

where

$$p_{x_t|x_{t-1}}(\boldsymbol{\theta}) := P(X_t = x_t|X_{t-1} = x_{t-1}) = \sum_{k=0}^{\min(x_t,x_{t-1})} \binom{x_{t-1}}{k} \alpha^k (1-\alpha)^{x_{t-1}-k} \frac{e^{-\lambda}\lambda^{x_t-k}}{(x_t-k)!}.$$

The CML estimator, as described above, is consistent and asymptotically normal (see Freeland and Mccabe [11]).

Simulation

In order to test if constructed estimates are plausible, simulations were performed. MM, CLS and CML methods were applied for the same simulated data sets. It was chosen to use two samples with different lengths, to see how methods perform, first set length $N = 200$ and the second $N = 500$. 200 independent experiments performed for both data sets, parameters were foxed for both simulation sets $\alpha = 0.8$ and $\lambda = 3$.

To compare the simulation results we calculate *Bias* (13) and Root Mean Square Error (RMSE) (14):

$$Bias = \frac{1}{N}\sum_{t=1}^{N}(\hat{\theta}_t - \theta_t), \tag{13}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{\theta}_t - \theta_t)^2}. \tag{14}$$

It is worth stressing, to find the CML estimations some initial parameters must be provided. In this case, CLS estimates were used as the primary values. The CML approach is used to locate global extremes, choosing unreasonable values for the parameters may result in finding local rather than global extremes, resulting in biased results. From the simulation results Table 1 we can notice, that in both cases $N = 200$ and $N = 500$ MM and CLS estimates are very similar. But looking at *Bias* and RMSE of CML estimator we can see the better performance than MM and CLS for both sets of simulations.

		MM estimation		CLS estimation		CML estimation	
Size	Real parameter	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
N=200	$\alpha = 0.8$	0.019	0.004	0.015	0.003	0.0014	0.0015
	$\lambda = 3$	-0.307	0.054	-0.247	0.052	-0.034	0.023
N=500	$\alpha = 0.8$	0.004	0.0008	0.0036	0.0008	0.0007	0.0003
	$\lambda = 3$	-0.067	0.0133	-0.0566	0.0131	-0.0122	0.005

Table 1: Simulation results of Poisson INAR(1)

Forecasting

From equation (5) we can derive h-step-ahead forecast formula for INAR(1) with Poisson distribution:

$$\mathbb{E}(X_{t+h}|\mathcal{F}_t) \overset{d}{=} \alpha^h X_t + \lambda \frac{1-\alpha^h}{1-\alpha}. \tag{15}$$

3.2.1.2 Negative Binomial Distribution

Lets say we have $\varepsilon_t \sim NB(r,p)$ with probability mass function

$$\mathbb{P}(\varepsilon_t = k) = \binom{k+r-1}{r-1} \cdot (1-p)^k p^r, \quad k = 0, 1, ... \tag{16}$$

where r is the number of successes, k is the number of failures, and p is the probability of success. Mean and variance is known $\mathbb{E}(\varepsilon_t) = \frac{pr}{1-p}$, $\mathbb{V}ar(\varepsilon_t) = \frac{pr}{(1-p)^2}$.

Method of moments

First, lets consider method of moments estimators for unknown parameters α, p and θ. Having the first two moments, mean and variance, it is easy to derive estimations for parameters r and p. It can be done just by expressing one from another, and because $\alpha = \rho_1$ we have estimators:

$$\hat{\alpha}^{MM} = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)}, \quad \hat{p}^{MM} = 1 - \frac{\mathbb{E}(\varepsilon_t)}{\mathbb{V}ar(\varepsilon_t)}, \quad \hat{r}^{MM} = \frac{\mathbb{E}(\varepsilon_t)^2}{\mathbb{V}ar(\varepsilon_t) - \mathbb{E}(\varepsilon_t)}. \tag{17}$$

Conditional least squares

Parameters α, p and r also can be estimated with CLS, since now we need to estimate three parameters instead of two, we will use two-step conditional least squares method [19]. The first step is to find estimates for α and μ_ε, they are obtained by minimizing the following expression:

$$\sum_{t=2}^{n} [X_t - E(X_t|\mathcal{F}_{t-1})]^2 = \sum_{t=2}^{n} [X_t - \alpha X_{t-1} - \mu_\varepsilon]^2. \tag{18}$$

Analogical as in the Poisson distribution case, we took partial derivatives and express α and μ_ε through each other leading to equations:

$$\hat{\alpha}^{CLS} := \frac{(n-1)\sum_{t=2}^{n} X_t X_{t-1} - \sum_{t=2}^{n} X_t \sum_{t=2}^{n} X_{t-1}}{(n-1)\sum_{t=2}^{n} X_{t-1}^2 - (\sum_{t=2}^{n} X_{t-1})^2}, \tag{19}$$

$$\hat{\mu_\varepsilon}^{CLS} := \frac{1}{n-1}(\sum_{t=2}^{n} X_t - \hat{\alpha}^{CLS} \sum_{t=2}^{n} X_{t-1}). \tag{20}$$

For the second step it is necessary to find estimation for σ_ε^2. Lets define a new random variable $V_n = (X_t - \mathbb{E}(X_t|\mathcal{F}_{t-1}))^2 = (X_t - \alpha X_{t-1} - \mu_\varepsilon)^2$, and more $\mathbb{E}(V_t|\mathcal{F}_{t-1}) = \mathbb{V}ar(X_t|\mathcal{F}_{t-1}) = \mathbb{V}ar(\alpha \circ X_t + \varepsilon_t|\mathcal{F}_{t-1}) = \alpha(1-\alpha)X_{t-1} + \sigma^2$.

The CLS estimator for σ_ε^2 can be found by minimizing the sum of squares:

$$S_n(\sigma_\varepsilon^2) = \sum_{t=2}^{n}(V_t - \mathbb{E}(V_t|\mathcal{F}_{t-1})^2 = \sum_{t=2}^{n}(V_n - \alpha(1-\alpha)X_{t-1} - \sigma_\varepsilon^2)^2.$$

Taking partial derivative and equating to zero:

$$\frac{\partial S_n}{\partial \sigma_\varepsilon^2} = -2\sum_{t=2}^{n}(V_t - \alpha(1-\alpha)X_{t-1} - \sigma_\varepsilon^2) = 0$$

Now we can find estimate of the parameter σ_ε^2 and use estimates of $\hat{\alpha}$ from equation (19) and $\hat{\mu}_\varepsilon$ (20)

$$\begin{aligned}\hat{\sigma_\varepsilon^2}^{CLS} &= \frac{\sum_{t=2}^{n} V_t - \hat{\alpha}(1-\hat{\alpha})\sum_{t=2}^{n} X_{t-1}}{n-1} \\ &= \frac{\sum_{t=2}^{n}(X_t - \hat{\alpha}X_{t-1} - \hat{\mu_\varepsilon})^2 - \hat{\alpha}(1-\hat{\alpha})\sum_{t=2}^{n} X_{t-1}}{n-1}\end{aligned} \tag{21}$$

Lastly, the CLS for the parameters p and r are:

$$\hat{p}^{CLS} = 1 - \frac{\hat{\mu}_\varepsilon^{CLS}}{\hat{\sigma_\varepsilon^2}^{CLS}} \quad \text{and} \quad \hat{r}^{CLS} = \frac{\hat{\mu^2}_\varepsilon^{CLS}}{\hat{\sigma_\varepsilon^2}^{CLS} - \hat{\mu}_\varepsilon^{CLS}}. \tag{22}$$

Conditional maximum likelihood

Assume we have $X_1, X_2, ..., X_n$ with fixed X_1, be a random sample of size n from a Negative Binomial INAR(1) process, with parameters $\boldsymbol{\theta} = (\alpha, p, r)^T$. To estimate the unknown parameters we maximize the conditional log-likelihood function:

$$\hat{\boldsymbol{\theta}}^{CML} = \underset{\boldsymbol{\theta}}{\text{argmax}} \sum_{t=2}^{n} \log p_{x_t|x_{t-1}}(\boldsymbol{\theta}), \tag{23}$$

where

$$\begin{aligned} p_{x_t|x_{t-1}}(\boldsymbol{\theta}) &:= P(X_t = x_t|X_{t-1} = x_{t-1}) \\ &= \sum_{k=0}^{\min(x_t,x_{t-1})} \binom{x_{t-1}}{k} \alpha^k (1-\alpha)^{x_{t-1}-k} \binom{x_{t-1}-k+r-1}{r-1} \cdot (1-p)^{x_{t-1}-k} p^r. \end{aligned}$$

Simulation

Similarly as in previous example we ran two sets of 200 independent simulations for proposed estimates. The lengths for series were chosen $N = 200$ and $N = 500$, model parameters fixed $\alpha = 0.8$, $p = 0.3$ and $r = 2$. To compare the accuracy we use *Bias*, described in equation (13) and RMSE (14). CLS estimates were used as the initial values for the CML estimations. Results are presented in Table 2.

From *Bias* and RMSE we can observe that parameters estimated by CML are the most accurate, values are smallest. Naturally in all cases, simulations with larger sample were more accurate. As in Poisson INAR(1) case, MM and CLS do not differ a lot.

		MM estimation		CLS estimation		CML estimation	
Size	Real parameter	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
N=200	$\alpha = 0.8$	0.004	0.001	0.003	0.0008	0.0005	0.0004
	$r = 2$	-0.164	0.032	-0.159	0.030	-0.045	0.014
	$p = 0.3$	-0.152	0.011	-0.151	0.010	-0.003	0.001
N=500	$\alpha = 0.8$	0.0017	0.0004	0.0014	0.0003	0.0002	0.0001
	$r = 2$	-0.076	0.015	-0.073	0.012	-0.017	0.006
	$p = 0.3$	-0.096	0.007	-0.094	0.006	-0.001	0.0005

Table 2: Simulation results of Negative Binomial INAR(1)

Forecasting

From equation (5) we can derive h-step-ahead forecast formula for INAR(1) with Negative Binomial distribution:

$$\mathbb{E}(X_{t+h}|\mathcal{F}_t) \overset{d}{=} \alpha^h X_t + \frac{pr}{1-p}\frac{1-\alpha^h}{1-\alpha}. \tag{24}$$

3.2.1.3 Generalized Poisson Distribution

Lets say we have $\varepsilon_t \sim GP(\lambda, \eta)$ with probability mass function

$$\mathbb{P}(\varepsilon_t = k) = \lambda(\lambda + \eta k)^{k-1} \frac{e^{-\lambda - \eta k}}{k!}, \quad k = 0, 1, ... \tag{25}$$

Mean and variance is known $\mathbb{E}(\varepsilon_t) = \frac{\lambda}{1-\eta}$, $\mathbb{V}ar(\varepsilon_t) = \frac{\mathbb{E}(\varepsilon_t)}{(1-\eta)^2}$.

Method of moments

Lets consider method of moments estimators for unknown parameters α, η and λ. Having the first two moments, mean and variance, we can derive estimators for parameters η and λ. It is done just by expressing one from another, and because $\alpha = \rho_1$ we have expressions:

$$\hat{\alpha}^{MM} = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)}, \quad \hat{\eta}^{MM} = 1 - \sqrt{\frac{\mathbb{E}(\varepsilon_t)}{\mathbb{V}ar(\varepsilon_t)}}, \quad \hat{\lambda}^{MM} = \mathbb{E}(\varepsilon_t)\sqrt{\frac{\mathbb{E}(\varepsilon_t)}{\mathbb{V}ar(\varepsilon_t)}}. \tag{26}$$

Conditional least squares

Parameters α, η and λ also can be estimated with CLS, because now we need to estimate three parameters instead of two we will use two-step conditional least squares method, just like it was shown for Negative Binomial case. Using equations for $\hat{\alpha}^{CLS}$ from equation (19), $\hat{\mu}_\varepsilon^{CLS}$ (20) and $\hat{\sigma_\varepsilon^2}^{CLS}$ (21) we derive CLS estimators for η and λ:

$$\hat{\eta}^{CLS} = 1 - \sqrt{\frac{\hat{\mu}_\varepsilon^{CLS}}{\hat{\sigma_\varepsilon^2}^{CLS}}}, \quad \hat{\lambda}^{CLS} = \hat{\mu}_\varepsilon^{CLS}\sqrt{\frac{\hat{\mu}_\varepsilon^{CLS}}{\hat{\sigma_\varepsilon^2}^{CLS}}} \tag{27}$$

Conditional maximum likelihood

Let $X_1, X_2, ..., X_n$ with fixed X_1, be a random sample of size n from a Generalized Poisson INAR(1) process, with parameters $\boldsymbol{\theta} = (\alpha, \lambda, \eta)^T$. To estimate the unknown parameters we maximize the conditional log-likelihood function:

$$\hat{\boldsymbol{\theta}}^{CML} = \underset{\boldsymbol{\theta}}{\text{argmax}} \sum_{t=2}^{n} \text{log} p_{x_t|x_{t-1}}(\boldsymbol{\theta}), \tag{28}$$

where

$$\begin{aligned} p_{x_t|x_{t-1}}(\boldsymbol{\theta}) &:= P(X_t = x_t | X_{t-1} = x_{t-1}) \\ &= \sum_{k=0}^{\min(x_t, x_{t-1})} \binom{x_{t-1}}{k} \alpha^k (1-\alpha)^{x_{t-1}-k} \frac{\lambda[\lambda + (x_t - k)\eta]^{x_t - k - 1} e^{-\lambda - (x_t - k)\eta}}{(x_t - k)!}. \end{aligned}$$

Simulation

Likewise in Poisson and Negative Binomial distributions sections, we ran two sets of 200 independent simulations for checking estimators. The lengths for series were chosen $N = 200$ and $N = 500$, model parameters fixed $\alpha = 0.8$, $\eta = 0.2$ and $\lambda = 3$. For the accuracy comparison we use *Bias*, described in equation (13) and RMSE (14). CLS estimates were used as the initial values for the CML estimations. Results are presented in Table 3.

From *Bias* and RMSE we can observe that parameters estimated by CML are the most accurate, values are the smallest. The worst performance showed MM, accuracy increases as the number of observations increase, nevertheless, it still has the highest *Bias* and RMSE. As in previous simulation cases, MM and CLS do not differ a lot.

		MM estimation		CLS estimation		CML estimation	
Size	Real parameter	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
N=200	$\alpha = 0.8$	0.0207	0.003	0.017	0.003	0.004	0.002
	$\eta = 0.2$	0.057	0.111	0.062	0.011	0.013	0.005
	$\lambda = 3$	-0.62	0.103	-0.589	0.0977	-0.147	0.046
N=500	$\alpha = 0.8$	0.003	0.0008	0.0022	0.0008	0.0008	0.0005
	$\eta = 0.2$	0.008	0.0026	0.008	0.002	0.004	0.0015
	$\lambda = 3$	-0.094	0.023	-0.085	0.013	-0.034	0.0126

Table 3: Simulation results of Generalized Poisson INAR(1)

Forecasting

From equation (5) we can derive h-step-ahead forecast formula for INAR(1) with Generalized Poisson:

$$\mathbb{E}(X_{t+h}|\mathcal{F}_t) \overset{d}{=} \alpha^h X_t + \frac{\lambda}{1-\eta}\frac{1-\alpha^h}{1-\alpha}$$

3.3 Seasonal INAR

Often in the data, we observe seasonality, which can be monthly, weekly, etc. For this reason, in this section we present the seasonal INAR process, define how to estimate parameters and how to get the point forecast. The first-order seasonal non-negative INAR model is defined following [9].

Definition 3. A discrete-time non-negative integer-valued stochastic process $\{X_t\}_{t\in\mathbb{Z}}$ is said to be a first-order seasonal INAR process with seasonal period s (INAR$(1)_s$) if is satisfies the following equation:

$$X_t = \alpha \circ X_{t-s} + \varepsilon_t, \quad t \in \mathbb{Z}, \tag{29}$$

where $\alpha \in [0,1]$, $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an innovation sequence of i.i.d. non-negative integer-valued random variables not depending on past values of $\{X_t\}_{t\in\mathbb{Z}}$ and $s \in \mathbb{N}$ denotes the seasonal period. It is also assumed that the Bernoulli variables that define $\alpha \circ X_{t-s}$, that is, the Bernoulli variables from which X_t are obtained, are independent of the Bernoulli variables from which other values of the series are calculated. Moreover, we assume that all Bernoulli variables defining the thinning operations are independent of the innovation sequence $\{\varepsilon_t\}_{t\in\mathbb{Z}}$.

Process described in equation (29) has two random components: $\alpha \circ X_{t-s}$ number of survivals of the process at time $t-s$ with survival probability α and innovations ε_t describing elements that entered the system at time interval $(t-s, t]$. Note, that in case $s = 1$ we have previously described INAR(1) (1) model.

Lets assume we have seasonal Poisson INAR(1) process. Then the CLS estimator $\hat{\boldsymbol{\theta}}_{CLS} = (\hat{\alpha}_{CLS}, \hat{\lambda}_{CLS})^T$ of $\boldsymbol{\theta} = (\alpha, \lambda)^T$ is given by

$$\hat{\boldsymbol{\theta}}_{CLS} \coloneqq \underset{\boldsymbol{\theta}}{\operatorname{argmin}}(\sum_{t=s+1}^{n} [X_t - \mathbb{E}(X_t|\mathcal{F}_{t-1})])^2.$$

We know $\mathbb{E}(X_t|\mathcal{F}_{t-1}) = \mathbb{E}(X_t|X_{t-s})] = \alpha X_{t-s} + \lambda$. By making analogous calculations as in Poisson INAR(1) case, we can find the CLS estimators for α and λ respectively, as:

$$\hat{\alpha}_{CLS} \coloneqq \frac{(n-s)\sum_{t=s+1}^{n} X_t X_{t-s} - \sum_{t=s+1}^{n} X_t \sum_{t=s+1}^{n} X_{t-s}}{(n-s)\sum_{t=s+1}^{n} X_{t-s}^2 - (\sum_{t=s+1}^{n} X_{t-s})^2}, \tag{30}$$

$$\hat{\lambda}_{CLS} \coloneqq \frac{1}{n-s}(\sum_{t=s+1}^{n} X_t - \hat{\alpha}_{CLS} \sum_{t=s+1}^{n} X_{t-s}). \tag{31}$$

Considering the forecasts for $X_{t+h}, h \in \mathbb{N}$, the distribution of X_{t+h} can be expressed as

$$X_{t+h} \overset{d}{=} \alpha^q \circ X_{t-r} + \sum_{j=0}^{q-1} \alpha^j \circ \varepsilon_{t+h-js}, \quad h \in \mathbb{N}. \tag{32}$$

Here $q \coloneqq \lceil h/s \rceil$ and $r \coloneqq qs - h$, $r \in 1, ..., s-1$, since $\lceil y \rceil$ denoting the upper integer part of $y \in \mathbb{R}$, $\lceil y \rceil \coloneqq \min\{n \in \mathbb{Z} | y \leq n\}$. Based on the observed sample $X_1, X_2, ..., X_t$, h-step-ahead conditional expectation is given by:

$$\mathbb{E}(X_{t+h}|\mathcal{F}_t) = \alpha^q X_{t-r} + \mu_\varepsilon \frac{1-\alpha^h}{1-\alpha}, \quad h \in \mathbb{N}. \tag{33}$$

Similarly as in INAR(1) case, CML estimators can be derived for INAR$(1)_s$. Applying the $s-$step Markov property of the process, conditional likelihood function is

$$L(\theta|X_1,...,X_n)=\prod_{t=s+1}^{n}P(X_t|X_{t-s}).$$

The CML estimators are obtained by maximizing the conditional log-likelihood function

$$l(\theta|X_1,...,X_n)=\sum_{t=s+1}^{n}\text{log}P(X_t|X_{t-s}).$$

3.4 BINAR

Another extension of simple INAR processes is bivariate INAR. Sometimes the data have a bivariate structure, so it is more appropriate to analyse it together.

Definition 4. Let $X_t = [X_{1,t}, X_{2,t}]^T, t \in \mathbb{Z}$ be stationary non-negative integer-valued bivariate time series and $\varepsilon_t = [\varepsilon_{1,t}, \varepsilon_{2,t}]^T, t \in \mathbb{Z}$ be a non-negative integer-valued bivariate random sequence independent from X_t. Then the process X_t is a bivariate INAR process if it satisfies the equation:

$$X_t = A \circ X_{t-1} + \varepsilon_t = \begin{bmatrix}\alpha_1 & 0\\ 0 & \alpha_2\end{bmatrix} \circ \begin{bmatrix}X_{1,t-1}\\ X_{2,t-1}\end{bmatrix} + \begin{bmatrix}\varepsilon_{1,t}\\ \varepsilon_{2,t}\end{bmatrix}, \quad t \in \mathbb{Z}, \tag{34}$$

here $\alpha_i \in [0,1), i = 1,2$.

Similarly we can define non-negative seasonal integer-value process:

$$X_t = A \circ X_{t-s} + \varepsilon_t = \begin{bmatrix}\alpha_1 & 0\\ 0 & \alpha_2\end{bmatrix} \circ \begin{bmatrix}X_{1,t-s}\\ X_{2,t-s}\end{bmatrix} + \begin{bmatrix}\varepsilon_{1,t}\\ \varepsilon_{2,t}\end{bmatrix}, \quad t \in \mathbb{Z}. \tag{35}$$

Lets assume innovations $\varepsilon_t = [\varepsilon_{1,t}, \varepsilon_{2,t}]^T, t \in \mathbb{Z}$ have a bivariate Poisson distribution with parameters λ_1, λ_2 and ϕ with probability mass function:

$$\begin{aligned} f(k,l) = \mathbb{P}(Z_1 = k, Z_2 = l) &= e^{-(\lambda_1-\lambda_2-\phi)}\frac{(\lambda_1-\phi)^k}{k!}\frac{(\lambda_2-\phi)^l}{l!}\\ &\times \sum_{m=0}^{\min(k,l)}\binom{k}{m}\binom{l}{m}m!\left(\frac{\phi}{(\lambda_1-\phi)(\lambda_2-\phi)}\right)^m. \end{aligned}$$

The mean, variance and covariance is known:

$$\mu(Z_i) = \mathbb{V}ar(Z_i) = \mathbb{E}(Z_i) = \lambda_i + \phi, \quad i = 1, 2$$
$$\mathbb{C}ov(Z_1, Z_2) = \phi.$$

Now we will derive some properties for seasonal BINAR, but it also applies for simple case, when $s = 1$.

1. $\mathbb{E}(X_{i,t}) = \dfrac{\lambda_i + \phi}{1 - \alpha_i}, \quad i = 1, 2;$
2. $\mathbb{E}(X_{i,t}|X_{i,t-s}) = \alpha_i X_{i,t-s} + \lambda_i + \phi, \quad i = 1, 2;$
3. $\mathbb{V}ar(X_{i,t}) = \dfrac{\lambda_i + \phi}{1 - \alpha_i}, \quad i = 1, 2;$
4. $\mathbb{C}ov(X_{i,t}, X_{j,t}) = \dfrac{\phi}{1 - \alpha_i \alpha_j}, \quad i \neq j.$

Proof:

1.

$$\begin{aligned}\mathbb{E}(X_{i,t}) &= \mathbb{E}(\sum_{k=0}^{\infty} \alpha_i^k \circ \varepsilon_{t-ks,i}) = \sum_{k=0}^{\infty} \mathbb{E}(\alpha_i^k \circ \varepsilon_{t-ks,i}) = \sum_{k=0}^{\infty} \alpha_i^k \mathbb{E}(\varepsilon_{t-ks,i}) \\ &= \sum_{k=0}^{\infty} \alpha_i^k (\lambda_i + \phi) = \frac{\lambda_i + \phi}{1 - \alpha_i}.\end{aligned}$$

2.

$$\begin{aligned}\mathbb{E}(X_{i,t}|X_{i,t-s}) &= \mathbb{E}(\alpha \circ X_{i,t-s} + \varepsilon_{i,t}|X_{i,t-s}) = \mathbb{E}(\alpha \circ X_{i,t-s}|X_{i,t-s}) + \mathbb{E}(\varepsilon_{i,t}|X_{i,t-s}) \\ &= \alpha_i \mathbb{E}(X_{i,t-s}|X_{i,t-s}) + \mathbb{E}(\varepsilon_{i,t}) = \alpha_i X_{i,t-s} + \lambda_i + \phi.\end{aligned}$$

3.

$$\begin{aligned}\mathbb{V}ar(X_{i,t}) &= \mathbb{V}ar(\sum_{k=0}^{\infty} \alpha_i^k \circ \varepsilon_{t-ks,i}) = \sum_{k=0}^{\infty} \mathbb{V}ar(\alpha_i^k \circ \varepsilon_{t-ks,i}) \\ &= \sum_{k=0}^{\infty} (\alpha_i^{2k} \mathbb{V}ar(\varepsilon_{t-ks,i}) + \alpha_i^k (1 - \alpha^k s_i) \mathbb{E}(\varepsilon_{t-ks,i}) \\ &= \sum_{k=0}^{\infty} (\alpha_i^{2k} (\lambda_i + \phi) + \alpha_i^k (1 - \alpha_i^k)(\lambda_i + \phi)) \\ &= \frac{\lambda_i + \phi}{1 - \alpha_i^2} + \frac{\lambda_i + \phi}{1 - \alpha_i} - \frac{\lambda_i + \phi}{1 - \alpha_i^2} = \frac{\lambda_i + \phi}{1 - \alpha_i}.\end{aligned}$$

4.

$$\begin{aligned}
\mathbb{C}ov(X_{i,t}, X_{j,t}) &= \mathbb{C}ov(\sum_{k=0}^{\infty} \alpha_i^k \circ \varepsilon_{t-ks,i}, \sum_{l=0}^{\infty} \alpha_j^l \circ \varepsilon_{t-l,j}) \\
&= \sum_{k,l=0}^{\infty} \alpha_i^k \alpha_j^l \mathbb{C}ov(\varepsilon_{t-ks,i}, \varepsilon_{t-ls,j}) \\
&= \sum_{k=0}^{\infty} \alpha_i^k \alpha_j^k \mathbb{C}ov(\varepsilon_{t-ks,i}, \varepsilon_{t-ks,j}) \\
&= \frac{\mathbb{C}ov(\varepsilon_{t-ks,i}, \varepsilon_{t-ks,j})}{1-\alpha_i\alpha_j} = \frac{\phi}{1-\alpha_i\alpha_j}.
\end{aligned}$$

To find unknown parameters for seasonal BINAR we will use CLS method, which minimizes the squared differences:

$$Q(\alpha_j, \lambda_j) = \min_{\alpha_j, \lambda_j} \sum_{t=s+1}^{n} (X_{j,t} - \alpha_j X_{j,t-s} - \lambda_j)^2.$$

We get the following system by taking partial derivatives of $Q(\alpha_j, \lambda_j)$ and equating them to zero:

$$\begin{cases} \frac{\partial Q}{\partial \alpha_j} & = -2\sum_{t=2}^{n} (X_{j,t} - \alpha_j X_{j,t-s} - \lambda_j) X_{j,t-s} = 0 \\ \frac{\partial Q}{\partial \lambda_j} & = -2\sum_{t=2}^{n} (X_{j,t} - \alpha_j X_{j,t-s} - \lambda_j) = 0 \end{cases}, \quad j = 1, 2. \tag{36}$$

Now from (36) we get the following expressions:

$$\sum_{t=s+1}^{n} X_{j,t} X_{j,t-s} - \alpha_j \sum_{t=s+1}^{n} X_{j,t-s}^2 - \lambda_j \sum_{t=s+1}^{n} X_{j,t-s} = 0 \tag{37}$$

and

$$\lambda_j = \frac{\sum_{t=s+1}^{n} (X_{j,t} - \alpha_j X_{j,t-s})}{n-s}. \tag{38}$$

By substituting expression from (37) into (38) and multiplying by $(n-s)$ we get:

$$(n-s)\sum_{t=s+1}^{n} X_{j,t}X_{j,t-s} - \alpha_j(n-s)\sum_{t=s+1}^{n} X_{j,t-s}^2 - \sum_{t=s+1}^{n} X_{j,t}\sum_{t=s+1}^{n} X_{j,t-s} + \alpha_j(\sum_{t=s+1}^{n} X_{j,t-s})^2 =$$

$$\alpha_j((\sum_{t=s+1}^{n} X_{j,t-s})^2 - (n-s)\sum_{t=s+1}^{n} X_{j,t-s}^2) + (n-s)\sum_{t=s+1}^{n} X_{j,t}X_{j,t-s} - \sum_{t=s+1}^{n} X_{j,t}\sum_{t=s+1}^{n} X_{j,t-s} = 0.$$

Now we have CLM estimators for α_j and λ_j, $j = 1, 2$.

$$\hat{\alpha}_j^{CLS} := \frac{(n-s)\sum_{t=s+1}^{n} X_{j,t}X_{j,t-s} - \sum_{t=s+1}^{n} X_{j,t}\sum_{t=s+1}^{n} X_{j,t-s}}{(n-s)\sum_{t=s+1}^{n} X_{j,t-s}^2 - (\sum_{t=s+1}^{n} X_{j,t-s})^2}, \tag{39}$$

$$\hat{\lambda}_j^{CLS} := \frac{1}{n-s}(\sum_{t=s+1}^{n} X_{j,t} - \hat{\alpha}_j^{CLS}\sum_{t=s+1}^{n} X_{j,t-s}). \tag{40}$$

However, these CLS estimates do not consider the dependence parameter for innovations assumed to have bivariate distribution. Based on suggestion by Padeli [18] it is proven, that models residuals are equal to the covariance of the innovations. Therefore, by minimizing the squared differences between model residuals and the innovation covariance the dependence parameter ϕ can be estimated:

$$Q(\mathbb{C}ov(\varepsilon_1,\varepsilon_2)) = \min_{\mathbb{C}ov(\varepsilon_1,\varepsilon_2)} \sum_{t=s+1}^{n} ((X_{1,t} - \alpha_1 X_{1,t-s} - \lambda_1)(X_{2,t} - \alpha_2 X_{2,t-s} - \lambda_2) - \mathbb{C}ov(\varepsilon_1,\varepsilon_2))^2.$$

Assuming $\mathbb{C}ov(\varepsilon_1,\varepsilon_2) = \phi$ it becomes:

$$Q(\phi) = \min_{\phi} \sum_{t=s+1}^{n} ((X_{1,t} - \alpha_1 X_{1,t-s} - \lambda_1)(X_{2,t} - \alpha_2 X_{2,t-s} - \lambda_2) - \phi)^2.$$

By taking the derivative in respect of ϕ and equating it to zero and expressing ϕ we get:

$$\phi = \frac{1}{n-s}\Big\{ \sum_{t=s+1}^{n} (X_{1,t} - \alpha_1 X_{1,t-s})(X_{2,t} - \alpha_2 X_{2,t-s}) - \lambda_1 \sum_{t=s+1}^{n} (X_{2,t} - \alpha_2 X_{2,t-s}) -$$
$$\lambda_2 \sum_{t=s+1}^{n} (X_{1,t} - \alpha_1 X_{1,t-s}) + \lambda_1\lambda_2 \Big\}.$$

Now by substituting $\hat{\lambda}_j^{CLS}, j = 1, 2$ from (40) we get CLS estimator for ϕ:

$$\begin{aligned}\hat{\phi}^{CLS} &= \frac{1}{n-s}\sum_{t=s+1}^{n}(X_{1,t}-\hat{\alpha}_1^{CLS}X_{1,t-s})(X_{2,t}-\hat{\alpha}_2^{CLS}X_{2,t-s})\\ &-\frac{1}{(n-s)^2}\sum_{t=s+1}^{n}(X_{1,t}-\hat{\alpha}_1^{CLS}X_{1,t-s})\sum_{t=s+1}^{n}(X_{2,t}-\hat{\alpha}_2^{CLS}X_{2,t-s})\\ &-\frac{1}{(n-s)^2}\sum_{t=s+1}^{n}(X_{2,t}-\hat{\alpha}_2^{CLS}X_{2,t-s})\sum_{t=s+1}^{n}(X_{1,t}-\hat{\alpha}_1^{CLS}X_{1,t-s})\\ &+\frac{1}{(n-s)^2}\sum_{t=s+1}^{n}(X_{1,t}-\hat{\alpha}_1^{CLS}X_{1,t-s})\sum_{t=s+1}^{n}(X_{2,t}-\hat{\alpha}_2^{CLS}X_{2,t-s})\\ &=\frac{1}{n-s}\Bigg\{\sum_{t=s+1}^{n}(X_{1,t}-\hat{\alpha}_1^{CLS}X_{1,t-s})(X_{2,t}-\hat{\alpha}_2^{CLS}X_{2,t-s})\\ &-\frac{1}{n-s}\sum_{t=s+1}^{n}(X_{1,t}-\hat{\alpha}_1^{CLS}X_{1,t-s})\sum_{t=s+1}^{n}(X_{2,t}-\hat{\alpha}_2^{CLS}X_{2,t-s})\Bigg\}.\end{aligned}$$

Here $\hat{\lambda}_j^{CLS}$ represents $\mathbb{E}(\varepsilon_{j,t})$. Since we use innovations that are distributed under bivariate Poisson distribution $(\varepsilon_1, \varepsilon_2) \sim BivariatePois(\lambda_1, \lambda_2, \phi)$, the expectations are calculated $\mathbb{E}(\varepsilon_{j,t}) = \lambda_j + \phi$. Therefore, the true estimator for λ_j is:

$$\hat{\lambda}_j^{*CLS} = \hat{\lambda}_j^{CLS} - \hat{\phi}^{CLS}, \quad j = 1, 2.$$

4 Covid-19 data

In this thesis INAR models are applied to COVID-19 data: deaths and incidences in Lithuania, Estonia, Croatia and Israel. First, we present an overview of the situation, then analyse the relationship between deaths and incidences and lastly apply models and produce a two-week forecast.

4.1 Data description

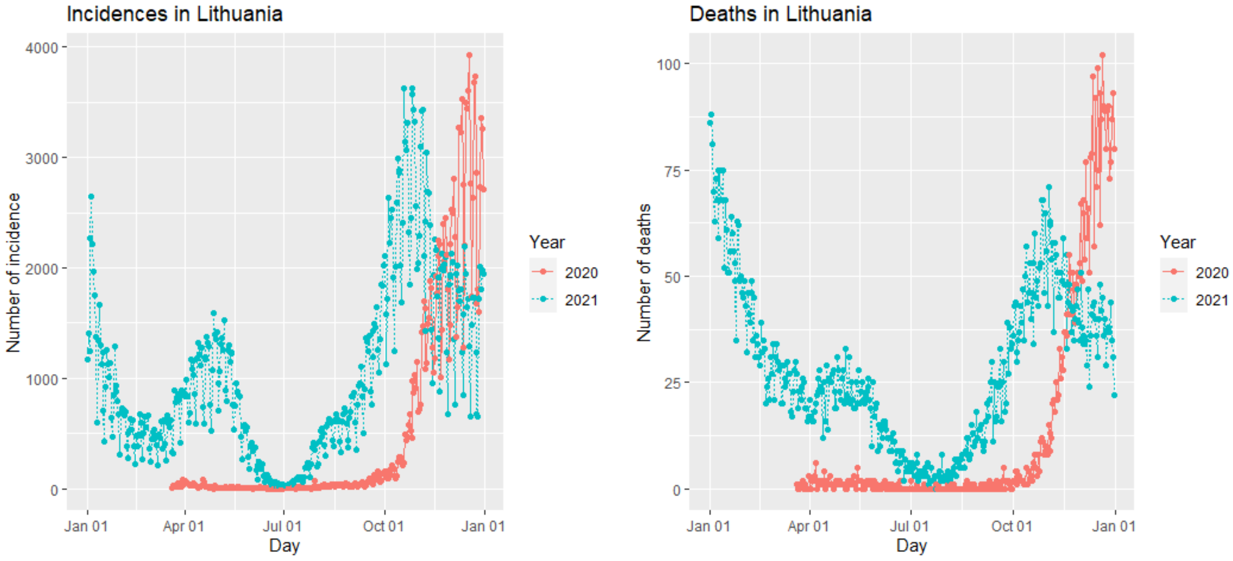
Data for Lithuania is taken from Lithuania Official Statistics Portal Open Data [2] and covers the period from the first registered case on February 1st, 2020, to December 31st 2021. Data set contains daily registered COVID-19 cases, deaths, vaccinations. The number of deaths is broken down into three groups based on the definition:

1. Number of deaths with COVID-19 as the leading cause of death. The indicator is calculated by summing all registered records of medical form E106 (unique persons), in which the main cause of death is ICD disease codes U07.1 or U07.2.

2. Number of deaths with COVID-19 of any cause of death. The indicator is calculated by summing all registered records of the medical form E106 (unique persons), in which the ICD disease codes U07.1, U07.2, U07.3, U07.4, U07.5 are indicated as the main, direct, intermediate cause of death or other important pathological condition.

3. Number of deaths of COVID-19 or COVID-19 deaths due to any cause and deaths due to non-external causes within 28 days. The indicator is calculated by summing all registered records of the medical form E106 (unique persons), in which the ICD disease codes U07.1, U07.2, U07.3, U07.4, U07.5 are indicated as the main, direct, intermediate cause of death or other important pathological condition, and all records in medical form E106 (unique individuals) who died within the last 28 days after receiving a positive diagnostic response to the SARS-CoV-2 test or had an entry in medical form E025 with ICD disease code U07.2 or U07.1.

For the analysis and calculations we chose the third definition, which is the widest of all and comparable with other countries. Looking at the Figure 1 we can observe that a better epidemic situation is during the warm season - late spring to early autumn. Also, it is possible to see three waves, and notice that the current wave started to increase earlier compared with 2020. Just to remind, in 2020 schools and universities were on remote teaching, and working from home, if possible, was highly recommended. Of course, a series of preventions were made so that the spread be stopped, the most important dates are presented in Table 4. There were two quarantines, which respectively lasted three and seven and a half months, another important date is the start of mass vaccination on May 31th 2020, before that the priority was for vulnerable parts of the population, such as older people or having some chronic disease, and for workers at the most important sections, like medics or teachers.



(a)

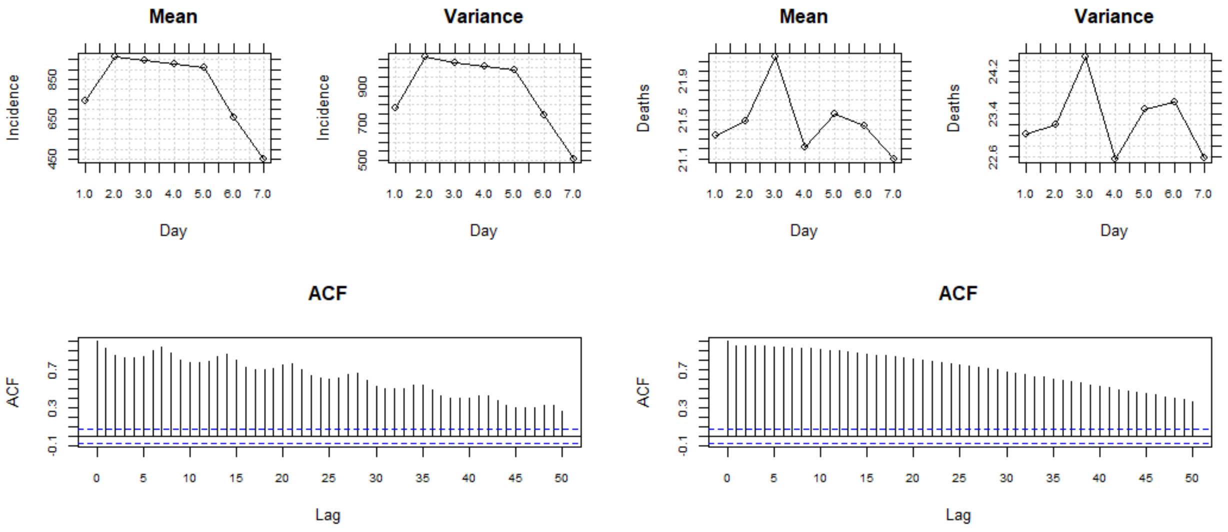
(b)

Figure 1: COVID-19 cases (a) and deaths (b) in Lithuania

Date	Event
2/28/2020	The first case detected
3/16/2020	Beginning of the first quarantine
6/17/2020	End of first quarantine
11/7/2020	Beginning of the second quarantine
12/27/2020	Vaccination started
5/31/2021	Start of mass vaccination
7/1/2021	End of the second quarantine
9/13/2021	Activities bounded by the National Certificate

Table 4: COVID-19 restrictions in Lithuania

By looking at the data we can suspect the seasonality for incidence data, which closely depends on the number of tests done. Usual during Friday and weekend testing slows down and that reflects on confirmed cases. It would be possible to use the derived indicator - share of positive tests from all tests, however, since we want to work with integer-value data it can not be a choice. To check seasonally we calculated daily means and variance, from Figure 2 (a) we can see that weekly seasonality do exist, this hypothesis is also confirmed by ACF, significant lag at 7 days period. The same hypothesis can be checked for deaths data, as seen in Figure 2 (b) there is no evidence for seasonality, which is expected.



(a)

(b)

Figure 2: Mean, variance and ACF for cases (a) and deaths (b) in Lithuania

Data for other countries are taken from the World Health Organisation website [6]. A few countries are selected: Estonia, Israel and Croatia. The first registered COVID-19 cases for all countries are different, and the last data point is 31st December, 2021. The dynamics for Lithuania, Estonia and Croatia are similar, lowers numbers of infected during late spring - early autumn, while during the colder season we observe an increase (see Appendix A 6). A little different situation can be observed in Israel data, here a fast increase in the summer of 2021 and a decrease in late autumn. Of course, the control of pandemic depends on taken precaution actions from the government, the vaccination rate.

As well as in Lithuania data, weekly seasonality in incidence exists and in other countries (see Appendix A 6), while there is no seasonality in deaths time series, as expected. Since we analyse a few distributions, we need to figure out what kind of data we have.

Country	Sample size	Mean	Variance	Min	Max	Overdispersion
Croatia	676	1049.8195	2178988.7	0	7315	2075.5841
Estonia	673	358.7043	193204.9	0	2300	538.6189
Israel	680	2027.1882	6766747.2	0	11345	3337.9965
Lithuania	652	800.1411	825264.4	0	3925	1031.3986

Table 5: Incidence data summary

Country	Sample size	Mean	Variance	Min	Max	Overdispersion
Croatia	676	18.480769	473.42630	0	92	25.6172
Estonia	673	2.870728	16.63951	0	32	5.7963
Israel	680	12.123529	180.79768	0	75	14.9130
Lithuania	652	21.466258	537.04494	0	102	25.0181

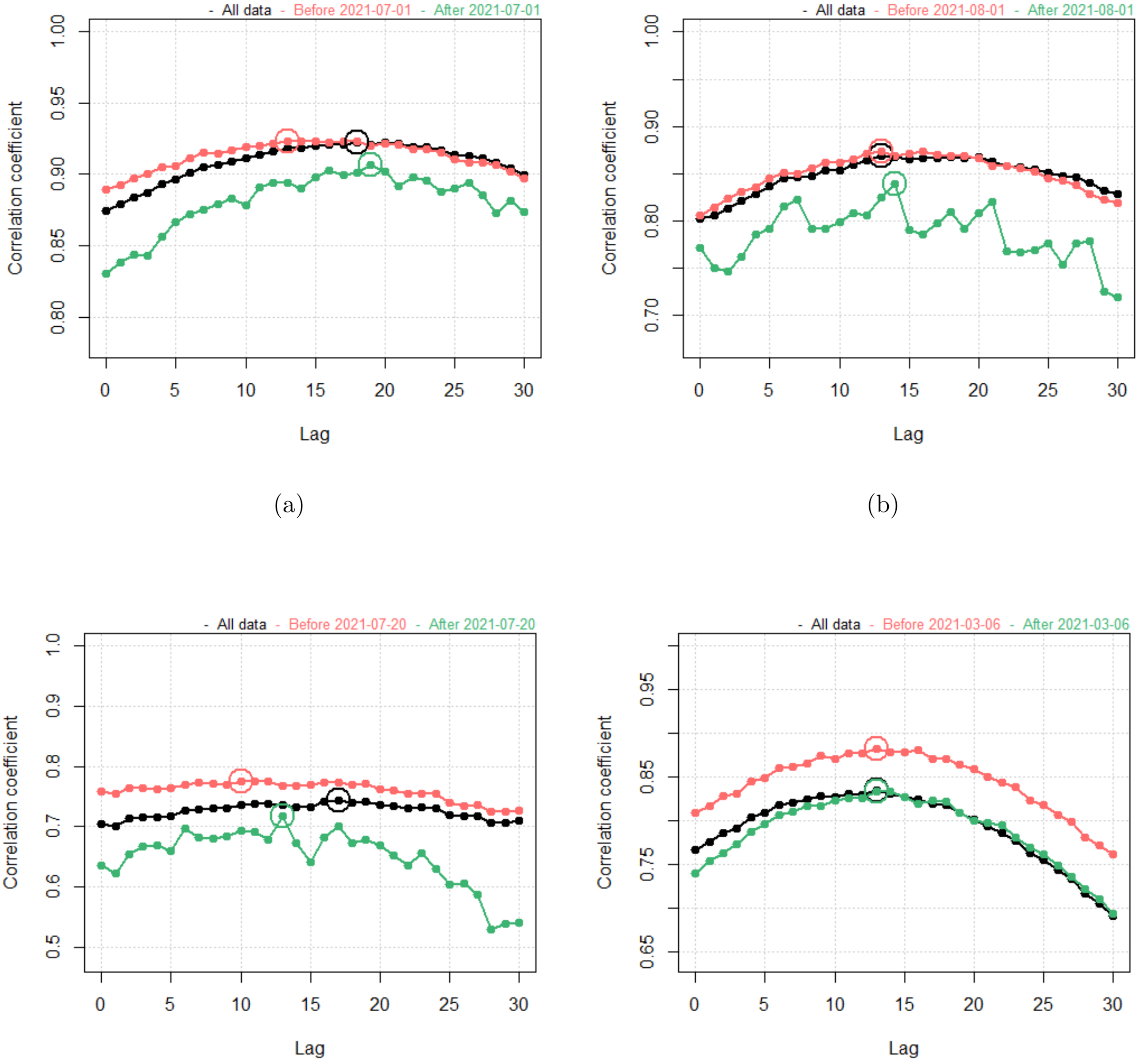
Table 6: Deaths data summary

Tables 5 and 6 show information about the data. Mortality is significantly highest in Lithuania, having in mind that the population in Croatia and Israel respectively are ≈4.1 and ≈9.3 million while the population in Lithuania are ≈2.7 million, according to the UN. Another thing we can observe from these two tables is the overdispersion, which means Poisson distribution might not be the best choice.

4.2 Incidence and deaths

What about the dependence between the number of incidences and the number of causalities? It is clear that deaths depend on confirmed cases, what is unknown with what lag the dependence is the highest, in other words after getting infected how many days passed till the person might die. To figure this out we calculate the Pearson correlation coefficients for 1 to 30 days. Also, since we know some information about vaccination in each country, additionally we split the time series into two parts: before vaccination and after vaccination. However, vaccination is rather a slow process, consequently, there is no exact date to use. Because of it, for each country selected the date was when a share of the vaccinated population in greater than 40%.

From the Figure 3 we can observe a slight difference between pre-vaccination and post-vaccination (assumption is 40% of the population are vaccinated), but it seems like the period is just a little bit longer, for instance in Lithuania before July 1st, 2021 the highest correlation coefficient was at day 13, while after July 1st, 2021 it is 19 days. Only for Israel in all three cases is the same number of days - 13 days. We could conclude that on average it is 15 days, so when we have a peak of number incidence, approximately after 15 days, we can expect the peak of the number of deaths.



(a)

(b)

(c)

(d)

Figure 3: Correlation between deaths and incidences in Lithuania (a), Croatia (b), Estonia (c) and Israel (d)

4.3 Two weeks forecast

Before applying the models we need to prepare the data. As seen in the graphs earlier we can observe some specifics: a wide range of values, high variability and trend. Knowing that the INAR type models are used for stationary data, we make a few data transformations. First, we take the logarithm of data, however, just by taking the logarithm, we get non-integer values meaning we can not apply models. Several articles proposed to take logarithm and then use only the integer part of values. Using this proposition we take ceilings of logarithmic data, that is $\lceil y \rceil := \min\{n \in \mathbb{Z} | y \leq n\}$. Also in order not to lose a lot of information, before taking the ceiling, we multiply data by 100, so the full transformation is:

$$\lceil 100 \cdot log(y + 10) \rceil.$$

The constant is added for dealing with zero values. The next step is to remove the trend. Just by looking at graphs, we see the trend, which is also confirmed by the Augmented Dickey-Fuller test, a p-value greater than 0.05 shows that the data has unit root. For trend elimination we calculate seven days median, however, we can not straightforward subtract it from the data because we would get negative values. To avoid it, a trend was shifted lower so that there are no negative values after subtraction. Performed Augmented Dickey-Fuller test showed (see Table 7) that after transformations there are no proof of unit root, the null hypothesis is rejected for both variables, all countries have a p-value less than 0.05.

		Before transformation		After transformation	
Country	Variable	Test statistic	p-value	Test statistic	p-value
Lithuania	Incidence	-1.72	0.69	-5.52	0.01
	Deaths	-1.24	0.9	-5.87	0.01
Estonia	Incidence	-2.4	0.41	-6.9	0.01
	Deaths	-2.02	0.58	-9.38	0.01
Croatia	Incidence	-1.66	0.72	-5.58	0.01
	Deaths	-1.67	0.71	-6.63	0.01
Israel	Incidence	-1.89	0.62	-3.72	0.02
	Deaths	-1.63	0.73	-6.12	0.01

Table 7: Unit root test

4.3.1 INAR

Having prepared data we can apply models and calculate a two-week forecast. For incidence data, we use INAR$(1)_7$ model since weekly seasonality was confirmed. INAR(1) will be used for the number of deaths. From the results presented in Tables 5 and 6 we can assume, that Poisson distribution is not a good fit, therefore we also take Negative Binomial and Generalized Poisson distributions, which are used for overdispersed data. In previous sections, we concluded the best estimation method is CML, accordingly, parameters will be estimated using this method.

	Distribution		
Parameter	Poisson	NB	GP
α	0.11	0.12	0.13
λ	90.2		68.96
r		148.72	
p		0.63	
η			0.21
RMSE	12.193	12.375	12.185
AIC	5145.872	5112.303	5086.734
BIC	5154.810	5120.641	5100.142

(a)

	Distribution		
Parameter	Poisson	NB	GP
α	0.018	0.04	0.04
λ	99.15		64.97
r		77.96	
p		0.45	
η			0.33
RMSE	14.377	14.452	14.381
AIC	5775.571	5629.469	5540.957
BIC	5784.573	5638.312	5554.461

(b)

	Distribution		
Parameter	Poisson	NB	GP
α	0.11	0.17	0.16
λ	89.16		55.49
r		64.65	
p		0.54	
η			0.34
RMSE	14.296	14.572	14.272
AIC	5712.404	5582.982	5484.284
BIC	5721.415	5596.499	5497.801

(c)

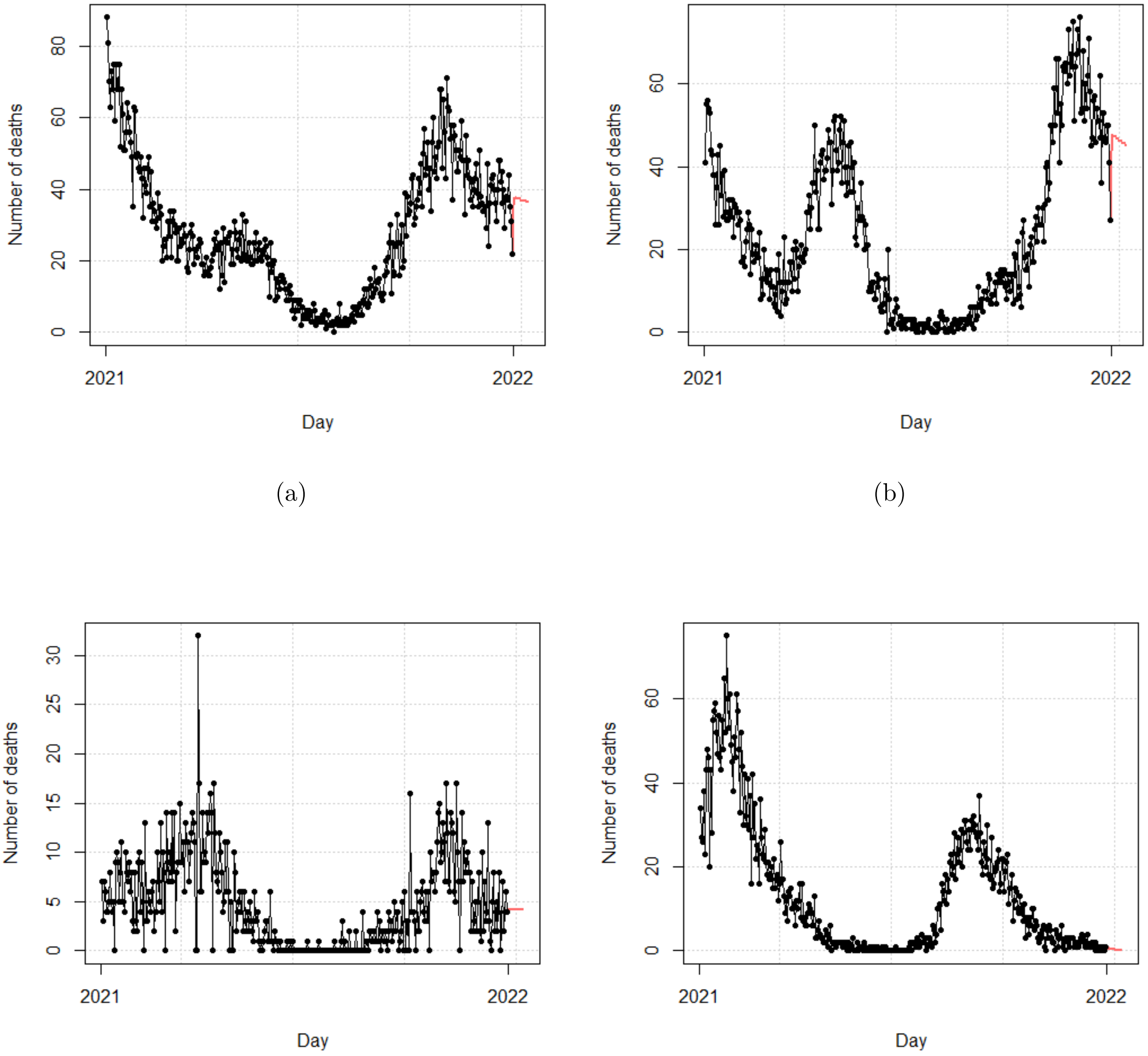
	Distribution		
Parameter	Poisson	NB	GP
α	0.09	0.12	0.12
λ	90.15		61.28
r		82.49	
p		0.51	
η			0.30
RMSE	13.673	13.770	13.662
AIC	5651.759	5575.839	5477.909
BIC	5660.783	5589.374	5491.445

(d)

Table 8: INAR(1) model results for number of deaths in Lithuania (a), Croatia (b), Estonia (c) and Israel (d)

Table 8 shows the results for all countries. To determine which model is the best, RMSE, Akaike Information Criterion (AIC) and Bayesian information criterion (BIC) was calculated. Based on these three indicators it can be concluded that INAR(1) model with Generalized Poisson distribution is the best fit for deaths data. In all four countries, AIC and BIC were the smallest for the model with Generalized Poisson distribution, while RMSE for Estonia and Croatia were slightly smaller for the model with the Poisson distribution.

Assuming the best choice is INAR(1) with Generalized Poisson distribution a two-week forecast is calculated, which is presented in Figure 5, the black line is the historical data, red - forecasted values. The dynamic for all countries is very similar - a slight decrease. If we look at the dynamic of new registered cases data for Lithuania and Croatia from the time $t - 15$, here t is December 31st, 2021, we also observe a slight decrease, which confirms that forecasts are quite reasonable. In Estonia and Israel past few weeks shows an increase in new cases, which might be because of a new fast-spreading COVID-19 variant Omicron. Having this in mind we could expect an increase in the number of deaths, however, this model does not take into account incidence data.



(a)

(b)

(c)

(d)

Figure 4: Two week forecast for Lithuania (a), Croatia (b), Estonia (c) and Israel (d)

A similar situation is in the number of new cases data, where INAR$(1)_7$ was used. If we look at the Table 9 it can be noticed that the smallest RMSE values are for the model with Poisson distribution for Lithuania, Estonia and Israel. But considering AIC and BIC unanimously the best fit for this data is INAR$(1)_7$ with Generalized Poisson distribution.

Having the results, 14-day forecasts were calculated by using the model with Generalized

Poisson distribution. It is visible that for Lithuania, Estonia and Israel cases will increase, while in Croatia will drop a little. Weekly fluctuations are also seen in the forecasts. Of course to calculate an accurate forecast is quite impossible since it depends on many factors, such as virus mutation, which could be highly transmissible, vaccination rate etc.

	Distribution		
Parameter	Poisson	NB	GP
α	0.47	0.64	0.63
λ	156.43		48.96
r		66.19	
p		0.56	
η			0.54
RMSE	30.506	31.74	31.43
AIC	6475.471	6213.179	5915.363
BIC	6484.410	6226.587	5928.770

(a)

	Distribution		
Parameter	Poisson	NB	GP
α	0.24	0.19	0.21
λ	227.29		55.48
r		65.27	
p		0.22	
η			0.33
RMSE	33.526	33.6455	33.610
AIC	7603.348	6825.739	6641.678
BIC	7612.351	6939.243	6655.182

(b)

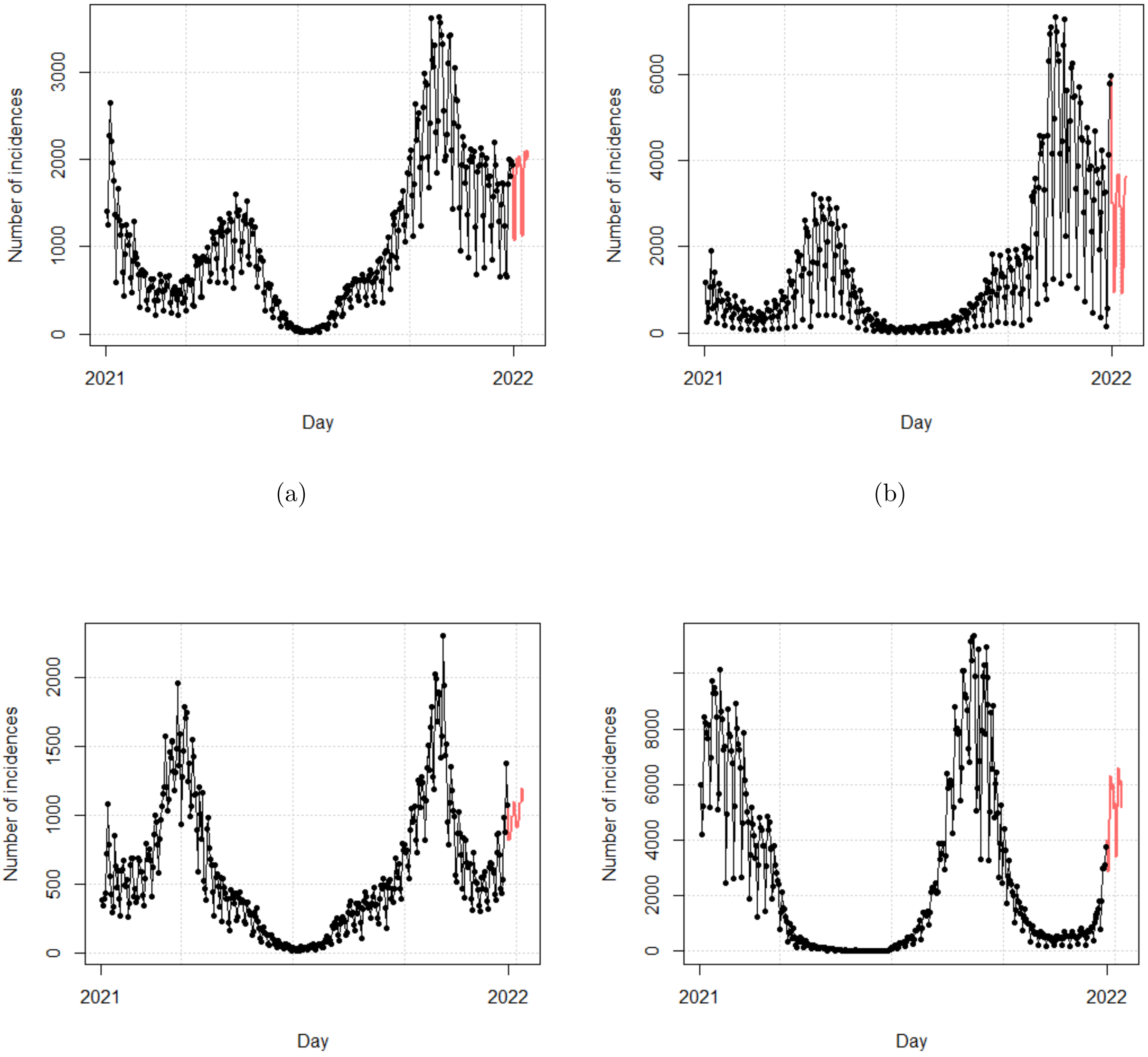
	Distribution		
Parameter	Poisson	NB	GP
α	0.11	0.17	0.16
λ	89.16		55.49
r		64.65	
p		0.54	
η			0.34
RMSE	56.549	62.085	55.118
AIC	10713.707	8694.388	7502.765
BIC	10722.718	8106.963	7516.282

(c)

	Distribution		
Parameter	Poisson	NB	GP
α	0.47	0.58	0.57
λ	155.51		51.06
r		65.90	
p		0.55	
η			0.59
RMSE	35.282	36.856	35.653
AIC	7501.892	7169.428	6486.868
BIC	7510.915	7182.964	6500.403

(d)

Table 9: INAR(1) model results for number of incidences in Lithuania (a), Croatia (b), Estonia (c) and Israel (d)



(a)

(b)

(c)

(d)

Figure 5: Two week forecast for Lithuania (a), Croatia (b), Estonia (c) and Israel (d)

4.3.2 BINAR

In previous sections the dependence between deaths and incidences was confirmed, also it could be assumed data have similar contributory factors: various preventions measures, weather or season etc., therefore, it is plausible to analyse time series as a bivariate process by applying bivariate INAR model. Since it is already known that incidence data has seasonality, we would like to include it in the model. Lets use BINAR(p) model, denote D_t - number of deaths, I_t - number of incidences:

$$\begin{cases} D_t & = \beta_1 \circ D_{t-1} + ... + \beta_p \circ D_{t-p} + \varepsilon_t^{(1)}, \\ I_t & = \alpha_1 \circ I_{t-1} + ... + \alpha_p \circ I_{t-p} + \varepsilon_t^{(2)}. \end{cases}$$

However, we would like to have restricted BINAR(p). Assume $p = 7$, since we have weekly seasonality, and coefficients $\{\beta_2, ..., \beta_7, \alpha_1, ..., \alpha_6\} = 0$, using binomial thinning property $0 \circ Z = 0$, we get the following model:

$$\begin{cases} D_t & = \beta_1 \circ D_{t-1} + \varepsilon_t^{(1)}, \\ I_t & = \alpha_7 \circ I_{t-7} + \varepsilon_t^{(2)}. \end{cases}$$

For simplicity instead of β_1 and α_7 we use β, α. Then the conditional means are:

$$\mathbb{E}(D_t|\mathcal{F}_{t-1}) = \mathbb{E}(\beta \circ D_{t-1} + \varepsilon_t^{(1)}|\mathcal{F}_{t-1}) = \beta D_{t-1} + \lambda_1,$$
$$\mathbb{E}(I_t|\mathcal{F}_{t-7}) = \mathbb{E}(\alpha \circ I_{t-7} + \varepsilon_t^{(2)}|\mathcal{F}_{t-7}) = \alpha I_{t-7} + \lambda_2.$$

Using the calculation provided in section 3.4 we will use CLS method to estimate the unknown parameters for $\beta, \alpha, \lambda_j, \phi$, here $j = 1, 2$.

$$\hat{\beta}^{CLS} := \frac{(n-1)\sum_{t=2}^n D_t D_{t-1} - \sum_{t=2}^n D_t \sum_{t=2}^n D_{t-1}}{(n-1)\sum_{t=2}^n D_{t-1}^2 - (\sum_{t=2}^n D_{t-1})^2}, \quad \hat{\lambda}_1^{CLS} := \frac{\sum_{t=2}^n D_{j,t} - \hat{\beta}^{CLS}\sum_{t=2}^n D_{t-1}}{n-1}, \tag{41}$$

$$\hat{\alpha}^{CLS} := \frac{(n-7)\sum_{t=8}^n I_t I_{t-7} - \sum_{t=8}^n I_t \sum_{t=8}^n I_{t-7}}{(n-7)\sum_{t=8}^n I_{t-7}^2 - (\sum_{t=8}^n I_{t-7})^2}, \quad \hat{\lambda}_2^{CLS} := \frac{\sum_{t=8}^n I_t - \hat{\alpha}^{CLS}\sum_{t=8}^n I_{t-7}}{n-7}. \tag{42}$$

The next step is to minimize the squared differences of models residuals and the covariance:

$$Q(\mathbb{C}ov(\varepsilon^{(1)}, \varepsilon^{(2)})) = \min_{\mathbb{C}ov(\varepsilon^{(1)}, \varepsilon^{(2)})} \sum_{t=8}^n ((D_t - \beta D_{t-1} - \lambda_1)(I_t - \alpha I_{t-7} - \lambda_2) - \phi)^2.$$

Again by taking the derivative in respect of ϕ and equating it to zero we get:

$$(n-7)\phi = \sum_{t=8}^n (I_t - \alpha I_{t-7})(D_t - \beta D_{t-1}) - \lambda_1 \sum_{t=8}^n (D_t - \beta D_{t-1}) - \lambda_2 \sum_{t=8}^n (I_t - \alpha I_{t-7}) + \lambda_1 \lambda_2.$$

Now by substituting $\hat{\lambda}_j^{CLS}, \quad j = 1, 2$, we get CLS estimate for ϕ:

$$\hat{\phi}^{CLS} = \frac{1}{n-7}\Bigg\{ \sum_{t=8}^{n}(I_t - \hat{\alpha} I_{t-7})(D_t - \hat{\beta} D_{t-1}) - \frac{\sum_{t=8}^{n}(I_t - \hat{\alpha} I_{t-7})\sum_{t=8}^{n}(D_t - \hat{\beta} D_{t-1})}{n-7} -$$
$$\frac{\sum_{t=8}^{n}(I_t - \hat{\alpha} I_{t-7})\sum_{t=8}^{n}(D_t - \hat{\beta} D_{t-1})}{n-1} + \frac{\sum_{t=8}^{n}(I_t - \hat{\alpha} I_{t-7})\sum_{t=8}^{n}(D_t - \hat{\beta} D_{t-1})}{(n-1)(n-7)}\Bigg\}.$$

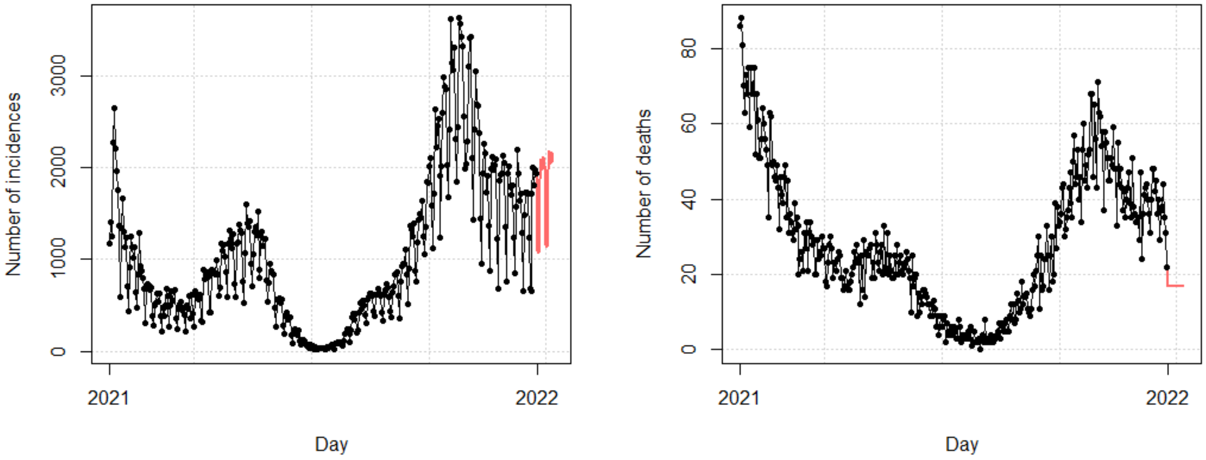
The final estimate for $\hat{\lambda}_j^{*CLS} = \hat{\lambda}_j^{CLS} - \hat{\phi}^{CLS}, j = 1, 2.$

Now having all the necessary information about the model, restricted BINAR will be applied to the data. The results are presented in the Table 10, since the parameters were estimated using the CLS method, RMSE is used for measuring model accuracy. Interesting results can be noticed by looking at the ϕ estimate, for Lithuania $\phi = 1.3$ which means only approximately one incidence/causality is caused by the same factors (season, virus mutation, etc.) for both time series, while for Croatia and Estonia it is around 26. If we compare the RSME of the BINAR model and individual model, calculated earlier (Tables 8 and 9), it can be concluded that individual models worked better, RMSE are lower for all countries for both time series. For deaths data separate models RMSE was around 12-14.5, meanwhile, for the BINAR model it varies from 14.3 - 34.3, a similar situation is for incidence data: RMSE from the separate model were 30 - 62, and for BINAR it is significantly higher 297 - 1324. However, it is essential to remember that for separate models Generalized Poisson distribution was used, which is more suitable for overdispersed data, moreover different estimation methods were used, from simulations we concluded CML presents more accurate estimators.

Parameter		Lithuania	Croatia	Estonia	Israel
β		0.14	0.17	0.018	0.14
α		0.67	0.81	0.39	0.71
λ_1		24.31	44.76	34.93	18.23
λ_2		97.75	78.63	153.71	74.51
ϕ		1.31	26.17	26.21	10.46
RMSE	Deaths	25.37	34.25	32.38	14.26
	Incidences	446.03	819.69	297.01	1323.54

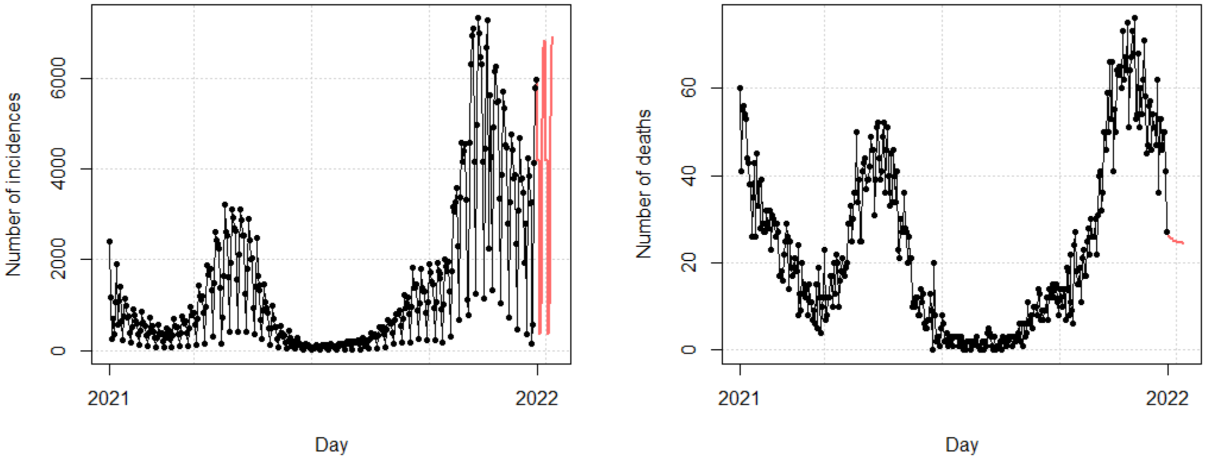
Table 10: BINAR(1) model results

As well as in previous cases, a 14-day forecast can be calculated. Figures 6-9 below illustrates the results, black line represents historical data, red - forecasted values. The forecasted number of incidences are increasing in all four countries, in Croatia a higher variability might be observed. The fastest increase in two weeks is for Israel, which is expected since the last week was rather a large growth. Taking a look at the number of deaths forecast all, except for Israel, are slightly decreasing.



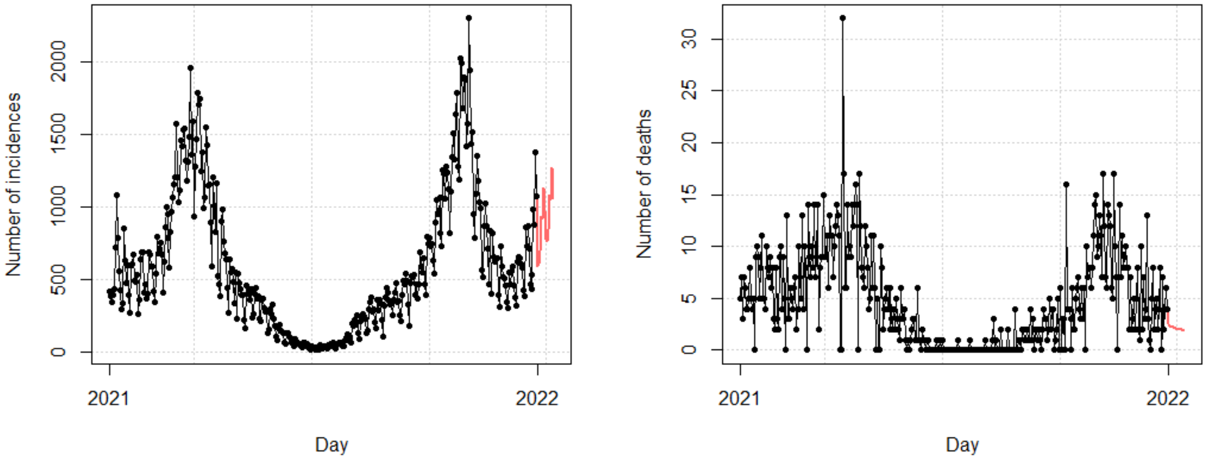
(a) (b)

Figure 6: Two week forecast using BINAR for Lithuania: incidences (a) and deaths (b)



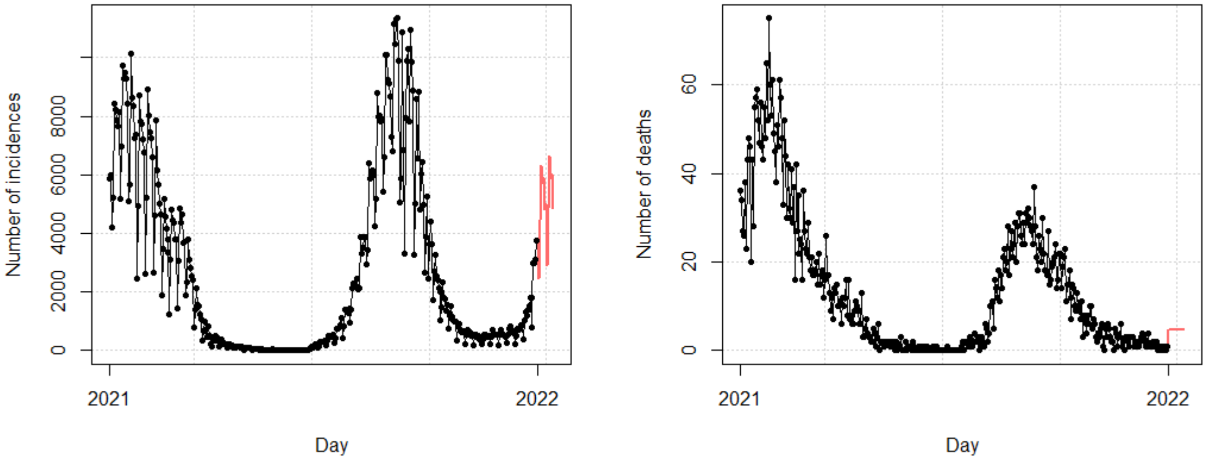
(a) (b)

Figure 7: Two week forecast using BINAR for Croatia: incidences (a) and deaths (b)



(a) (b)

Figure 8: Two week forecast using BINAR for Estonia: incidences (a) and deaths (b)



(a) (b)

Figure 9: Two week forecast using BINAR for Israel: incidences (a) and deaths (b)

5 Conclusions

In this thesis some non-negative integer-valued autoregressive models, that are easy to estimate and apply, were introduced. Furthermore, INAR models can be extended by including seasonality (INAR$_s$) or analysing data with bivariate structure (BINAR). Three estimators for the model parameters are considered: the method of moments, conditional least squares and conditional maximum likelihood. The efficiency of the estimators has been tested on simulated data and evaluated using Bias and RMSE. Considered models with different distributions: Poisson, Negative Binomial and Generalized Poisson.

Models for COVID-19 data: incidence and deaths in Lithuania, Estonia, Croatia and Israel were applied. The start day is unique for each country (early spring of 2020), while the last data point is in 31st December, 2021. In all four countries, 7 days seasonality in incidence data was observed, while no seasonality in deaths data, as expected. First, models were applied separately: INAR(1) for deaths and INAR(1)$_7$ for incidences. Several transformations were made: we took logarithm and ceilings to keep integer-valued series, after that the trend was removed so that data be stationary. By comparing models with different distributions, the best result, in respect of RMSE, showed models with Generalized Poisson innovations, which is reasonable having in mind overdispersion in the data.

Another part of the analysis was made by taking deaths and incidence as bivariate data case, where innovations are bivariate Poisson distribution. BINAR model was constructed as restricted BINAR(7), where some coefficients were equated to zero. Based on model accuracy and forecasts BINAR model did not outperform previously explored individual models.

Predicting COVID-19 cases and deaths are quite impossible, it depends on a bunch of various factors, such as vaccination, preventions like quarantine, social distancing or masks. Also at any time, a new more aggressive mutation might appear, such as a variant called Omicron, which is experienced currently at the beginning of 2022. The most recent researches state the Omicron has a higher transmissibility rate than previous mutations and might predominate in most places, moreover it is more resistant to vaccines, so the boost shot is recommended.

For future work, model modification might be applied, such as the incorporation of some lagged number of incidence, cause the more people get infected the more will die. For this case, it could be interesting to find a solution, if it exists, for model $D_t = \beta_1 \circ D_{t-1} + \alpha \circ I_{t-s}$, where s is highest correlation between deaths and lags of incidences, as seen earlier it is around 15 days. Another model that could be applied is the Poisson model with varying intensity.

References

[1] COVID-19 Dashboard by the Center of Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). https://coronavirus.jhu.edu/map.html. Accessed: 2021-12-21.

[2] COVID-19 in Lithuania. https://experience.arcgis.com/experience/cab84dcfe0464c2a8050a78f817924ca/page/Atviri-duomenys/. Accessed: 2021-12-21.

[3] Naming the coronavirus disease (COVID-19) and the virus that causes it. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it. Accessed: 2021-12-21.

[4] Scientific brief: SARS-COV-2 transmission. https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/sars-cov-2-transmission.html. Accessed: 2021-12-21.

[5] UNWTO World Tourism Barometer. https://www.e-unwto.org/loi/wtobarometereng. Accessed: 2021-06-01.

[6] WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int/table. Accessed: 2021-01-01.

[7] Agosto, A., and Giudici, P. A Poisson Autoregressive Model to Understand COVID-19 Contagion Dynamics. Risks 8, 3 (2020).

[8] Al-Osh, M. A., and Alzaid, A. A. First-Order Integer-Valued Autoregressive (INAR(1)) Process. Journal of Time Series Analysis 8, 3 (May 1987), 261–275.

[9] Bourguignon, M., Vasconcellos, K. L., Reisen, V. A., and Ispány, M. A Poisson INAR(1) process with a seasonal structure. Journal of Statistical Computation and Simulation 86, 2 (2016), 373–387.

[10] Chowell, G., Tariq, A., and Hyman, J. A novel sub-epidemic modeling framework for short-term forecasting epidemic waves. BMC Medicine 17 (08 2019).

[11] Freeland, R., and McCabe, B. Analysis of low count time series by Poisson autoregression. Journal of Time Series Analysis 25 (07 2004), 701 – 722.

[12] He, S., Peng, Y., and Sun, K. SEIR modeling of the COVID-19 and its dynamics. Nonlinear Dynamics 101, 3 (Aug 2020), 1667–1680.

[13] Hu, B., Guo, H., Zhou, P., and Shi, Z.-L. Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol 19, 3 (Oct. 2020), 141–154.

[14] Magal, P., and Webb, G. Predicting the number of reported and unreported cases for the COVID-19 epidemic in South Korea, Italy, France and Germany, 03 2020.

[15] Mamode Khan, N., Bakouch, H. S., Soobhug, A. D., and Scotto, M. G. Insights on the trend of the Novel Coronavirus 2019 series in some Small Island Developing States: A Thinning-based Modelling Approach. Alexandria Engineering Journal 60, 2 (2021), 2535–2550.

[16] McKenzie, E. Some simple models for discrete variate time series. JAWRA Journal of the American Water Resources Association 21, 4 (1985), 645–650.

[17] Moriña, D., Puig, P., Ríos, J., Vilella, A., and Trilla, A. A statistical model for hospital admissions caused by seasonal diseases. Statistics in medicine 30 (11 2011), 3125–36.

[18] Pedeli, X., and Karlis, D. A bivariate INAR(1) process with application. Statistical Modelling 11 (08 2011), 325–349.

[19] Ristic, M., Nastic, A., and Bakouch, H. Estimation in an Integer-Valued Autoregressive Process with Negative Binomial Marginals (NBINAR(1)). Communication in Statistics-Theory and Methods 41 (02 2012), 606–618.

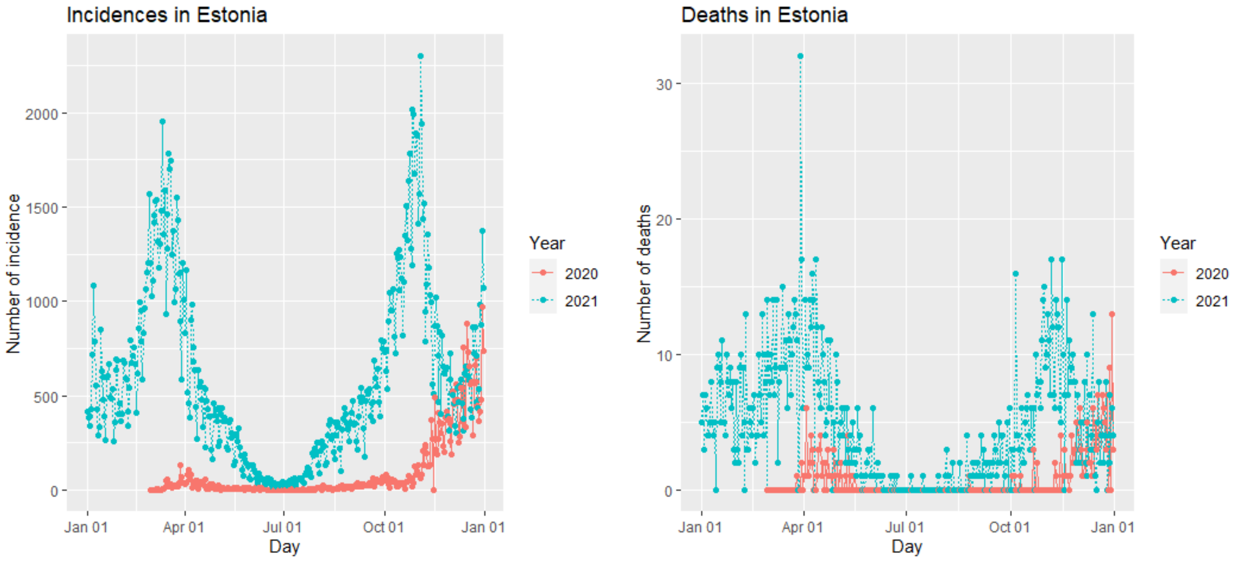
[20] Shengqi, T., Wang, D., and Cui, S. A seasonal geometric INAR(1) process based on negative binomial thinning operator. Statistical Papers 61 (12 2020).

[21] Viboud, C., Simonsen, L., and Chowell, G. A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks. Epidemics 15 (12 2015), 27–37.

[22] Weiß, C. An Introduction to Discrete-Valued Time Series. 01 2018.

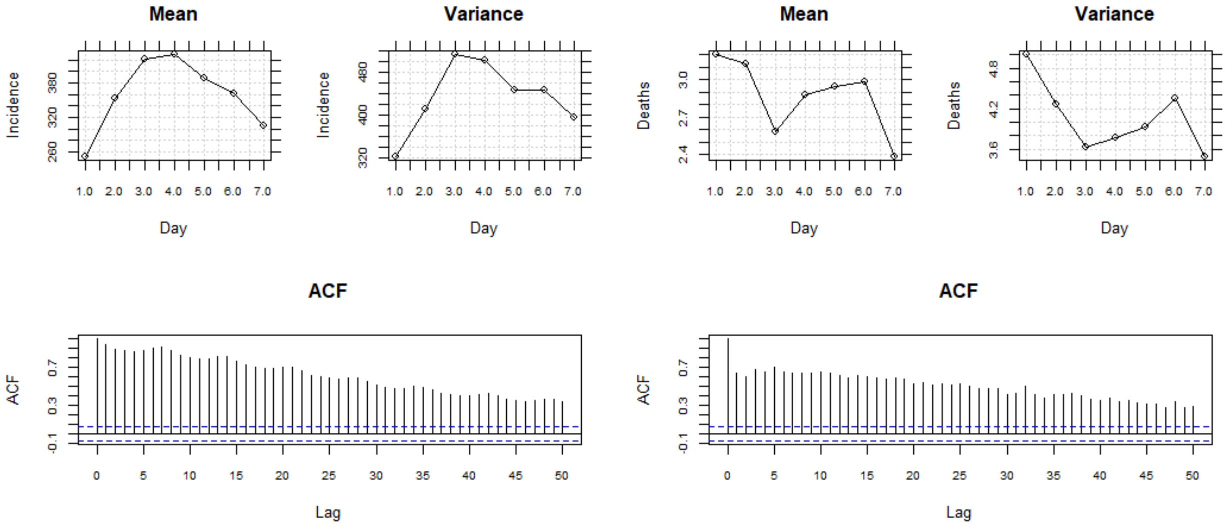
[23] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., and Tan, W. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine 382, 8 (2020), 727–733.

6 Appendix A



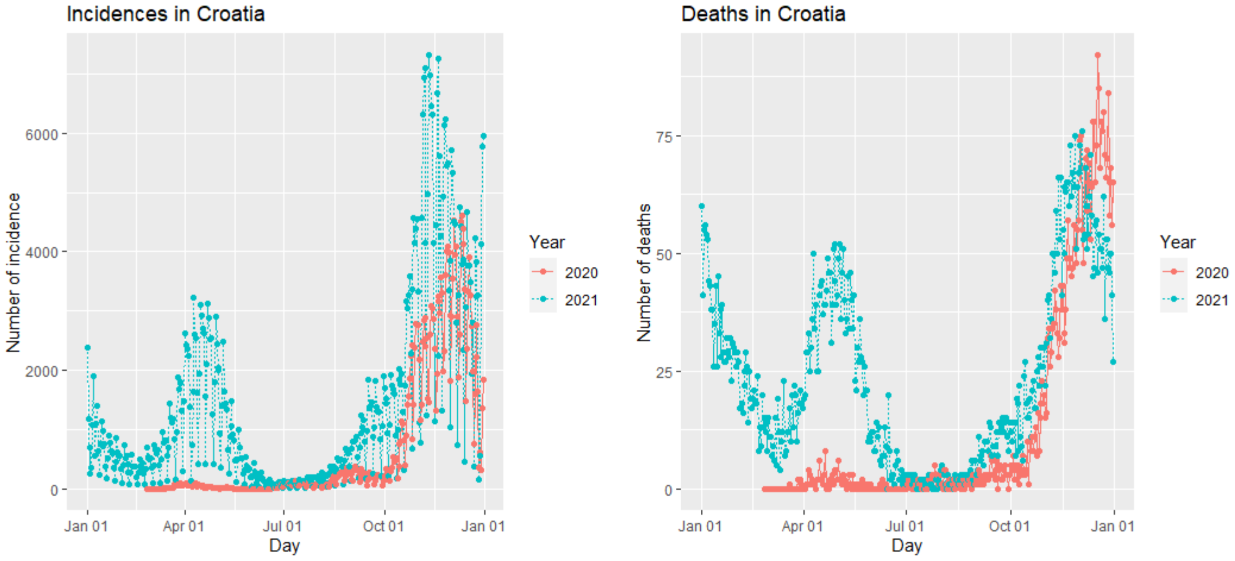
(a) (b)

Figure 10: COVID-19 cases (a) and deaths (b) in Estonia



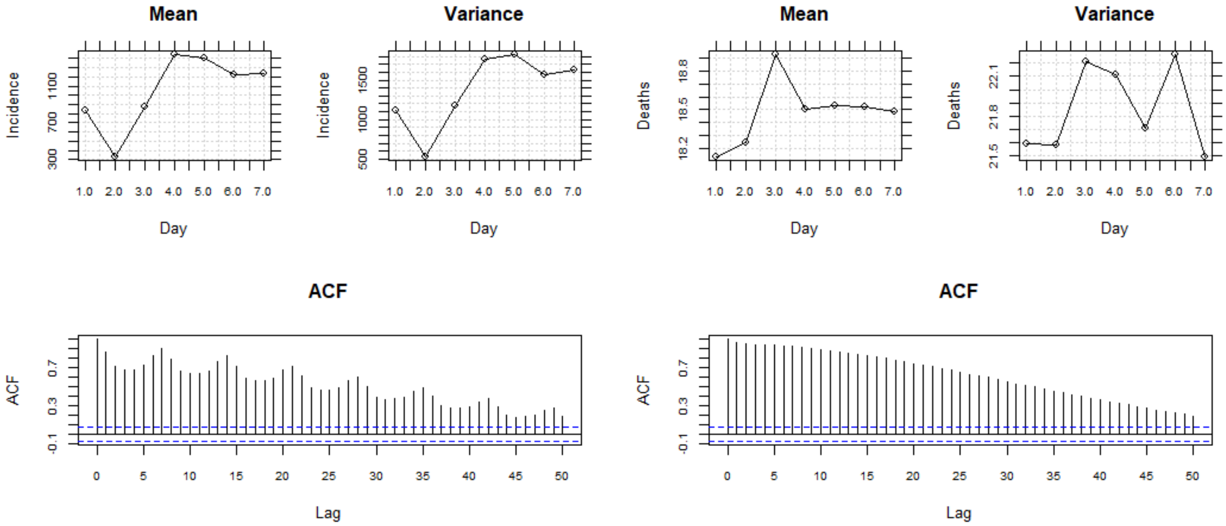
(a) (b)

Figure 11: Mean, variance and ACF for cases (a) and deaths (b) in Estonia



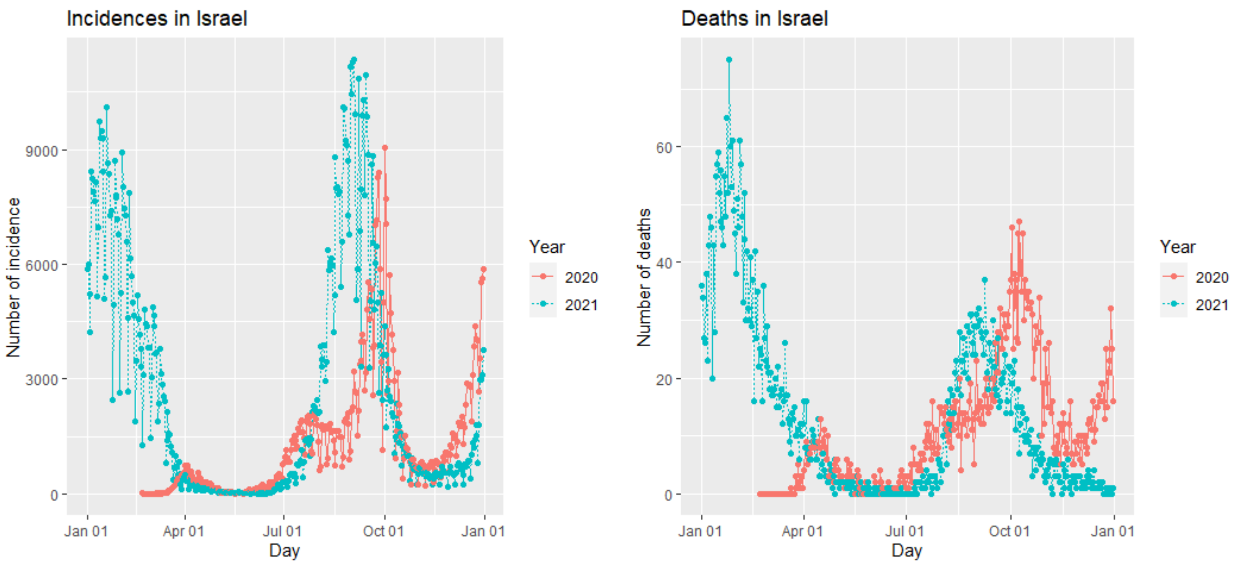
(a) (b)

Figure 12: COVID-19 cases (a) and deaths (b) in Croatia



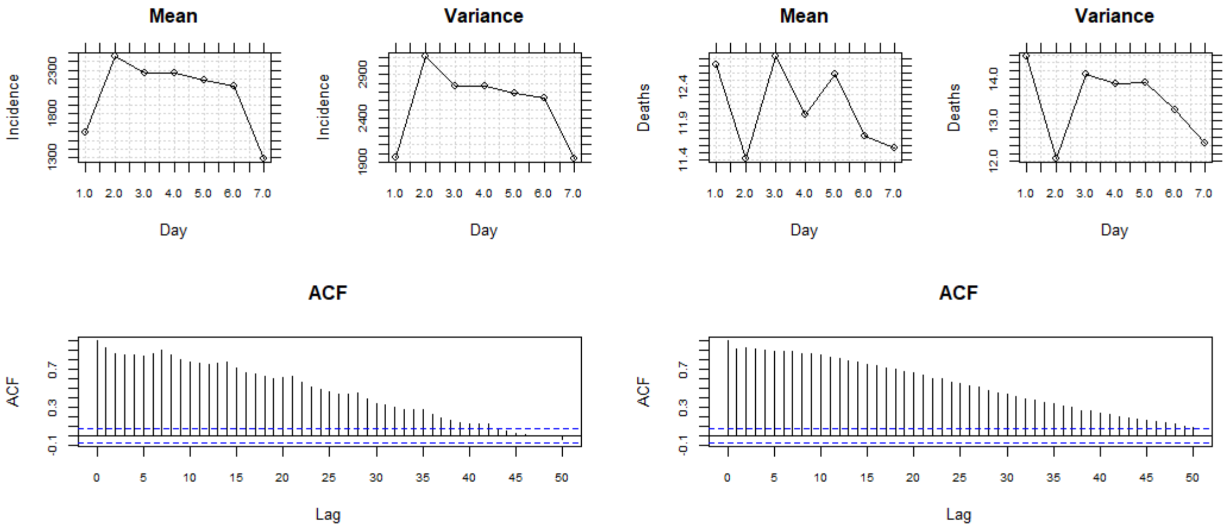
(a) (b)

Figure 13: Mean, variance and ACF for cases (a) and deaths (b) in Croatia



(a) (b)

Figure 14: COVID-19 cases (a) and deaths (b) in Israel



(a) (b)

Figure 15: Mean, variance and ACF for cases (a) and deaths (b) in Israel