ΠΜΣ ΒΙΟΣΤΑΤΙΣΤΙΚΗ

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΙΑΤΡΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

ΛΥΔΙΑ ΧΑΜΠΕΖΟΥ

Spatiotemporal clustering with an application on COVID-19 deaths in the provinces of the Netherlands

ΑΘΗΝΑ, 2022

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη

ΒΙΟΣΤΑΤΙΣΤΙΚΗ

που απονέμει η Ιατρική Σχολή και το Τμήμα Μαθηματικών του Εθνικού & Καποδιστριακού Πανεπιστημίου Αθηνών

Εγκρίθηκε την……. από την εξεταστική επιτροπή:

ΟΝΟΜΑΤΕΠΩΝΥΜΟ	***ΒΑΘΜΙΔΑ***	***ΥΠΟΓΡΑΦΗ***
1. …… ……	**…..……**	**……**
2. ……..……	**……**	**……**
3. ……	**……..……**	**……..**

Στην Περού, που με βγάζει πρωί-βράδυ βόλτα
για να ξεσκάω από το διάβασμα!

ΕΥΧΑΡΙΣΤΙΕΣ

Ολοκληρώνοντας τις μεταπτυχιακές σπουδές μου, θα ήθελα πρωτίστως να ευχαριστήσω τον κύριο Καρλή που όχι μόνο με εμπιστεύτηκε και βοήθησε καθοριστικά κατά τη διάρκεια της διπλωματικής μου εργασίας, αλλά και με ενέπνευσε για την περαιτέρω συνέχεια των σπουδών μου.

Επιπλέον, θα ήθελα να ευχαριστήσω τον μπαμπά μου που δεν μου χαλάει ποτέ χατίρι, τη μαμά μου που πάντα έχει μια καλή συμβουλή και την αδερφή μου που με κάνει να γελάω. Ευχαριστώ τη φίλη Δάφνη που μου ανεβάζει την αυτοπεποίθηση και τον αγαπημένο μου, Χαρίτωνα μου, που με την ευγενική κριτική του ματιά με βοηθάει να γίνομαι καλύτερη. Χάρη στους παραπάνω δεν ένιωσα ποτέ μόνη καθ'όλη τη διάρκεια της εκπόνησης της διπλωματικής μου εργασίας και τους ευχαριστώ θερμά.

Contents

Chapter 1

Introduction

1.1 COVID-19 pandemic

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. The first known case was identified in Wuhan, China, in December 2019. Henceforth, in less than three months, coronavirus was spread all over the world and on 11th of March, 2020, it was declared a global pandemic by the World Health Organization (WHO) (Li et al. 2020). As of August 2022, over 6.4 million people have been killed from coronavirus and over 581 million people have been diagnosed with COVID-19, globally. COVID-19 affects different people in different ways and thus, patients may present symptoms ranging from mild to severe or even be asymptomatic carriers. The most common symptoms include fever, cough, tiredness and loss of smell or taste, while other symptoms can be headache, diarrhoea, sore throat etc. (He et al. 2020). Serious illness is usually developed in older ages and in people with underlying medical conditions like chronic respiratory disease, cardiovascular disease, diabetes, or cancer but the majority of people infected by the virus won't need hospitalization (Garg et al. 2020).

In the Netherlands, the first person with COVID-19 was diagnosed on 27th of February, 2020, in the province of North Brabant, while a day after, the second case was identified in North Holland. Both cases had recently travelled to Lombardy, a region in Italy with increased virus load at that time. Residents of the Netherlands were strongly advised to use basic hygiene rules thoroughly and The National Institute for Public Health and the Environment was arranged for the isolation, tracing and general monitoring of the situation. During the first month, the majority of patients was identified in North Brabant. The first death occurred in 6th of March, 2020 and was an 86-year-old man who had been hospitalised. From 12th of March 2020, after 503 positive identified patients of novel coronavirus, measures were applied in the Netherlands to stop the rapid spread of coronavirus. Social distancing, closure of museums and venues in general, cancellation of events and mandatory online lectures in universities were some of the main restrictions. Three days after, on March 15th, despite the tries of the government to keep schools open, the inflation of the new patients led to the closure

of schools, bars, cafes etc. Additional measures were taken on 23rd of March and forced the prohibition of gatherings, while restriction at home was a strong indication. By August, 2022, over 8 million cases have been diagnosed positive and over 22 thousands have died from COVID-19, in the Netherlands.

In a more biological aspect, SARS-CoV-2 is an enveloped, single-stranded RNA virus of the genus Betacoronavirus and Coronaviridae family. Its virion is constructed from four main structural proteins, nucleocapsid (N), small envelope (E) glycoprotein, membrane (M) glycoprotein and spike (S) glycoprotein. The mechanism of entry of SARS-CoV-2 viral particles in host cells is mediated by S protein, localized in virion surface, which attaches to angiotensin-converting enzyme 2 (ACE2) receptor expressed mainly in lower respiratory tract organs (Lu et al. 2020, Shang et al. 2020). Since its emergence, SARS-CoV- 2 initial lineages (A and B) have continued to diversify forming the novel genetic "variants of concern (VOC)". These variants are evolved for increased transmissibility amongst humans, more severe disease, effective immune escape and reduction in neutralization from the host (Telenti et al. 2022). The S glycoprotein comprises the main target of neutralizing antibodies during host immune response to viral entry and is the main determinant of antigenic evolution thus leading to new variants (Duan et al. 2020). The different degrees of morbidity and mortality of COVID-19 in the population is due to the imbalanced early host response to SARS-CoV-2 infection at the cellular level since the transcriptional activation of Type I and III interferons (IFN-I and IFN-III, respectively) and subsequent upregulation of IFN-stimulated genes (ISGs), which are responsible for a balanced immune response and minimal tissue damage, are not properly achieved. The outcome of severe COVID-19 in these cases is multi-organ failure and death due to cytokine storm syndrome deriving from hyperinflammation (Mehta et al. 2020, Blanco-Melo et al. 2020). The virus spreads mainly between people, who are in close contact with each other, from an infected person's mouth or nose in small liquid particles, when they cough, sneeze, speak, sing or breathe, or through contaminated surfaces (Galbadage et al. 2020).

1.2 Spatiotemporal models

Considering the consequent spatiotemporal spread of COVID-19, spatiotemporal models have become particularly useful. Disease spreads evolve in time and space. Therefore, it is essential to consider time and space together in order to address changes in different geographical regions and understand why those changes happened. Spatiotemporal modeling arises when we have data in those two dimensions; time and space (Ibañez et al. 2021). Thus, those models can be employed to monitor the contagion dynamics of COVID-19 pandemic and provide a better understanding of the disease spread. Apart from COVID-19, spatiotemporal models could also be applied to crime counts (Liesenfeld et al. 2017, Wang & Brown 2012, Law et al. 2014), agricultural production (Müller et al. 2011, Diggle et al. 2002), infectious diseases (Bracher & Held 2020, Held et al. 2017), transportation (Rajabioun & Ioannou 2015,

Santos et al. 2018), environmental outcomes (Gusev 2008, Laurini 2019), etc. The general focus is to investigate the variation of an outcome in space and time over the area and time-period of interest.

There are several models developed to approach the spatiotemporal nature of an infectious disease. The basic transmission model for directly transmitted infectious diseases, is the Susceptible-Infected-Removed (SIR) model (Weiss 2013, Gaeta 2020). In the classical approach, the population, which is assumed homogeneous and isolated, is separated in three classes: susceptible, infected and removed. The model consists of three non-linear ordinary differential equations, from which useful information for the spread of the disease can be extracted (Smith et al. 2004). Numerous extensions of the SIR model have been developed during the years (Satsuma et al. 2004, Sadurní & Luna-Acosta 2021, Brugnano & Iavernaro 2020).

The Besag-York-Mollié (BYM) model is ideal for reliable estimations for relative risks for small areas or rare diseases. The overall variability is decomposed into a random Poisson component, a spatially structured region-specific random effect and an unstructured random part, across regional units, which allows the model to borrow the required information from the adjacent areas (D'Angelo et al. 2021, Alhdiri et al. 2017, Latouche et al. 2007).

Another type of spatiotemporal model, concerns the Multivariate Covariance Generalised Linear Models (MCGLM), which are used for multivariate count data and are specified through link functions for the mean vector of the outcome and linear predictors for the covariance matrix. MCGLM are flexible models, since they consider dependent variables of mixed types and allow covariance structures for longitudinal, spatial and spatiotemporal data (Ibañez et al. 2021, Bonat & Jørgensen 2016, Cressie & Zammit-Mangion 2016).

Other models, appropriate for spatiotemporal data, are Auto-Poisson models (Glaser 2017, Augustin et al. 2006), the Spatiotemporal Autoregressive Poisson model (P-SAR) (Glaser 2017, Rohimah et al. 2021), etc. A widely used and well developed model is the Endemic-Epidemic (EE) model, which will be utilized for this study.

The reasons for spatiotemporal modeling are various and can range from estimating the effect of a factor to a specific outcome, identifying clusters of areas with similar patterns or forecasting future observations (Ibañez et al. 2021). By means of spatiotemporal models, we are able to understand the past, which helps us to inform our understanding of the present and finally be able to make predictions about the future. Apparently, the information that arises from these models, can be utilized from governments and organizations, to decide and apply measurements for the control of the pandemic or the respective outcome of interest. The principles and the opportunities of the spatiotemporal models are comprehensively described in Meliker & Sloan (2011). Importantly, similar spatial and temporal patterns are being observed, a fact that urges

the need for clustering. At the same time, the increase in the size of data repositories has allowed the evolution of spatiotemporal clustering, which is a sub-field of data mining and can be processed by means of spatiotemporal modeling (Ansari et al. 2020).

1.3 Objectives and key results

In this thesis, our main objective is to provide model-based clustering through the endemic-epidemic model, which will be explored later in detail. Clustering is the task of grouping data sets by using a specific similarity measure. Cluster analysis is an unsupervised method of learning, since it doesn't require any a priori knowledge of the data sets. In spatial clustering, regional information is being used to create clusters, while in spatiotemporal clustering, which is an extension of spatial clustering, objects are grouped based on their spatial and temporal similarity (Madhulatha 2012, Ansari et al. 2020).

We analyzed data from January 2021 to August 2021, of the twelve provinces of The Netherlands and during this eight-month period we noticed similar patterns, related to the number of deaths, across the provinces. For this reason, we intended to incorporate the EM algorithm into the model, for clustering purposes. Finally, through this finite mixture model we ended up splitting the regions into groups that are more homogeneous within and more different to each other.

While aiming for a clustering method incorporated into the model, we went through the examination of several extensions of the endemic-epidemic model. Specifically, the simple form of the model is not always able to respond to the complex nature of a virus, while also technical issues, such as under-reporting or delayed reporting, further decrease the performance of the simple version of the endemic-epidemic model. For those reasons, we extended the epidemic-endemic model in order to study if and how the number of previous cases, previous deaths and seasonality affect the amount of deaths. The determination of the best extension for these data, according to some model selection criteria, constitutes the secondary goal of this work.

Therefore, our contribution in this work is two-fold. First, to provide a model for COVID-19 pandemic in The Netherlands, which includes important factors related to COVID-19 deaths. Since all extensions were exemplified to The Netherlands, the selected model will be able to predict and help the understanding of the disease spread there. Therefore, a model that contains all the important factors, would be able to facilitate decision-making and control of the rapid spread of the virus. The second and more outstanding contribution pertains to the generation of a model-based clustering method, which can be considered an important novel extension of the endemic-epidemic model. Especially when there is a large number of observations, clustering could stand in good stead. Moreover, in a wide range of fields, the outcome of interested can be monitored optimally, by using differentiated strategies for each of the clusters.

1.4 Overview of the paper

The rest of the paper is organized as follows. In the section of Methodology, mathematical details for the endemic-epidemic model, which is utilized in our work, as well as some important extensions, are provided. Moreover, our proposed methodology for clustering is described in depth. In the section of Data & Results, initially the data-set with information related to COVID-19 in the twelve provinces of The Netherlands for the time period between 1st of January to 31st of August, 2021, is comprehensively analyzed. Then we proceed with the application of the models that were described in the Methodology section to the above-mentioned data-set. More precisely, death counts are modeled not only by means of the extensions of the endemic-epidemic model, that already exist, but also by using the model-based clustering method we employed. In addition, in the same section, we present the findings derived from the models and the methodology that was employed. Finally, in the last section of Discussion, we commend on our work and results, outline some limitations and propose possible future improvements.

Chapter 2

Methodology

2.1 Endemic-epidemic (EE) model

2.1.1 Basic model

The endemic-epidemic (EE) framework is a class of time-series model created for the analysis of infectious diseases and it was proposed by Held et al. (2005). In specific, in their paper they suggest a statistical framework for univariate and multivariate infectious disease counts, while this framework can also be applied to non infectious diseases, with a slightly artificial interpretation. The simple version of the model represents a Poisson branching process model with immigration. In the multivariate formulation, in which we are interested, we consider the number of deaths in different spatial areas during eight months. We assume that we have k units ($k = 1, ..., K$), where each unit is one of the twelve provinces of The Netherlands. Several extensions have been proposed to develop the endemic-epidemic model (Bracher & Held 2020, Douwes-Schultz et al. 2022, Grimée et al. 2022, Held & Paul 2012).

The endemic epidemic model determines the mean value of the outcome of interest and uses incidence from the previous time-point $t - 1$ to explain the incidence in time-point t (Ibañez et al. 2021). The outcome of interest can be the newly infectious individuals of a contagious disease, deaths or in general discrete counts, while time can be measured in days, weeks, months, etc. In our case, we are interested in the number of deaths from COVID-19 and the time is measured in days. Specifically, we use daily data, during the time-period that begins on January 1st and ends on August 31st.

More precisely, we suppose that Y_{kt} denotes the number of confirmed deaths in a specific province $k = 1, ..., K$ in a country or in general in a region, at a specific day $t = 1, ..., T$. The counts, which are non negative and integer-valued, are usually supposed to follow Poisson or Negative Binomial distribution (Joe & Zhu 2005).

Since, we are working with count data, Poisson distribution is one of the candidate distributions to be used. In particular, Poisson distribution expresses the probability

of a number of events occurring over a specified period of time if these events are independent and occur with a known constant mean rate (Lambert 1992).

The probability mass function of a discrete random variable Y, which follows the Poisson distribution with parameter $\mu > 0$ is given below:

$$P(Y = k) = \frac{\mu^k}{k!} e^{-\mu}$$

where k is the number of events.

μ is equal to the expected number of Y, which is also assumed to be equal to the variance.

$$E(Y) = Var(Y) = \mu$$

In the case that deaths modelled in the endemic-epidemic model, are Poisson distributed, the mean value of them in the province k, at day t is expressed as:

$$Y_{kt}|Y_{k,t-1} \sim Poisson(\mu_{kt})$$
$$\mu_{kt} = E(Y_{kt}) = e_k v_{kt} + \lambda_{kt} Y_{k,t-1} + \phi_{kt} \sum_{q \neq k} w_{kq} Y_{q,t-1}$$

The details for the terms presented here, will be exhaustively described.

In practice, Poisson distribution is applied when we are interested in how many times an event of interest occurs based on one or more explanatory variables and the assumptions are that the events are independent, their rate through time is constant, and they cannot occur simultaneously.

The limitation of Poisson distribution is that mean and variance are assumed equal. However, especially in infectious diseases the first occurrence of an event makes a second occurrence more likely, leading to a variance greater than the mean value. We call this overdispersion and data related to contagious diseases are usually overdispersed. In this case, Poisson distribution is not appropriate and another one should be utilized in order to avoid a deflated standard error and inflated test statistics.

In this situation, Poisson distribution is replaced by the Negative Binomial distribution which provides an alternative more flexible option, allowing an overdispersion parameter ψ. The negative binomial distribution will converge to a Poisson distribution for large ψ ($\psi \to \infty$) (Yang & Berdine 2015) and therefore Poisson distribution can be considered to be a special case of the negative binomial distribution.

The probability mass function of the negative binomial distribution is:

$$P(Y = k) = \binom{k+\psi-1}{\psi-1}(1-p)^k p^\psi$$

where k is the number of deaths, ψ is the dispersion parameter and p is the probability of death.

Assuming that death counts follow the Negative Binomial distribution, the conditional mean $Y_{kt}|Y_{k,t-1}$ remains the same:

$$Y_{kt}|Y_{k,t-1} \sim NegBin(\mu_{kt}, \psi)$$

$$E(Y_{kt}) = e_k v_{kt} + \lambda_{kt} Y_{k,t-1} + \phi_{kt} \sum_{q \neq k} w_{kq} Y_{q,t-1}$$

where $\mu_{kt} = E(Y_{kt})$ and additionally, we have the overdispersion parameter ψ, where $\psi > 0$. However, the conditional variance increases:

$$Var(Y_{kt}) = \mu_{kt} \times (1 + \psi \mu_{kt})$$

Overdispersion parameter can either be considered the same among all provinces ($\psi_k = \psi$), or different for each region (ψ_k, where $\psi_k > 0$ for each $k = 0, ..., K$).

While Poisson and Negative Binomial distribution are the most common for endemic-epidemic models, alternative options are quasi-Poisson distributions (Ver Hoef & Boveng 2007)

Infectious disease surveillance data often display a mixture of endemic and epidemic behaviours. Therefore, it is reasonable that the endemic-epidemic model decomposes incidence into two components; the endemic and the epidemic one.

The first term is the endemic component ($e_k v_{kt}$) and describes information that is not directly linked to the outcome of interest. Namely, the endemic component refers to exogenous factors such as seasonality, temporal trends, socio-demographics, the size of the local population, etc (Celani & Giudici 2022). In the simplest version of the model, e_k denotes population fractions or the number population for each province $k = 1, .., K$ (Ibañez et al. 2021). The log-linear predictor (v_{kt}) can additionally capture information related to the day of the week, public holidays, temperature, humidity, seasonality, borders of each province, testing and vaccination proportions. By multiplying the v_{kt} by the size of the local population e_k, we get the mean of the endemic component for the respective region k.

The second component of the endemic-epidemic model, suggested by (Held et al. 2005), is the epidemic component and consists of two parts. The first part is called autoregressive and the second is the spatiotemporal part.

The autoregressive term ($\lambda_{kt}Y_{k,t-1}$), also called epidemic-within, describes how deaths in a specific province k, at a specific time t are affected by deaths that occurred at the previous time $t-1$, in the same province k. In this term, we therefore capture information related to the reproduction of the infectious disease, within the province, i.e temporal effect.

In the spatial/spatiotemporal part ($\sum_{q\neq k} w_{kq}Y_{q,t-1}$), otherwise mentioned as epidemic-between, we are interested in the transmission between provinces. It describes the way that death counts in a province k, at time t are affected by deaths recorded in the previous day, in other provinces $q \neq k$.

This effect of each province to the others is defined by the weight matrix (w_{ij}) where $i, j = 1, ..., k$. Weight matrices can be composed in various different ways, accounting for spatial distance. The most common one, is to consider an indicator function equal to 1 when the regions share common borders and 0 otherwise (Paul & Meyer 2016) as indicated below:

$$w_{ij} = \begin{cases} 1, & \text{if i and j are adjacent} \\ 0, & \text{if i and j are not adjacent} \end{cases} \quad i, j = 1, ..., k \tag{2.1}$$

In most cases those weights are normalised and restricted to be positive (Berlemann & Haustein 2020).

$$w_{ij}^{'} = \frac{w_i j}{\sum_j w_{ij}}$$

$$\sum_j w_{ij}^{'} = 1$$

Nevertheless, there are several interesting approaches.

For example, Celani & Giudici (2022) consider the inbound and outbound number of commuters and generate a categorical variable o_{ij} with six levels, where the larger the number of commuters, the lower the level. In this manner, "closeness" between two regions is defined by the movement between them and not by the physical borders. More precisely, the dependence of w_{ij} on o_{ij} is described by a power law function:

$$w_{ij} = o_{ij}^{-d}$$

where $d \geq 0$ is a decay parameter. The greater the d, the faster the decaying of the power law function.

Another conception is the one of Held et al. (2017), in which it is assumed that short-time travel behaviour follows approximately a power law with respect to distance.

The power of law model reflects the spread of COVID-19, between units, that occurs due to distance between them.

Therefore, weights are defined as:

$$w_{ij} = (o_{ij} + 1)^{-d} \tag{2.2}$$

where d is again the decay parameter.

Grimée et al. (2022) extend this idea even further. In their model, weights are time-dependent and therefore capture information related to changes of movement behaviours over time. Weights depend not only on the power law model, but also on Facebook mobility data and International Organization of Migration (IOM) data.

Facebook provides mobility data for the change of population. The weights are thus, updated to:

$$w_{ij,t} = w_{ij} \times (f_{i,t} + 1)$$

where w_{ij} is described in Equation 2.2 and $f_{i,t}$ is the Facebook movement change variable for region i at time t.

Eventually, border closures from the International Organization of Migration data were included in the final formula of weights:

$$w^{'}_{ij,t} = w_{ij,t} \times b_{ij,t}$$

where $b_{ij,t}$ denotes different border states:

$$b_{ij,t} = \begin{cases} 0.1, & \text{"Total restrictions" at time } t \\ 0.5, & \text{"Partial restrictions" at time } t \\ 1, & \text{otherwise} \end{cases}$$

In all three terms discussed above, the log-linear predictors of the endemic-epidemic equation can be modelled in a simple form:

$$\begin{aligned} \log(\upsilon_t) &= \alpha^{(\upsilon)}, \\ \log(\lambda_t) &= \alpha^{(\lambda)}, \\ \log(\phi_t) &= \alpha^{(\phi)} \end{aligned}$$

It is important to mention that all linear predictors v_t, λ_t, ϕ_t defined above, are constrained to be non negative. Furthermore, this form allows three different options, that could also be combined.

Firstly, the intercept can be fixed ($\alpha^{(i)}, i = v, \lambda$ or ϕ) and therefore, the same for all provinces. This approach is the one described above.

Secondly, intercepts can be region-specific ($b^{(\cdot)}$) and treated as fixed or random effects, considering that parameters vary across provinces.

$$\begin{aligned}
\log(v_{kt}) &= b_k^{(v)}, \\
\log(\lambda_{kt}) &= b_k^{(\lambda)}, \\
\log(\phi_{kt}) &= b_k^{(\phi)}.
\end{aligned}$$

Finally, additional exogenous explanatory variables (X_{kt}), in either of the three terms, may be included in the modeling of the log-linear predictors. Temporal trends, sine-cosine terms to account for seasonality, population fractions, population densities, vaccination coverage, borders and other covariates are therefore being covered in this part (Paul & Meyer 2016).

Accordingly, the linear predictors can be written in a general form:

$$\begin{aligned}
\log(v_{kt}) &= \alpha^{(v)} + b_k^{(v)} + \beta_i^{(v)} X_i, \quad \text{i=1,...N} \\
\log(\lambda_{kt}) &= \alpha^{(\lambda)} + b_k^{(\lambda)} + \beta_i^{(\lambda)} X_i, \quad \text{i=1,...N} \\
\log(\phi_{kt}) &= \alpha^{(\phi)} + b_k^{(\phi)} + \beta_i^{(\phi)} X_i, \quad \text{i=1,...N}
\end{aligned}$$

where $\alpha^{(v)}, \alpha^{(\lambda)}, \alpha^{(\phi)}$ are overall fixed intercepts (same for all provinces in each term),$b_k^{(v)}, b_k^{(\lambda)}, b_k^{(\phi)}$ are region-specific intercepts, $\beta_i^{(v)}, \beta_i^{(\lambda)}, \beta_i^{(\phi)}$ are coefficients for the respective covariates or indicator functions X_i and X_i are covariates or indicator functions.

An advantage of the endemic-epidemic model is that it provides, in general, an adequate fit and reliable one-step-ahead prediction intervals. Moreover, the important characteristic of the EE framework is that it does not require simulation-based inference such as computer-intensive Markov Chain Monte Carlo (MCMC) and parameters can be easily and rapidly estimated by Maximum Likelihood (ML), using generic optimization routines Held et al. (2005).

If the model contains random effects then inference is based on penalized quasi-likelihood procedures, as described in (Paul & Held 2011).

In our case, we utilize the methods from *optim*, which is a function that conducts general-purpose optimization. The main requirements of the *optim* function are the initial values, a function to get minimized (or maximized) and the method that will be employed. The default method is an implementation of that of Nelder & Mead (1965), which uses only function values at some points and does not form any gradient at those points. When this method doesn't converge, quasi-Newton method, which uses both function values and gradients, or the method of conjugate gradients are preferred. Other available methods are the method "L-BFGS-B", which allows box constraints and method "SANN", which is a variant of simulated annealing.

In the following sections, we will examine some extensions of the endemic-epidemic model.

2.1.2 Extension 1 - Order D

It is reasonable that every infectious disease is transmitted in a different rate. Therefore, in every infectious disease, deaths or cases on day t are affected by the number of deaths or cases in previous days. The simple version of the endemic-epidemic model accounts only for the previous day $t-1$. In the following extension we will examine a model that takes into account deaths of the last D days. In particular, in the epidemic component (autoregressive and spatiotemporal term) the sum of D previous days will be computed and thus, the number of deaths at time t in a specific province k will be assumed to be affected by the sum of deaths happened D days before in the same province k and in neighbouring provinces $q \neq k$ (Bracher & Held 2017).

The model will be the following:

$$E(Y_{kt}) = e_k v_t + \lambda_t \sum_{d=1}^{D} Y_{k,t-d} + \phi_t \sum_{q} \sum_{d=1}^{D} w_{qt} Y_{q,t-d}$$

and we assume the simple version of the log-linear predictors:

$$\begin{aligned} \log(v_t) &= \alpha^{(v)}, \\ \log(\lambda_t) &= \alpha^{(\lambda)}, \\ \log(\phi_t) &= \alpha^{(\phi)}. \end{aligned}$$

The above model suggests that each one of the D previous days has the same effect on the number of deaths. We can, thus further extend the model using different coefficients for each day-lag in the autoregressive term, in the spatiotemporal or in both of them.

Below, the extended models with linear predictors in their simple form, are presented.

Different coefficients in the autoregressive term:

$$E(Y_{kt}) = e_k v_t + \sum_{d=1}^{D} \lambda_{td} Y_{k,t-d} + \phi_t \sum_q \sum_{d=1}^{D} w_{qt} Y_{q,t-D}, d = 1, 2, ...D$$

$$\begin{aligned} \log(v_t) &= \alpha^{(v)}, \\ \log(\lambda_{td}) &= \alpha_d^{(\lambda)}, \\ \log(\phi_t) &= \alpha^{(\phi)}. \end{aligned}$$

Different coefficients in the spatiotemporal term:

$$E(Y_{kt}) = e_k v_t + \lambda_t \sum_{d=1}^{D} Y_{k,t-d} + \sum_q \sum_{d=1}^{D} \phi_{td} w_{qt} Y_{q,t-D}, d = 1, 2, ...D$$

$$\begin{aligned} \log(v_t) &= \alpha^{(v)}, \\ \log(\lambda_t) &= \alpha^{(\lambda)}, \\ \log(\phi_{td}) &= \alpha_d^{(\phi)}. \end{aligned}$$

Different coefficient in both terms:

$$E(Y_{kt}) = e_k v_t + \sum_{d=1}^{D} \lambda_{td} Y_{k,t-d} + \sum_q \sum_{d=1}^{D} \phi_{td} w_{qt} Y_{q,t-D}, d = 1, 2, ...D$$

$$\begin{aligned} \log(v_t) &= \alpha^{(v)}, \\ \log(\lambda_{td}) &= \alpha_d^{(\lambda)}, \\ \log(\phi_{td}) &= \alpha_d^{(\phi)}. \end{aligned}$$

2.1.3 Extension 2 - Covariates

Apparently, the most significant impact on the amount of deaths caused by SARS-CoV-2 derives from the number of cases. Especially in infectious diseases such as COVID-19, larger numbers of cases result in an increased probability of consequent deaths. A model with the variable of cases in the endemic component was examined for different lags. We are interested only in the cases that were identified within the same province. The rest log-linear predictors are presented in their simple form with fixed intercepts. The inclusion of covariates in the EE model has been described in previous works (Paul & Held 2011, Meyer et al. 2017).

The extension with cases as covariate in the endemic part is:

$$E(Y_{kt}) = e_k v_{kt} + \lambda_t Y_{k,t-1} + \phi_t \sum_q w_{qt} Y_{q,t-1}$$

$$\log(v_{kt}) = \alpha^{(v)} + \beta^{(v)} \sum_{d=1}^{D} X_{k,t-d},$$

$$\log(\lambda_t) = \alpha^{(\lambda)},$$
$$\log(\phi_t) = \alpha^{(\phi)},$$

where X_{kt} is the number of confirmed cases in region k at time-point t and D is the lag examined each time.

Another extension would be to consider different coefficient for each lag:

$$E(Y_{kt}) = e_k v_{kt} + \lambda_t Y_{k,t-1} + \phi_t \sum_{q \neq k} w_{qt} Y_{q,t-1}$$

$$\log(v_{kt}) = \alpha^{(v)} + \beta_d^{(v)} \sum_{d=1}^{D} X_{k,t-d},$$

$$\log(\lambda_t) = \alpha^{(\lambda)},$$
$$\log(\phi_t) = \alpha^{(\phi)}$$

Moreover, a model which takes into account only the respective lag D (or lags) and not the sum of the last D days could be examined.

This model would be:

$$E(Y_{kt}) = e_k v_{kt} + \lambda_t Y_{k,t-1} + \phi_t \sum_{q \neq k} w_{qt} Y_{q,t-1}$$

$$\log(v_{kt}) = \alpha^{(v)} + \beta_i^{(v)} X_{k,t-d_i},$$

$$\log(\lambda_t) = \alpha^{(\lambda)},$$
$$\log(\phi_t) = \alpha^{(\phi)},$$

where d_i are the selected lags and β_i are the respective coefficients.

2.1.4 Extension 3 - Seasonality

Contagious diseases are strongly affected by seasonality and climate conditions in general. In COVID-19, a significant correlation between seasons and infections has been observed (Liu et al. 2021, Sajadi et al. 2020). The inclusion of seasonality with respect of sine-cosine terms in the model is therefore suggested. Since seasonality could be part of all three terms, we investigate all possible models. If seasonality is not computed in the log-linear predictors, then their simple form is considered. The respective model with seasonality terms has been exhaustively described in (Held & Paul 2012).

Below, we present the scenario that seasonality terms exist in all three log-linear predictors. Nevertheless, all combinations are possible.

Seasonality in all components:

$$E(Y_{kt}) = e_k v_t + \lambda_t Y_{k,t-1} + \phi_t \sum_{q \neq k} w_{qt} Y_{q,t-1}$$

$$\begin{aligned}
\log(v_t) &= \alpha^{(v)} + \gamma^{(v)} \sin(\omega t) + \delta^{(v)} \cos(\omega t), \omega = \tfrac{2\pi}{365}, \\
\log(\lambda_t) &= \alpha^{(\lambda)} + \gamma^{(\lambda)} \sin(\omega t) + \delta^{(\lambda)} \cos(\omega t), \omega = \tfrac{2\pi}{365}, \\
\log(\phi_t) &= \alpha^{(\phi)} + \gamma^{(\phi)} \sin(\omega t) + \delta^{(\phi)} \cos(\omega t), \omega = \tfrac{2\pi}{365},
\end{aligned}$$

where $\gamma^{(\phi)}$ and $\delta^{(\phi)}$ are unknown parameters and $\omega = \frac{2\pi}{freq}$ are Fourier frequencies ($freq$ = 365 for daily data)

2.1.5 Region-specific intercepts

As it was discussed before, intercepts in the log-linear predictors can be fixed or region-specific, accounting for heterogeneity between regions (Ibañez et al. 2021). For example, in the endemic component in which we include information for the population of the provinces, by considering region-specific intercepts we assume that for each province, its population has a different effect on the number of deaths reported there. The same interpretation applies to the other terms too. Region-specificity may appear in one or more terms.

The region-specific intercepts can be treated as fixed (e.g.,., $log(v_k) = \alpha_k$) or random effects (e.g.,., $log(v_k) = \alpha_0 + \alpha_k$). In the latter case, random effects (α_k) are assumed to be independent and identically distributed across k. However they can be correlated across the model components, following a normal distribution:

$$\alpha_k := (\alpha_k^{(v)}, \alpha_k^{(\lambda)}, \alpha_k^{(\phi)}) \sim N((0,0,0)^T, \sum_\alpha)$$

Below, region-specific extension, with region specificity in all terms in the simple version of the endemic-epidemic are presented.

$$E(Y_{kt}) = e_k v_{kt} + \lambda_t Y_{k,t-1} + \phi_t \sum_{q \neq k} w_{qt} Y_{q,t-1}$$

$$\begin{aligned} \log(v_{kt}) &= \alpha_k^{(v)}, \\ \log(\lambda_{kt}) &= \alpha_k^{(\lambda)}, \\ \log(\phi_{kt}) &= \alpha_k^{(\phi)}. \end{aligned}$$

Region-specific intercepts may also be considered in all possible combinations of those three terms. The same procedure is followed when region-specificity is applied to the extensions of the endemic-epidemic framework.

2.1.6 Implementation in R

The R package surveillance is broadly used for spatiotemporal modeling and the monitoring of epidemic phenomena (Höhle 2007). The package offers three endemic-epidemic modeling frameworks with tools for visualization, likelihood inference, and simulation. *hhh4()* (Paul & Meyer 2016, Meyer et al. 2016), fits a Poisson or Negative binomial distribution and estimates models for count time series. The function *oneStepAhead()* is used to compute the short-term predictions, while *hhh4addon()* is an add-on package to the surveillance package, which extents some features of the EE model. In general, with *hhh4()* fixed or random effects for the log-linear predictors can be considered. Moreover, the inclusion of seasonality or any linking covariates can occur. For the estimation of the maximum likelihood parameters, the quasi-Newton algorithm is utilized through the *nlminb()* or the *optim()*. The second endemic-epidemic framework is provided by *twinstim()* (Meyer et al. 2017), which estimates self-exciting point process models for a spatiotemporal point pattern of infective events (Meyer et al. 2012). Lastly, the third option is the *twinSIR()* for the susceptible-infectious-recovered (SIR) event history of a fixed population (Höhle 2009).

2.2 Finite Mixture Endemic Epidemic Model

2.2.1 What is the EM algorithm?

An Expectation–Maximization (EM) algorithm is an iterative method that attempts to find maximum likelihood estimates of parameters in statistical models (Dempster et al. 1977). Usually, when we want to find the maximum likelihood estimate (MLE) of θ, we use computational methods to calculate the θ that maximizes the likelihood function. It is also common that log-likelihood instead of the likelihood function will be maximized, since it is easier to be solved and log is a monotonically increasing function and thus

has the same solution. However, this procedure isn't always trivial and difficulty to compute the maximum value of the function arises, either using the likelihood or the log-likelihood function due to analytical, computational or both, difficulties. At this point, the EM algorithm provides a valuable solution to this problem.

In addition, the EM algorithm is broadly used in problems involving missing data or incomplete information. The terms "missing data" or "incomplete information" are used in a more wide way and represent mixtures, convolutions, random effects, censoring, truncated observations, missing data or grouping, among other schemes. In any case, the goal of the EM algorithm is to estimate the unknown parameters. The algorithm, basically converts the incomplete-data problem to a complete-data problem and therefore a problem which is easier to be solved (Ng et al. 2012).

Briefly, the way that the EM algorithm works is to make guesses about the complete data X and to solve for the θ that maximizes the log-likelihood of X over θ. Then, this estimation of θ is used to make a better guess about the complete data. Practically, the first step (Expectation step or E-step) computes the conditional expected value of the complete data log-likelihood and the second step (Maximization step or M-step) maximizes this expected value with respect to model parameters. This strategy is iterated until convergence. What the EM algorithm requires, is some observed data Y, a density function $P(Y|\theta)$ and a description of the incomplete data X (Gupta et al. 2011, Roick et al. 2021).

The EM algorithm will find a maximum of the likelihood function but in case there are multiple peaks, there is no guarantee that the algorithm will find the global maximum (Celeux & Govaert 1992). The peak that will be given depends on the initial values that the algorithm will start with. It is important thus, to give the appropriate initial values. Typically, the EM algorithm can start several times with different initial values and eventually the ones that give the largest likelihood in the last iteration should be chosen.

In general, convergence is fast and under general conditions the algorithm provides reliable results. Aitken's acceleration method (Aitken 1926) is utilized to determine the convergence of the algorithm. The method estimates the asymptotic maximum log-likelihood at each iteration of the algorithm. The acceleration at iteration i is given by:

$$\alpha^{(i)} = \frac{l^{(i+1)} - l^{(i)}}{l^{(i)} - l^{(i-1)}}$$

where $l^{(i)}$ is the log-likelihood value at iteration i.

The asymptotic estimate of the log-likelihood at iteration $i+1$ is the following:

$$l_{\infty}^{(i+1)} = l^{(i)} + \frac{1}{1-\alpha^{(i)}}(l^{(i+1)} - l^{(i)}$$

The Aitken's acceleration method decides whether the algorithm has converged. There are several stopping criteria such as the stopping criterion described in Lindsay (1995) that suggests that convergence is reached when:

$$l_{\infty}^{(i+1)} - l^{(i+1)} < \epsilon$$

where ϵ is a small value, for example $10^{(-10)}$

Another stopping criterion is the one proposed by McNicholas et al. (2010) and suggests convergence when

$$l_{\infty}^{(i+1)} - l^{(i+1)} \in (0, \epsilon)$$

where, again ϵ is a small value.

2.2.2 Proposed methodology

The purpose of the inclusion of the EM algorithm in the model is to create a finite mixture model to cluster regions. In our case, we have the number of deaths for K (K=12) provinces reported every day and we are interested in grouping those K provinces in G clusters based on their time series of the number of deaths. The unobserved data correspond to the unknown group membership labels $Z_1, ..., Z_K$ where $Z_k = (Z_{k1}, ...Z_{kG})$ and Z is an indicator function:

$$Z_{kg} = \begin{cases} 1, & \text{if } k \text{ belongs to group } g, \\ 0, & \text{otherwise} \end{cases} \qquad g = 1, ..., G,\ k = 1, ..., K$$

where $\bar{Y_i}$ are the observed data, namely the time series of the number of deaths for each day t, for province k.

Without the incorporation of the EM algorithm into the EE framework, we have that the number of deaths follows the Negative binomial distribution:

$$Y_{kt} \sim NegBin(\mu_{kt}, \psi)$$

where μ_{kt} is the mean value of deaths in province k at time t ($E(Y_{kt})$) and ψ is the overdispersion parameter.

Each region contributes to the log-likelihood:

$$\prod_{t=2}^{T} P(Y_{kt}; \mu_{kt}, \psi)$$

And for all provinces the log-likelihood becomes:

$$\prod_{k=1}^{K} \prod_{t=2}^{T} P(Y_{kt}; \mu_{kt}, \psi)$$

When we use the finite mixture form of the endemic-epidemic model, we have that each group g follows a distinct Negative binomial distribution with different formulation for the mean value $\mu_{kt}^{(g)}$ and different overdispersion parameter $\psi^{(g)}$ for each group. Therefore, assuming that we have G groups:

$$Y_{kt}|(Z_k = g) \sim NegBin(\mu_{kt}^{(G)}, \psi^{(G)})$$

where $g = 1, ..., G$ denotes the cluster.

The probability mass function (pmf) is now given by:

$$P(Y_{kt} = y_{kt}) = \sum_{g=1}^{G} p_g P_g(y_{kt})$$

where $P_g(y_{kt})$ is the pmf of a $NegBin(\mu_{kt}, \psi)$ and

$$\sum_{g=1}^{G} p_g = 1$$

Now, each province contributes to the log-likelihood:

$$L_k(\Theta) = \prod_{t=2}^{T} \sum_{g=1}^{G} p_g L_{kg}(\Theta_g)$$

where $L_{kg}(\Theta_g) = NegBin(\mu_{kt}^{(g)}, \psi^{(g)})$ for a specific province k and group g.

The total log-likelihood, thus becomes:

$$\prod_{k=1}^{K} \prod_{t=2}^{T} \sum_{g=1}^{G} p_g NegBin(\mu_{kt}^{(g)}, \psi^{(g)})$$

In the simple endemic-epidemic model, the mean value is calculated through:

$$E(Y_{kt}) = e_k v_t + \lambda_t Y_{k,t-1} + \phi_t \sum_{q \neq k} w_{qt} Y_{q,t-1}$$

$$\begin{aligned} \log(v_t) &= \alpha^{(v)}, \\ \log(\lambda_t) &= \alpha^{(\lambda)}, \\ \log(\phi_t) &= \alpha^{(\phi)}. \end{aligned}$$

The parameters to be estimated are four. Three parameters correspond to the endemic, autoregressive and spatiotemporal term and the last parameter concerns overdispersion.

In the simple region-specific endemic-epidemic model, where each province is considered unique, the mean value is calculated through:

$$E(Y_{kt}) = e_k v_{kt} + \lambda_{kt} Y_{k,t-1} + \phi_{kt} \sum_{q \neq k} w_{qt} Y_{q,t-1}$$

$$\begin{aligned} \log(v_{kt}) &= \alpha_k^{(v)}, \\ \log(\lambda_{kt}) &= \alpha_k^{(\lambda)}, \\ \log(\phi_{kt}) &= \alpha_k^{(\phi)}. \end{aligned}$$

Assuming 12 provinces, the parameters to be estimated are 37. Twelve parameters for each component; the endemic, autoregressive and spatiotemporal; and the last parameter concerns overdispersion.

In the extended proposed simple model, if we assume that we have G groups, there are $4 \times G - G \times (G-1)$ parameters to be estimated. The four estimates, as in the simple model, for each one of the G clusters, plus $G \times (G-1)$ probabilities.

At first, some initial values for the probabilities and the rest parameters to be estimated are given. The expectation step computes the weights and the maximization step solves G weighted log-likelihood problems. More specific, in the E-step, Z values (weights) are updated using their conditional expected values:

$$Z_{kg} = \frac{p_g L_{kg}(\Theta_g)}{\sum_{g=1}^{G} p_g L_{kg}(\Theta_g)} = \frac{p_g L_{kg}(\Theta_g)}{L_k(\Theta)}$$

Then, M-step maximizes the expected value of the complete-data log-likelihood. In each iteration, the mixing proportions are first updated by means of the following formula:

$$p_g = \frac{k_g}{K}$$

where K is the total number of provinces, $g = 1, ..., G$ and

$$k_g = \sum_{k=1}^{K} Z_{kg}$$

Then the weighted log-likelihood is maximized using the *optim()* function in R and the procedure that was discussed in the previous subsection is being followed.

The log-likelihood to be maximized is given below:

$$L^g(\Theta) = \sum_{k=1}^{K} Z_{kg} L_{kg}(\Theta_g)$$

We repeat the EM algorithm for different numbers of groups G and for various initial values and in the end we select the number of clusters and the initial values that provide the largest likelihood in the final iteration. The output of the algorithm includes the estimates of parameters, the mixing proportions, the final log-likelihood and the weights.

In our application, the EM algorithm is initially incorporated in the simple endemic-epidemic model. Afterwards, for each of the discussed extensions, the best model, according to some model selection criteria, is selected to be upgraded into a finite mixture model.

2.3 Model selection criteria

Model selection is the task of selecting a statistical model from a set of candidate models. Especially when candidate models are not nested, information criteria provide useful insight. Some of them are Akaike information criterion (AIC) (Bozdogan 1987, Burnham & Anderson 2004), Bayesian information criterion (BIC) (Burnham & Anderson 2004), Bridge criterion (BC) (Ding et al. 2017), Deviance information criterion (DIC) (Spiegelhalter et al. 2014, Van Der Linde 2005), Focused information criterion (FIC) (Claeskens & Hjort 2003, Behl et al. 2012), Hannan–Quinn information criterion (HQC) (Hannan & Quinn 1979), Minimum description length (MDL) (**??**), etc. In case that parameters are estimated through maximum likelihood (ML), which is our case too, the procedure of model selection goes as follows. The candidate models are defined and their parameters are estimated by means of maximum likelihood. Then, using those estimations, a log-likelihood-based score is calculated for each candidate model. Information criteria, by including information related to those models, provide eventually a score for each model. For the computation of that score, the most popular criteria use the number of parameters of each model (k), the log-likelihood (LL) and the sample size (n) of the model. The best model is supposed to be the one with the lower score. Some basic characteristics of the information criteria are that, when log-likelihood increases, the score decreases, while the score decreases when more parameters are added to the model, considering a penalty for complexity. Appropriate measures for the selection of the best time series model are Akaike information criterion (AIC), Bayesian Information Criterion (BIC) and Hannan-Quinn Information Criterion (HQIC).

The most common information criterion to estimate the quality of a model, is the Akaike information criterion (AIC) (Bozdogan 1987). AIC, based on information theory, estimates the relative amount of information that is lost from the given model. Therefore, the selection of the ”best” model becomes a minimization problem and according to the AIC, the ”best” model will be the one that neither over-fits nor under-fits. The formula of AIC is:

$$AIC = 2k - 2\log L$$

where k is the number of parameters and L is the likelihood of the respective model.

AIC's assumptions are that compared models use the same data, the outcome variable is the same across them and that the sample size is large enough.

After Akaike information criterion, many alternative criteria were formulated to deal with various problems. For example, for small sample sizes the corrected Akaike information criterion (AICc) (Brewer et al. 2016) gives a more accurate solution to the model selection problem. In particular, AICc is usually used when the ratio sample size over the number of parameters is less that 40 and is calculated through:

$$AICc = -2(\log L) + 2k + \frac{2k(k+1)}{n-k-1}$$

which is equal to:

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

where k is the number of parameters and n is the total sample size.

The AIC and AICc criteria can also be modified and adjust for overdispersion ψ. Those criteria are called Quasi-AIC (QAIC) and Quasi-AICc (QAICc) respectively (Richards 2008). Below, the updated formulas are presented:

$$QAIC = 2k - \frac{2\log L}{\psi}$$

$$QAICc = QAIC + \frac{2k(k+1)}{n-k-1}$$

where $\psi > 1$ is the overdispersion parameter, k is the number of parameters and n is the total sample size.

The second most commonly used information criterion is the Bayesian Information Criterion (BIC) (Burnham & Anderson 2004) or the Schwarz information criterion (SIC). While closely related, BIC penalizes the model complexity much more than Akaike's criteria.

The BIC for a given model is:

$$BIC = -2\log L + k\log(n)$$

where L is the maximized value of the likelihood function of the model, k is the number of parameters in the model and n is the sample size.

The score should again be minimized. Moreover, in order to use BIC, the sample size n should be much larger than the number of parameters k and the outcome variable should be the same in all candidate models.

Another variation of the AIC is the consistent Akaike's Criterion Information (CAIC) (Anderson et al. 1998), which extends BIC by considering an additional penalty for more parameters in the model.

CAIC is defined as:

$$CAIC = -2 + k(\log(n) + 1) = BIC + k$$

where k is the number of parameters and n is the number of observations.

Another statistical measure for the comparative evaluation among time series models is Hannan-Quinn Information Criterion (HQIC or HQC) (Maïnassara & Kokonendji 2016). HQIC provides a smaller penalty than BIC, but larger than AIC and is given by the following equation:

$$HQIC = -2 \log L + 2k \log(\log(n))$$

where log is natural logarithm, n is the number of data, L is the maximum likelihood of the model and k is the number of parameters that have to be estimated in the model.

Model	Notation
Simple	$\mathsf{E}(Y_{kt}) = e_k v_{kt} + \lambda_{kt} Y_{k,t-1} + \phi_{kt} \sum_{q \neq k} w_{kq} Y_{q,t-1}$ $\log(v_t) = \alpha^{(v)}$, $\log(\lambda_t) = \alpha^{(\lambda)}$, $\log(\phi_t) = \alpha^{(\phi)}$
Order D (same coef.)	$\mathsf{E}(Y_{kt}) = e_k v_t + \lambda_t \sum_{d=1}^{D} Y_{k,t-d} + \phi_t \sum_q \sum_{d=1}^{D} w_{qt} Y_{q,t-d}$ $\log(v_t) = \alpha^{(v)}$, $\log(\lambda_t) = \alpha^{(\lambda)}$, $\log(\phi_t) = \alpha^{(\phi)}$
Order D (different coef.)	$\mathsf{E}(Y_{kt}) = e_k v_t + \sum_{d=1}^{D} \lambda_{td} Y_{k,t-d} + \sum_q \sum_{d=1}^{D} \phi_{td} w_{qt} Y_{q,t-D}, d = 1, 2, ...D$ $\log(v_t) = \alpha^{(v)}$, $\log(\lambda_{td}) = \alpha_d^{(\lambda)}$, $\log(\phi_{td}) = \alpha_d^{(\phi)}$
Covariate-cases (sum & same coef.)	$\mathsf{E}(Y_{kt}) = e_k v_{kt} + \lambda_t Y_{k,t-1} + \phi_t \sum_q w_{qt} Y_{q,t-1}$ $\log(v_{kt}) = \alpha^{(v)} + \beta^{(v)} \sum_{d=1}^{D} X_{k,t-d}$, $\log(\lambda_t) = \alpha^{(\lambda)}$, $\log(\phi_t) = \alpha^{(\phi)}$
Covariate-cases (sum & different coef.)	$\mathsf{E}(Y_{kt}) = e_k v_{kt} + \lambda_t Y_{k,t-1} + \phi_t \sum_{q \neq k} w_{qt} Y_{q,t-1}$ $\log(v_{kt}) = \alpha^{(v)} + \beta_d^{(v)} \sum_{d=1}^{D} X_{k,t-d}$, $\log(\lambda_t) = \alpha^{(\lambda)}$, $\log(\phi_t) = \alpha^{(\phi)}$
Covariate-cases (specific lags)	$\mathsf{E}(Y_{kt}) = e_k v_{kt} + \lambda_t Y_{k,t-1} + \phi_t \sum_{q \neq k} w_{qt} Y_{q,t-1}$ $\log(v_{kt}) = \alpha^{(v)} + \beta_i^{(v)} X_{k,t-d_i}$, $\log(\lambda_t) = \alpha^{(\lambda)}$, $\log(\phi_t) = \alpha^{(\phi)}$
Seasonality	$\mathsf{E}(Y_{kt}) = e_k v_t + \lambda_t Y_{k,t-1} + \phi_t \sum_{q \neq k} w_{qt} Y_{q,t-1}$ $\log(v_t) = \alpha^{(v)} + \gamma^{(v)} \sin(\omega t) + \delta^{(v)} \cos(\omega t), \omega = \frac{2\pi}{365}$, $\log(\lambda_t) = \alpha^{(\lambda)} + \gamma^{(\lambda)} \sin(\omega t) + \delta^{(\lambda)} \cos(\omega t), \omega = \frac{2\pi}{365}$, $\log(\phi_t) = \alpha^{(\phi)} + \gamma^{(\phi)} \sin(\omega t) + \delta^{(\phi)} \cos(\omega t), \omega = \frac{2\pi}{365}$

Table 2.1: Summary table with all models

Chapter 3

Data and Results

3.1 Data

The data used for this study are daily reported deaths occurred in the provinces of The Netherlands. The Netherlands can be organised in 380 municipalities or in 12 provinces; Drenthe, Gelderland, Groningen, Flevoland, Friesland, Limburg, North-Brabant, North-Holland, Overijssel, Utrecht, Zeeland and South-Holland. The provinces represent the administrative layer between the national government and the local municipalities and have responsibilities with respect to sub-national and regional importance. Other partitions of The Netherlands could be security regions, which are 25 in total or the 11 regional organizations of acute care (ROAZ). Figure 3.1 shows the spatial location of the provinces of The Netherlands and Figure 3.2 provides information for the distribution of population ($\times 100.000$ people) in those 12 provinces. The most populous and the most densely populated province is South-Holland with over 3.7 million inhabitants in 2021. The less populous is Zeeland, while the least densely populated province is Drenthe.

According to the National Institute for Public Health and the Environment, by the end of 2020 there were 11389 deaths in The Netherlands, attributed to coronavirus disease and 794604 confirmed cases.

From the beginning of 2021 until 28th of April, 2021, lockdown was enforced. This means that schools, restaurants and bars, museums, public venues, theaters, etc were closed, while there was a restriction in the number of visitors in houses. Remote work was a strong indication and citizens were urgently advised not to travel until 15th of May, 2021. At times when concern for new variants, such as Omicron and Delta, was raised, flight bans for and from dangerous regions were imposed. Moreover, night-time curfew was forced from 23rd of January until 28th of April. Some levels of education reopened before the end of lockdown at various times. However, from 28th of April, the first big steps for reopening started to take place and until 26th of June almost all measures had been lifted. Some restrictions for restaurants, cafes, bars etc, took effect again on 10th of July, because of the increase of new infections.

Table 3.1 provides information related to the population of each province and the mean daily incidence (per 100 people) of deaths and cases, in province-level, caused by COVID-19 pandemic in the Netherlands during the first eight months of 2021. This

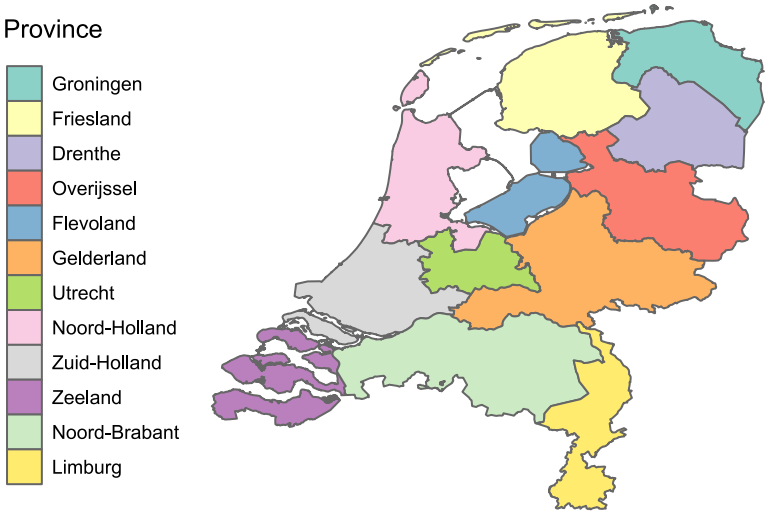

Figure 3.1: The 12 provinces of The Netherlands.

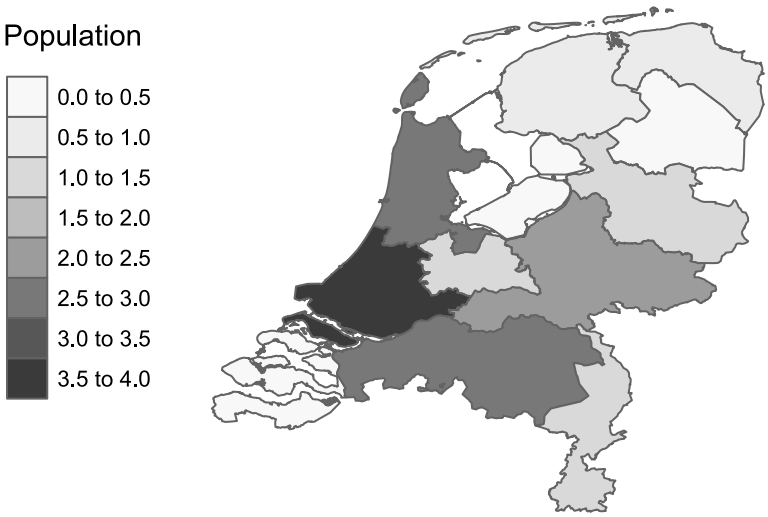

Figure 3.2: Population distribution ($\times 100.000$ people) in the 12 provinces of The Netherlands.

Province	Population	DI of deaths	DI of cases
Drenthe	494771	0.009	1.019
Flevoland	428226	0.004	0.958
Friesland	651435	0.013	1.482
Gelderland	2096603	0.032	5.212
Groningen	586937	0.007	1.35
Limburg	1115872	0.007	3.34
North-Brabant	2573949	0.031	7.544
North-Holland	2888486	0.044	8.185
Overijssel	1166533	0.019	2.961
Utrecht	1361153	0.018	3.413
Zeeland	385400	0.005	0.945
South-Holland	3726050	0.063	10.718

Table 3.1: COVID-19 mean daily incidence (DI) of deaths (per 100 people) and cases (per 100 people) and total population for the 12 provinces of the Netherlands.

information is depicted in Figures 3.3 and 3.4. The highest mean incidence, regarding deaths and cases, is observed in South-Holland, which is also the most populous province of the Netherlands. The next three more populous provinces; North-Holland, North-Brabant and Gelderland, have the highest mean incidence of deaths and cases after South-Holland, indicating that about three to four people are dying daily on average, due to COVID-19.

The data for the population distribution of The Netherlands in 2021 is obtained by Statista Research Department, which is an online portal providing data (https://www.statista.com/). Data for deaths and confirmed cases of coronavirus disease are available for a total of 243 days, from 01/01/2021 until 31/08/2021. It is provided by the National Institute for Public Health and the Environment (RIVM), municipal health services (GGDs) and hospitals, and is accessible to everyone (https://coronadashboard.government.nl/). Since the first confirmed case of SARS-CoV-2 was identified in the province of North-Brabant, data for deaths and cases in municipality, security region, ROAZ region and province level is provided and updated daily. Health policies due to COVID-19 pandemic during 2021 were obtained from the official site of the government of the Netherlands (https://www.government.nl/topics/coronavirus-covid-19/news).

In Figure 3.5 the epidemic curve, which shows the overall daily incidence of cases from coronavirus disease in the Netherlands is plotted, while in Figure 3.6 we can see the daily incidence of deaths attributed to the virus during the same period of time. It is clear that lockdown and curfew measures led to the reduction of cases and deaths, since the epidemic curves fall until last days of May. The lifting of the measures increased again the number of cases and deaths during July but some restrictions that took effect again, eventually controlled the contagion.

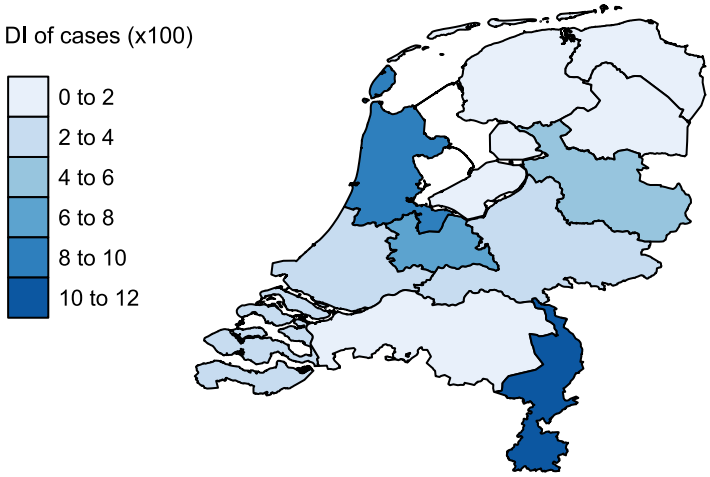

Figure 3.3: Daily incidence of COVID-19 infections in the 12 provinces of The Netherlands.

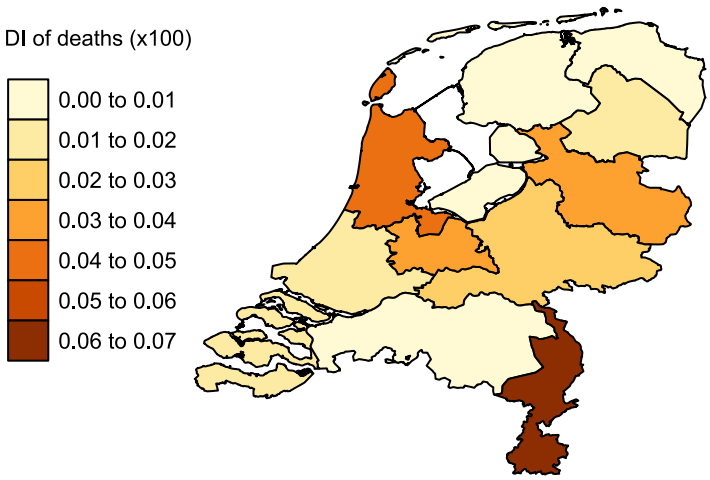

Figure 3.4: Daily incidence of COVID-19 deaths in the 12 provinces of The Netherlands.

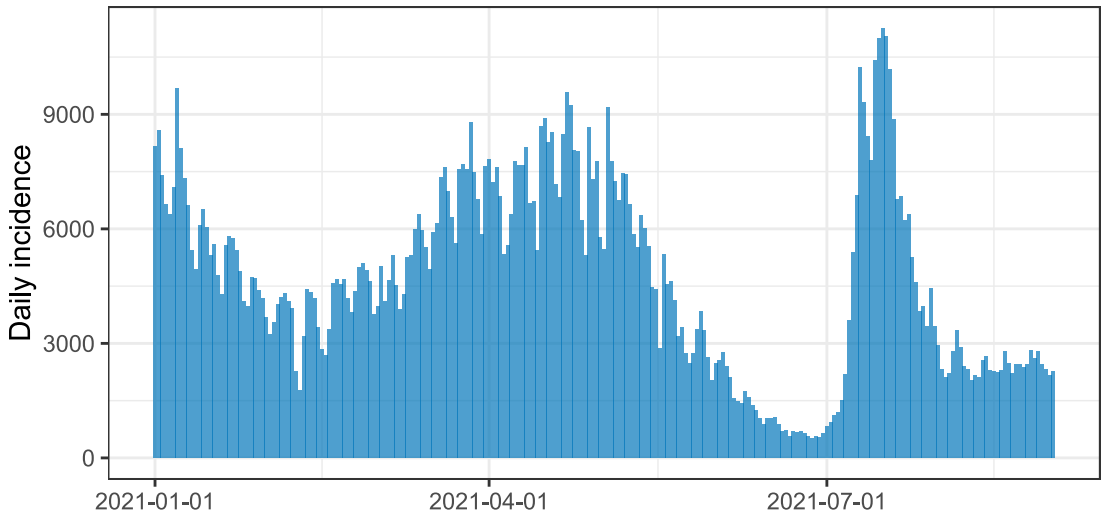

Figure 3.5: Daily incidence of COVID-19 infections in the Netherlands.

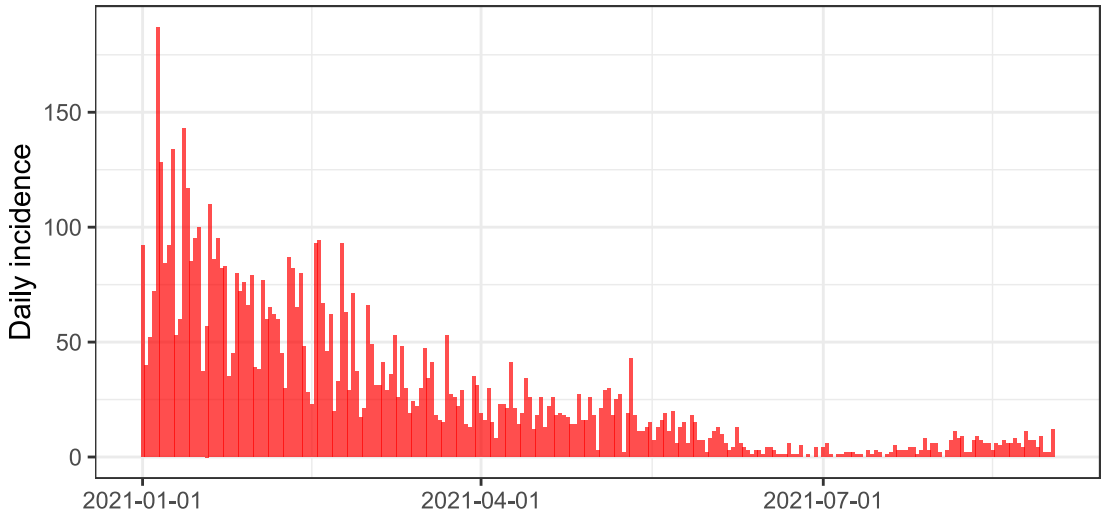

Figure 3.6: Daily incidence of COVID-19 deaths in the Netherlands.

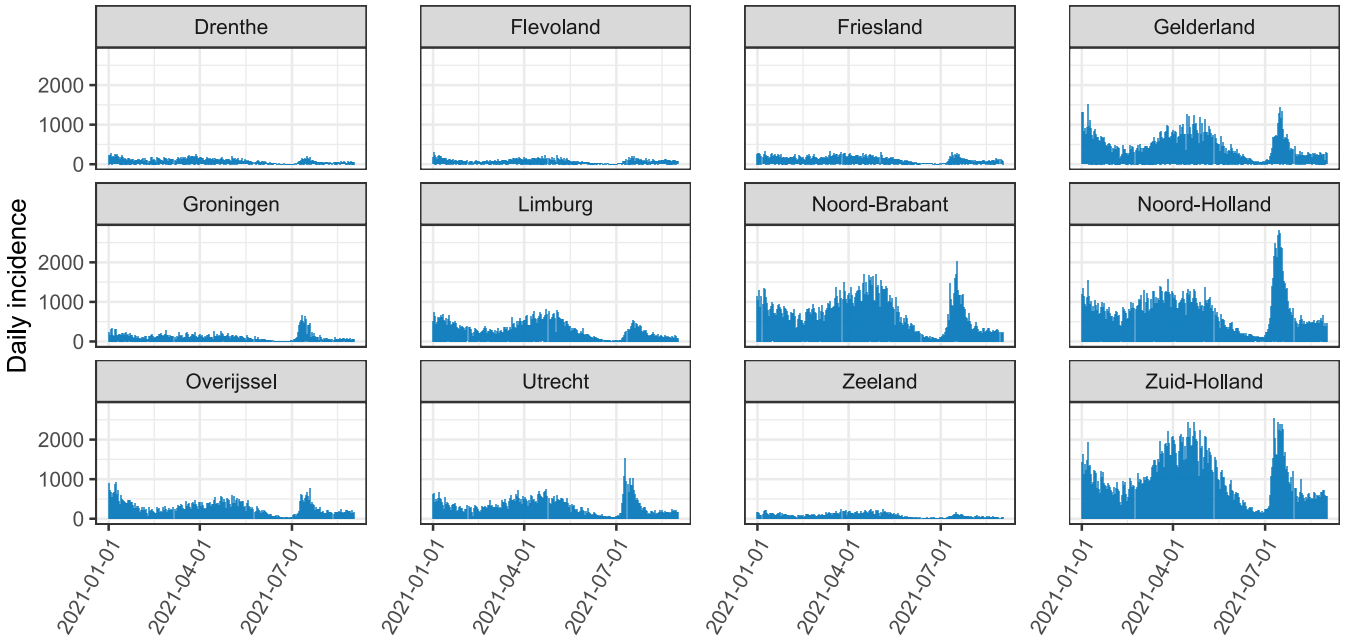

Figure 3.7: Daily incidence of COVID-19 cases in the 12 provinces of the Netherlands.

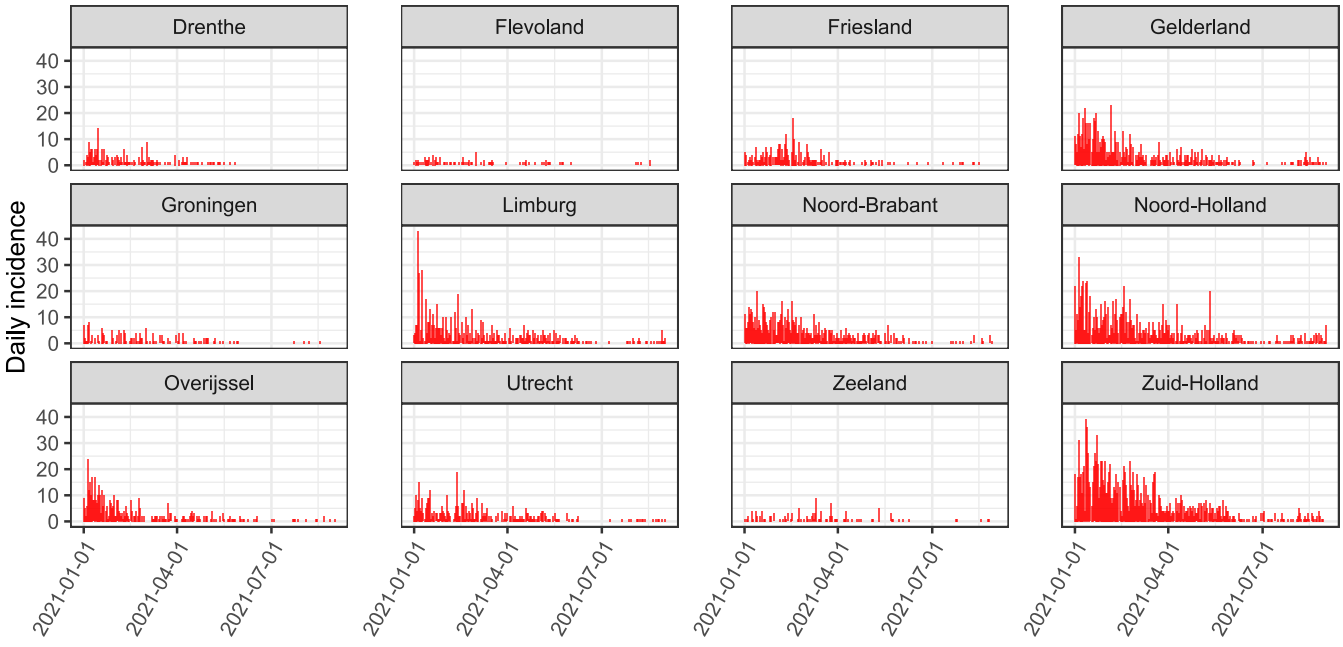

Figure 3.8: Daily incidence of COVID-19 deaths in the 12 provinces of the Netherlands

In Figure 3.7 and Figure 3.8, daily incidence for each province separately, for cases

and deaths respectively, are presented.

	Dr	Fl	Fr	Ge	Gr	Li	N-Br	N-Ho	Ov	Ut	Ze	Z-Ho
Drenthe	0	0	1	0	1	0	0	0	1	0	0	0
Flevoland	0	0	1	1	0	0	0	1	1	1	0	0
Friesland	1	1	0	0	1	0	0	1	1	0	0	0
Gelderland	0	1	0	0	0	1	1	0	1	1	0	1
Groningen	1	0	1	0	0	0	0	0	0	0	0	0
Limburg	0	0	0	1	0	0	1	0	0	0	0	0
North-Brabant	0	0	0	1	0	1	0	0	0	0	1	1
North-Holland	0	1	1	0	0	0	0	0	0	1	0	1
Overijssel	1	1	1	1	0	0	0	0	0	0	0	0
Utrecht	0	1	0	1	0	0	0	1	0	0	0	1
Zeeland	0	0	0	0	0	0	1	0	0	0	0	1
South-Holland	0	0	0	1	0	0	1	1	0	1	1	0

Table 3.2: Adjacency matrix of the 12 provinces of the Netherlands.

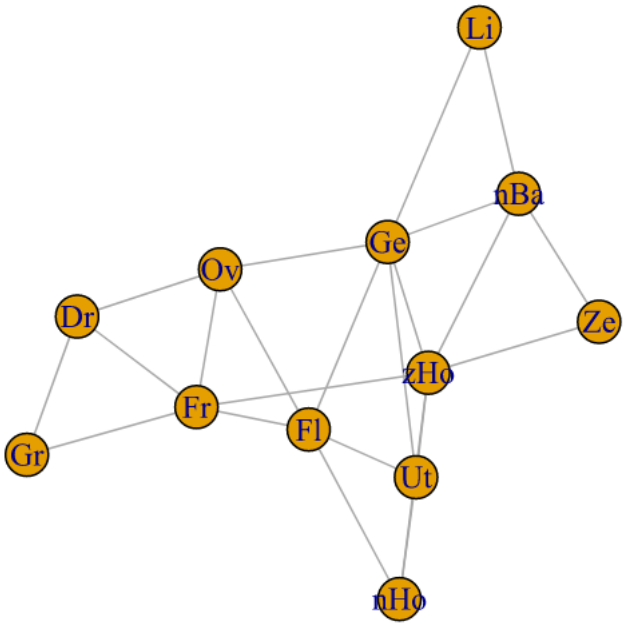

Figure 3.9: Graph obtained by the adjacency matrix for the 12 provinces of the Netherlands.

3.2 Fixed-intercept models

As stated in previous sections, we consider that the number of daily deaths in the province k, on the day t, Y_{kt} is described by the endemic-epidemic framework.The general form of the endemic-epidemic framework is:

$$E(Y_{kt}) = e_k v_{kt} + \lambda_{kt} Y_{k,t-1} + \phi_{kt} \sum_{q \neq k} w_{qt} Y_{q,t-1}$$

In the endemic part in the above equation, we assume that the e_k denotes the amount of population of the province k, which is shown in Table 3.1.

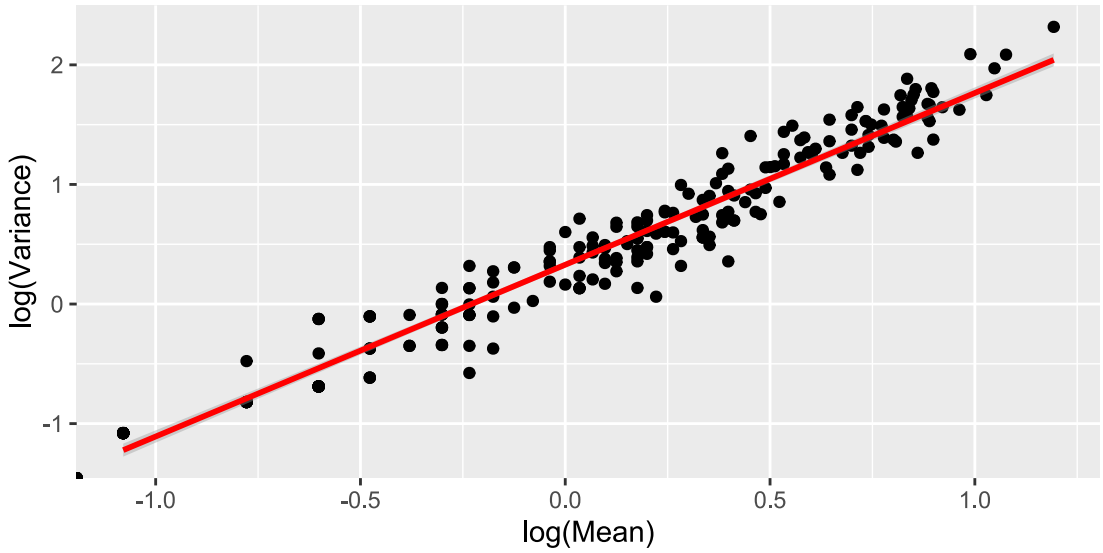

Figure 3.10: Mean vs variance of the daily number of deaths of the 12 provinces of the Netherlands, during 243 days. Number of deaths are indicated in a log scale with black dots. The red line is the smoothing line of the linear model.

As can be seen in Figure 3.10 there is clear overdispersion in the data set, since the variance of the death counts is constantly greater than the mean. Therefore, we will assume that deaths follow Negative binomial distribution and the overdispersion parameter will be considered the same for all provinces. Since all provinces belong to the same country, this is a reasonable assumption.

As presented in Table 3.2 and Figure 3.9, the weight matrix, considered in our application, indicates geographic borders and is equal to 1 if regions have common borders and 0 if there is no adjacency. In the Netherlands, during the first eight months of 2021, there was strong indication not to travel and at some periods flight bans were imposed. Also, until 24th of April, lockdown and curfew were in force. Therefore, we consider that this conception of the weight matrix is representative for the mobility in the Netherlands.

In the following subsections, we will exemplify different formulations of the log-linear predictors of the endemic and epidemic components, in the twelve provinces of the Netherlands.

3.2.1 Simple model

The simple version of the endemic epidemic model assumes fixed intercepts across provinces, for all the log-linear predictors and thus, along with the oversdispersion parameter, four parameters to be estimated.

$$E(Y_{kt}) = e_k v_t + \lambda_t Y_{k,t-1} + \phi_t \sum_{q \neq k} w_{qt} Y_{q,t-1}$$

$$\begin{aligned} \log(v_t) &= \alpha^{(v)}, \\ \log(\lambda_t) &= \alpha^{(\lambda)}, \\ \log(\phi_t) &= \alpha^{(\phi)} \end{aligned}$$

	Model 1	
	Estimate	Standard Error
$\alpha^{(v)}$	-3.354	0.073
$\alpha^{(\lambda)}$	-8.174	-
$\alpha^{(\phi)}$	-1.751	0.045
ψ	0.952	0.023
npar:	4	
Log-likelihood:	-4626.923	
AIC:	9261.846	
BIC:	9285.758	
CAIC:	9289.758	
HQIC:	9270.459	

Table 3.3: Estimations for the four parameters along with their standard errors of the simple model. In addition, model selection criteria and the number of parameters (npar) are provided.

According to the above simple version of the endemic-epidemic model, all of the constituent components indicate relatively weak dependence on the number of daily deaths, with the largest one the spatiotemporal term ($e^{-1.751} = 0.173$).

3.2.2 Extension 1 - Order D

Considering that the number of deaths in the k-th province on day t is not only affected by the deaths occurred at the previous day $t-1$, but there is a lagged effect, we examine the possibility of a delay.

$$E(Y_{kt}) = e_k v_{kt} + \sum_{d=1}^{D} \lambda_{td} Y_{k,t-d} + \sum_{q} \sum_{d=1}^{D} \phi_{td} w_{qt} Y_{q,t-D}, \qquad d = 1, 2, ...D$$

$$\begin{aligned} \log(v_t) &= \alpha^{(v)}, \\ \log(\lambda_{td}) &= \alpha_d^{(\lambda)}, \\ \log(\phi_{td}) &= \alpha_d^{(\phi)} \end{aligned}$$

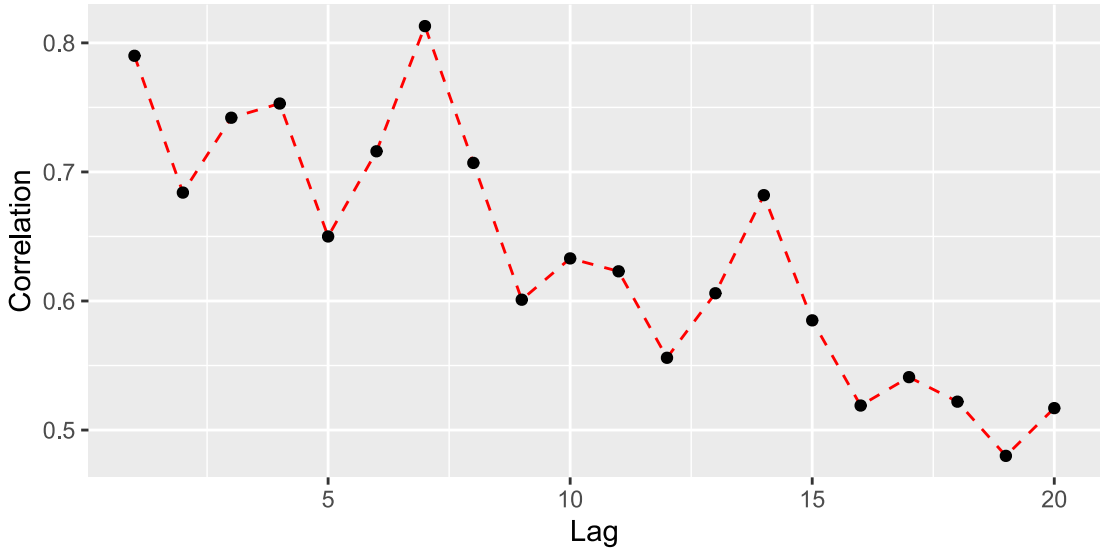

Figure 3.11: Cross-correlogram within the series of daily deaths. It shows the Pearson correlation as a function of the displacement (days) of deaths relative to subsequent deaths.

The cross-correlogram in Figure 3.11, indicates strong correlation between days at time t with days at time $t-1$, which consists the simple model. At the same time, there is even stronger correlation between day t and day $t-7$. We will therefore consider the model that accounts for lag 1 and lag 7.

$$E(Y_{kt}) = e_k v_t + \lambda_{1t} Y_{k,t-1} + \lambda_{2t} Y_{k,t-7} + \phi_{1t} \sum_{q \neq k} w_{qt} Y_{q,t-1} + \phi_{2t} \sum_{q \neq k} w_{qt} Y_{q,t-7}$$

$$\begin{aligned} \log(v_t) &= \alpha^{(v)}, \\ \log(\lambda_{dt}) &= \alpha_d^{(\lambda)}, \\ \log(\phi_{dt}) &= \alpha_d^{(\phi)} \end{aligned}$$

where $d = 1$ or $d = 7$.

	Model 1	
	Estimate	Standard Error
$\alpha^{(v)}$	-4.335	0.0131
$\alpha_1^{(\lambda)}$	-1.351	0.259
$\alpha_2^{(\lambda)}$	-1.083	0.339
$\alpha_1^{(\phi)}$	-4.836	0.008
$\alpha_2^{(\phi)}$	-3.105	0.045
ψ	1.21	0.065
npar:	6	
Log-likelihood:	-4626.923	
AIC:	9299.726	
QAIC:	7687.807	
BIC:	9335.594	
CAIC:	9341.594	
HQIC:	9312.646	

Table 3.4: Estimations for the parameters of the model that accounts for a delay of 1 and 7 days in the effect of deaths. Model selection criteria and the number of parameters to be estimated (npar) are also provided.

In Table 3.4, it is clearly concluded that the autoregressive term has a stronger effect than the endemic and the spatiotemporal term. More precisely, λ_1 is equal to 0.259 ($e^{-1.351}$) and λ_2 is equal to 0.338 ($e^{-1.083}$), which are also very close to each other. Additionally, not only in the autoregressive term, but also in the spatiotemporal, both lag one and lag seven display similar effect on the number of deaths at day t.

We will further examine the scenario that the sum of previous days is taken into account. Specifically, the sum of the deaths of the previous D days will replace the effect of the previous day $t-1$. Possible additive delay of three, four and five days will be examined.

We will call Model 2.1.1 the model that assumes lag of two days (order 2) and has the same coefficient for both days. Same coefficients reflect same impact from those days to day t. Model 2.1.2 assumes that lag-days have different effect on the number of deaths that happen in the same province, but the same effect to other provinces. Model 2.1.3 corresponds to the opposite assumption, while Model 2.1.4 attributes different effects of the days in both components. Model 2.2.1 is the model with three days delay and so on.

Below, the Table 3.5 displays the estimations with their standard errors only for the order-5 scenario, while the rest scenarios are presented in Tables A.1 and 3.5 in the Appendix section.

	Model 2.3.1.	Model 2.3.2.	Model 2.3.3.	Model 2.3.4.
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
$\alpha^{(v)}$	-4.689 (0.155)	-4.682 (0.154)	-4.701 (0.154)	-4.699 (0.155)
$\alpha_1^{(\lambda)}$	-1.899 (0.039)	-1.584 (0.118)	-1.91 (0.039)	-1.58 (0.121)
$\alpha_2^{(\lambda)}$		-2.553 (0.27)		-2.473 (0.251)
$\alpha_3^{(\lambda)}$		-1.67 (0.124)		-1.691 (0.129)
$\alpha_4^{(\lambda)}$		-1.855 (0.146)		-1.92 (0.158)
$\alpha_5^{(\lambda)}$		-2.101 (0.181)		-2.099 (0.182)
$\alpha_1^{(\phi)}$	-5.057 (0.14)	-5.066 (0.14)	-4.658 (0.516)	-4.745 (0.581)
$\alpha_2^{(\phi)}$			-14.839 (58.232)	-11.576 (11.669)
$\alpha_3^{(\phi)}$			-4.725 (0.623)	-4.704 (0.635)
$\alpha_4^{(\phi)}$			-4.199 (0.388)	-4.29 (0.44)
$\alpha_5^{(\phi)}$			-8.71 (26.199)	-6.509 (3.785)
ψ	1.73	1.76	1.74	1.77
npar:	4	8	8	12
Log-likelihood:	-4450.252	-4442.177	-4442.658	**-4435.453**
AIC:	8908.504	8900.354	8901.316	**8894.906**
QAIC:	5152.8	5063.928	5122.503	**5035.811**
BIC:	**8932.416**	8948.178	8949.14	8966.642
CAIC:	**8936.416**	8956.178	8957.14	8978.642
HQIC:	**8917.117**	8917.581	8918.543	8920.746

Table 3.5: Estimates and their standard error (SE), whenever it is available, for the parameters of the order 5 models. In the second part of the table the model selection criteria are provided, while npar denotes the number of parameters to be estimated each time.

After exploring the effect of the sum up to five days before, we notice that the best considered lag is five days. When we consider three or four days delay, the best model selected by the majority of the criteria is the one which assigns different coefficients in both components. In the case of order-five model, three out of five model selection criteria suggest equal effect of those five days. For instance, the 8 extra coefficients, increase the value of BIC almost 35 units, while CAIC increases almost 43 units.

The results of the abovementioned models are presented in Tables A.1, A.2 and 3.5. Without exceptions, in all models the autoregressive parameters indicate the stronger dependence, while the endemic and the spatiotemporal terms present a weak effect. This evidence is reasonable since lockdown and curfew measures prevented the motion between provinces and therefore there isn't strong influence between them. Whilst low, the effect of previous days, within provinces, is inevitable.

Overall, order-5 extension provides the best model according to model selection criteria and will be further analyzed in the upcoming sections.

3.2.3 Extension 2 - Cases as covariate

An important extension of the endemic-epidemic model, is the one that adds explanatory variables. Those variables can be COVID-19 cases, vaccination proportions, measures, etc. For our application we will consider the covariate of new cases in the endemic part. Similarly with the number of deaths, a delay of the effect is a conceivable scenario. In the following models, we examine possible lags of previous days. Model 3.1 assumes lag one, Model 3.2 lag two and so on.

$$E(Y_{kt}) = e_k v_{kt} + \lambda_t Y_{k,t-1} + \phi_t \sum_{q \neq k} w_{qt} Y_{q,t-1}$$

$$\log(v_{kt}) = \alpha^{(v)} + \beta^{(v)} \sum_{d=1}^{D} X_{k,t-d},$$

$$\log(\lambda_t) = \alpha^{(\lambda)},$$
$$\log(\phi_t) = \alpha^{(\phi)}$$

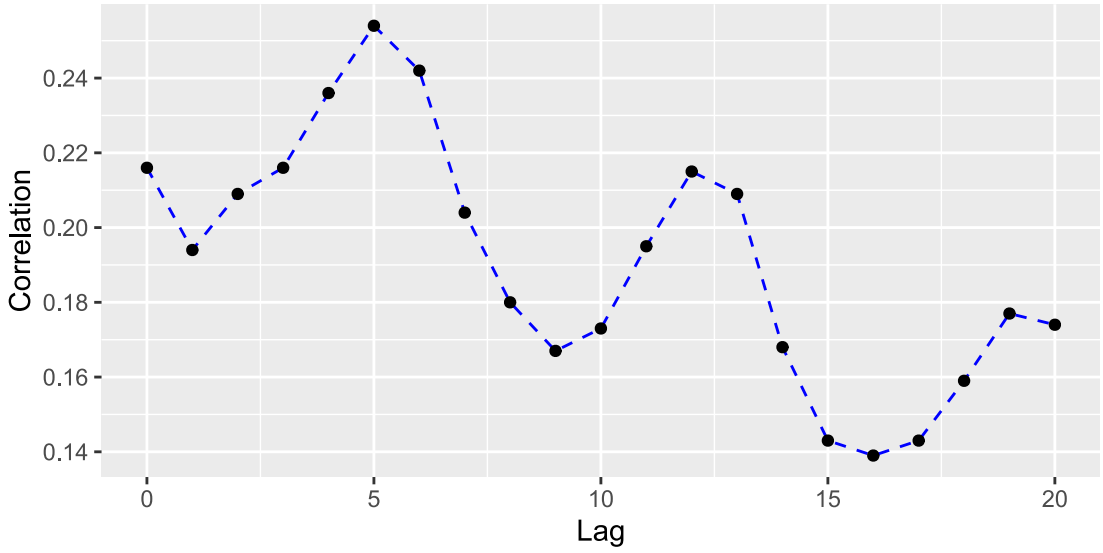

Figure 3.12: Cross-correlogram between daily new deaths and daily new cases. It shows the Pearson correlation between both series as a function of the displacement of daily COVID-19 deaths to daily positive cases

The cross-correlogram in Figure 3.12 provides information related to the lagged effect of newly infections to COVID-19 deaths. In accordance with the cross-correlogram, we would expect that components of the endemic-epidemic model reflect death events affected from infections from further in the past. A reason for this delay is that the reported number of cases is subject to reporting delay or even under-reporting. Another reason is that the virus does not provoke sudden death, but in case of death, it happens some days after the infection. Figure 3.12 suggests that the largest correlation between deaths and cases is when we account for a lag of five days in the number of cases. In our application, we considered six possibilities for the lags. In Table 3.6, the estimates, along with their standard errors are presented. All model selection criteria confirm that lag 5 is the appropriate consideration for the covariate of cases. As in previous models, the autoregressive term, has again the most influencial role in the number of deaths, while the inclusion of cases in the endemic part, increases its effect. Moreover, it is observed that in comparison with the simple model, the selected model from this extension provides a better score of all model selection criteria.

	Model 3.1.	Model 3.2.	Model 3.3.	Model 3.4.	Model 3.5.	Model 3.6.
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
$\alpha^{(v)}$	-15.297 (0.09)	-15.32 (0.09)	-3.819 (0.1)	-3.833 (0.1)	-3.846 (0.1)	-3.851 (0.098)
$\beta^{(v)}$	-8.01 (0.34)	-8.543 (0.29)	-8.897 (0.27)	-9.12 (0.26)	-9.286 (0.24)	-9.438 (0.237)
$\alpha^{(\lambda)}$	-0.766 (0.06)	-0.772 (0.06)	-0.775 (0.06)	-0.778 (0.06)	-0.78 (0.06)	-0.787 (0.066)
$\alpha^{(\phi)}$	-2.542 (0.08)	-2.544 (0.08)	-2.54 (0.08)	-2.54 (0.08)	-2.54 (0.08)	-2.534 (0.078)
ψ	1.33 (0.004)	1.33 (0.004)	1.33 (0.004)	1.33 (0.004)	1.34 (0.004)	1.32 (0.074)
npar:	5	5	5	5	5	5
Log-lik.:	-4622.734	-4621.26	-4620.993	-4619.83	**-4619.365**	-4622.714
AIC:	9255.468	9252.52	9251.986	9249.66	**9248.73**	9255.428
QAIC:	6961.48	6959.263	6958.862	6957.113	**6904.575**	6909.573
BIC:	9285.358	9282.41	9281.876	9279.55	**9278.62**	9285.318
CAIC:	9290.358	9287.41	9286.876	9284.55	**9283.62**	9290.318
HQIC:	9266.235	9263.287	9262.753	9260.427	**9259.497**	9266.195

Table 3.6: Estimates and their standard error (SE) for the parameters of the lagged models with COVID-19 cases as covariate. In the second part of the table the model selection criteria and the amount of parameters (npar) are provided.

3.2.4 Extension 3 - Seasonality

Seasonality has always been a important factor in the transmission of infectious diseases. Several studies indicate that during cold months, deaths attributed to COVID-19 are increased, while during summer they have been showed to be less. In agreement with the above statement, Figure 3.3 and Figure 3.4 display significant high amount of death events during January and February, but very low during summer months, despite

the inflation of the infections during both periods. The respective model notation for this case is the following:

$$E(Y_{kt}) = e_k v_t + \lambda_t Y_{k,t-1} + \phi_t \sum_{q \neq k} w_{qt} Y_{q,t-1}$$

$$\begin{aligned}
\log(v_t) &= \alpha^{(v)} + \gamma^{(v)} sin(\omega t) + \delta^{(v)} cos(\omega t), \\
\log(\lambda_t) &= \alpha^{(\lambda)} + \gamma^{(\lambda)} sin(\omega t) + \delta^{(\lambda)} cos(\omega t), \\
\log(\phi_t) &= \alpha^{(\phi)} + \gamma^{(\phi)} sin(\omega t) + \delta^{(\phi)} cos(\omega t),
\end{aligned}$$

where $\omega = \frac{2\pi}{365}$

The above is the model in the most general form. When we account for seasonality in a log-linear predictor, the respective coefficients $\gamma^{(\cdot)}$ and $\delta^{(\cdot)}$ should be estimated, otherwise they are set to be equal to zero. We will therefore assume that Model 4.1 has seasonality only in the endemic term, Model 4.2 only in the autoregressive, Model 4.3 only in the spatiotemporal part, etc. In Table 3.7, the results of the models are provided. We notice, that again the autoregressive term preforms the largest effect in comparison with the other terms. However its effect is less when seasonality is added to the spatiotemporal term. In this case, spatiotermporal term indicates strong dependence to the number of deaths. It is also noticed that when models have seasonality terms in the endemic term, the coefficient of the spatiotemporal part is tremendously decreased - standard errors are quite big, though. At the same time, when we consider seasonality in the endemic term, the overdispersion parameter increases markedly. According to BIC and CAIC, the consideration of seasonality terms in the endemic and in the spatiotemporal part, provides the best model, while AIC, QAIC and HQC select the model with seasonality terms in all log-linear predictors as the best model. Eventually, after adjusting for seasonality, despite the four additional parameters, AIC, BIC, QAIC and CAIC have a decrease of about 700 units each.

	Model 4.1.	Model 4.2.	Model 4.3.	Model 4.4.	Model 4.5.	Model 4.6.	Model 4.7.
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
$\alpha^{(v)}$	-2.473 (0.04)	-3.425 (0.07)	-3.526 (0.07)	-2.588 (0.05)	-2.488 (0.04)	-3.421 (0.07)	-2.578 (0.05)
$\beta^{(v)}$	0.41 (0.05)			0.674 (0.07)	0.313 (0.06)		0.541 (0.08)
$\gamma^{(v)}$	1.355 (0.04)			1.162 (0.05)	1.468 (0.04)		1.31 (0.06)
$\alpha^{(\lambda)}$	-1.691 (0.12)	-1.304 (0.16)	-3.441 (0.22)	-1.505 (0.16)	-1.765 (0.13)	-1.132 (0.14)	-1.52 (0.14)
$\beta^{(\lambda)}$		0.421 (0.17)		-1.292 (0.38)		0.071 (0.17)	-1.084 (0.37)
$\gamma^{(\lambda)}$		0.917 (0.13)		0.928 (0.28)		0.832 (0.12)	0.712 (0.27)
$\alpha^{(\phi)}$	-13.17(45.56)	-2.8 (0.09)	0.952 (0.22)	-12.38(30.84)	-24.67(10.93)	-3.784 (0.24)	-27.38 (12.4)
$\beta^{(\phi)}$			0.986 (0.16)		16.696 (8.1)	1.378 (0.27)	18.731 (9.25)
$\gamma^{(\phi)}$			-0.893 (0.07)		-14.43 (7.59)	0.593 (0.16)	-16.2 (8.51)
ψ	2.376 (0.01)	1.447 (0.005)	1.361 (0.004)	2.433 (0.01)	2.453 (0.01)	1.465 (0.005)	2.502 (0.01)
npar:	6	6	6	8	8	8	10
L.L.:	-4285.856	-4580.73	-4582.11	-4277.271	-4266.355	-4551.753	**-4260.53**
AIC:	8583.712	9173.46	9176.22	8566.542	8548.71	9119.506	**8541.06**
QAIC:	3619.623	6343.348	6745.446	3532.047	3494.479	6229.997	**3425.699**
BIC:	8619.58	9209.328	9212.088	8618.366	**8596.534**	9167.33	8600.84
CAIC:	8625.58	9215.328	9218.088	8626.366	**8604.534**	9175.33	8610.84
HQC:	8596.632	9186.38	9189.14	8587.769	8565.937	9136.733	**8562.594**

Table 3.7: Estimates and their standard error (SE) for the parameters of all possible models that account for seasonality. In the second part of the table the model selection criteria and the number of parameters (npar) are provided. L.L symbolizes the log-likelihood.

3.3 Region-specific intercept models

Given the assumption of different characteristics between provinces, the adoption of region-specific parameters is fundamental to manage the COVID-19 pandemic. In the following subsections, we use the best model of each extension in order to examine the scenario in which every region requires unique manipulation.

3.3.1 Simple model

Firstly, region-specific parameters are examined in the simple version of the endemic-epidemic model.

	Endemic	Aut.	Spat.	End.&Aut.	End.&Spat.	Aut.&Spat.	All
npar:	15	15	15	26	26	26	37
LL:	-4584.434	-4583.505	-4460.234	-4551.702	-4444.643	-4452.499	**-4435.875**
AIC:	9198.868	9197.01	8950.468	9155.404	**8941.286**	8956.998	8945.75
QAIC:	6588.561	6431.543	5047.136	6129.039	4965.923	4933.331	**4864.362**
BIC:	9288.538	9286.68	9040.138	9310.831	9096.713	**8936.91**	9166.935
CAIC:	9303.538	9301.68	9055.138	9336.831	9122.713	**8940.91**	9203.935
HQIC:	9231.169	9229.311	8982.769	9211.392	8997.274	**8921.611**	9025.425

Table 3.8: Model selection criteria scores for each simple model when region-specific parameters appear in the Endemic (End.), Autoregressive (Aut.), Spatiotemporal (Spat.) or in combinations of those three. L.L symbolizes the log-likelihood and npar is the number of parameters

In the appendix section, each one of the Tables A.3 - A.9 considers region-specific parameters in different combinations of the three possible components. According to Table 3.8, BIC, CAIC and HQIC conclude region-specific parameters in the epidemic component (both autoregressive and spatiotemporal terms) provide the best model. This statement is reasonable since the spatial location, which is reflected in the spatiotemporal term, and the autoregressive behaviour, which is possibly influenced by the number of deaths, present substantial heterogeneity between provinces.

Therefore, the selected model is:

$$E(Y_{kt}) = e_k v_{kt} + \lambda_{kt} Y_{k,t-1} + \phi_{kt} \sum_{q \neq k} w_{qt} Y_{q,t-1}$$

$$\begin{aligned} \log(v_t) &= \alpha^{(v)}, \\ \log(\lambda_{kt}) &= \alpha_k^{(\lambda)}, \\ \log(\phi_{kt}) &= \alpha_k^{(\phi)} \end{aligned}$$

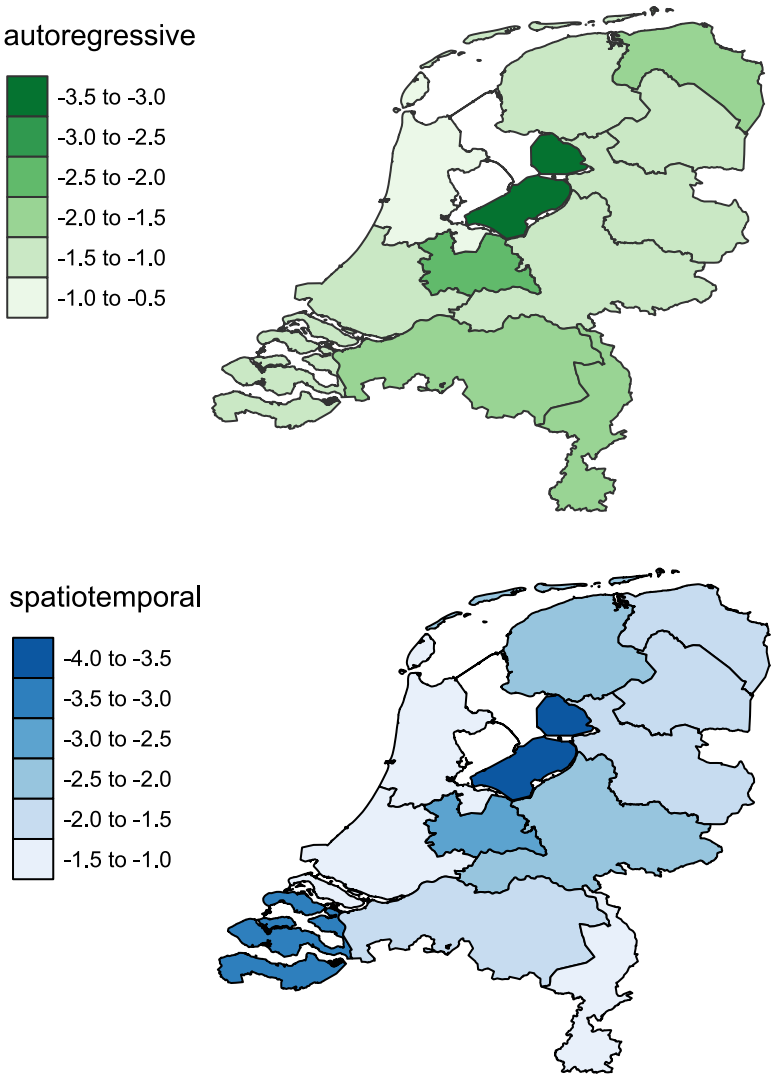

Figure 3.13: Map of the Netherlands coloured according to the region-specific autoregressive parameters (up) and region-specific spatiotemporal parameters (down) in the model that has those two terms as region-specific terms.

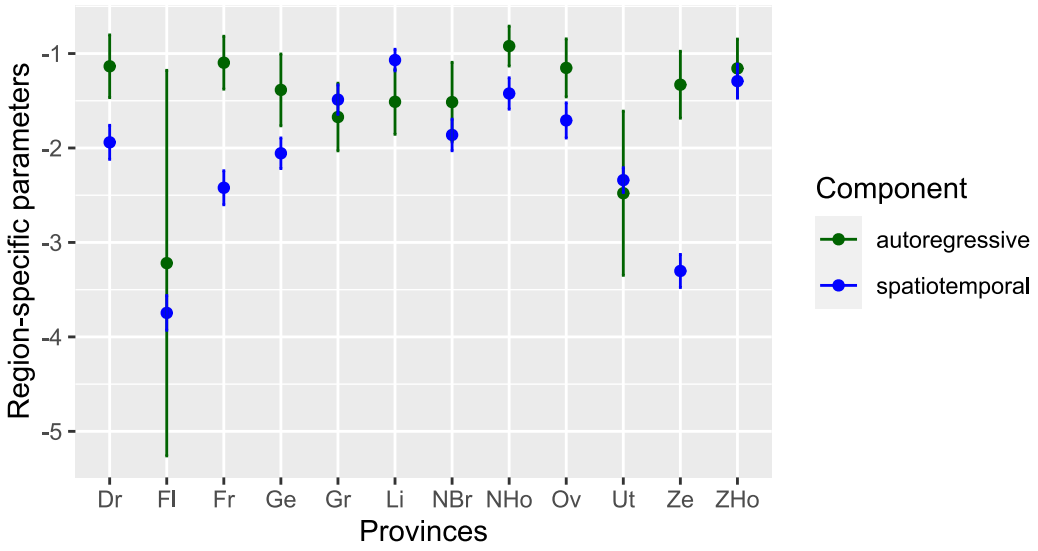

Figure 3.14: Region-specific parameters of the autoregressive ($\alpha_k^{(\lambda)}$) and the spatiotemporal ($\alpha_k^{(\phi)}$) terms of the provinces of the Netherlands.

For the selected model, maps of the Netherlands with the respective parameters for each province, both in autoregressive and spatiotemporal part are presented in Figure 3.13, while the parameters along with their standard error are shown in Figure 3.14. Flevoland appears with the lowest impact of the epidemic component on the number of deaths that happen there. Apart from Flevoland, Zeeland and Utrecht, the rest provinces have relatively larger effect derived from the spatiotemporal term. Flevoland and Utrecht are the provinces, which have the weakest dependence of the autoregressive component in the number of deaths. In addition, those two provinces have the largest standard errors. In comparison with the simple model, where all provinces have the same coefficient in each component, this model performs better. Specifically, althought the model adds 22 more parameters, model selection criteria are improved by almost 300 points.

3.3.2 Extension 1 - Order D

In the previous section, various order D models were examined. Finally, the one that accounts for five days further in the past for the number of deaths, is the one selected by the majority of the model selection criteria. In this model, we will additionally include region-specific parameters.

	Endemic	Aut.	Spat.	End.&Aut.	End.&Spat.	Aut.&Spat.	All
npar:	15	15	15	26	26	26	37
LL:	-4439.857	-4421.404	-4382.646	-4460.851	-4376.51	-4373.644	**-4365.005**
AIC:	8909.714	8872.808	8795.292	8973.702	8805.02	8799.288	**8804.01**
QAIC:	5150.942	4926.35	4585.765	4989.3	4573.188	**4521.743**	4525.815
BIC:	8999.384	8962.478	**8884.962**	9129.129	8960.447	8954.715	9025.195
CAIC:	9014.384	8977.478	**8899.962**	9155.129	8986.447	8980.715	9062.195
HQIC:	8942.015	8905.109	**8827.593**	9029.69	8861.008	8855.276	8883.685

Table 3.9: Model selection criteria scores for order 5 model when region-specific parameters appear in the Endemic (End.), Autoregressive (Aut.), Spatiotemporal (Spat.) or in combinations of those three. L.L symbolizes the log-likelihood and npar is the number of parameters.

Therefore, the selected model is:

$$E(Y_{kt}) = e_k v_t + \lambda_t \sum_{d=1}^{5} Y_{k,t-d} + \phi_{kt} \sum_{q} \sum_{d=1}^{5} w_{qt} Y_{q,t-d}$$

$$\begin{aligned} \log(v_t) &= \alpha^{(v)}, \\ \log(\lambda_t) &= \alpha^{(\lambda)}, \\ \log(\phi_{kt}) &= \alpha_k^{(\phi)} \end{aligned}$$

With a great difference, the selected model is the one that considers region-specific terms in the spatiotemporal component of the endemic-epidemic model. The location, the neighboring provinces and the mobility in them seem to play a major role in the transmission and therefore in the amount of deceases.

It is obvious from Figure 3.15 and the respective map in Figure 3.16 that the spatiotemporal terms in Flevoland, Friesland, Drenthe and Zeeland show weak effect on the number of deaths, while in Limburg, North-Holland and South-Holland the effect is relatively large. This large positive effect of the spatiotemporal term in those three provinces could be obviously explained. South-Holland is the most populous province in the Netherlands, Limburg has common borders with both Belgium and Germany, while the capital of the Netherlands, Amsterdam, belongs to North-Holland. Those factors make it more possible that the provinces attract people and therefore it is more likely that they have a high infection spread due to mobility reasons. The adoption of region-specific parameters in the spatiotemporal part, adds 11 parameters in the model

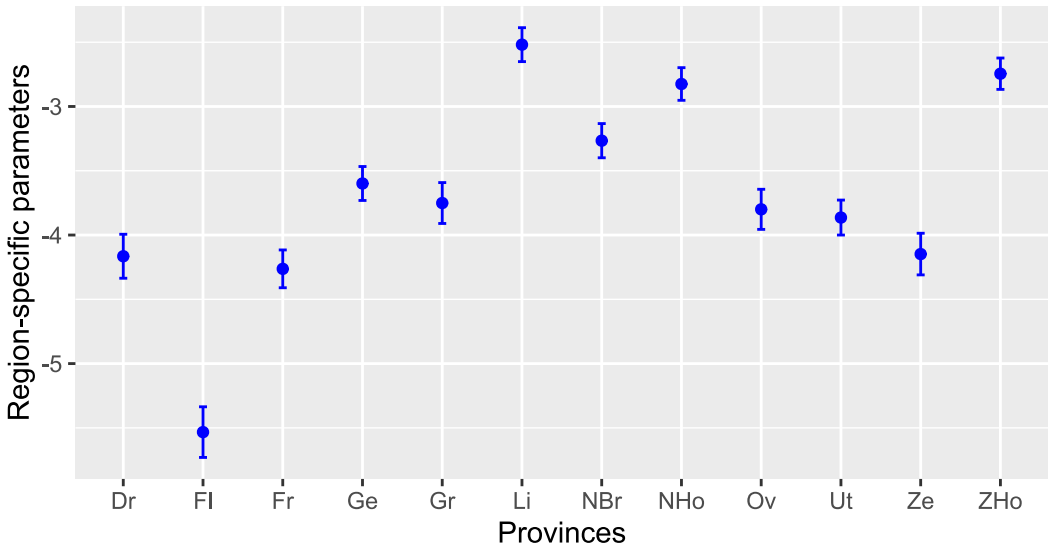

Figure 3.15: Region-specific intercepts of the spatiotemporal ($\alpha_k^{(\phi)}$) term of the provinces of the Netherlands.

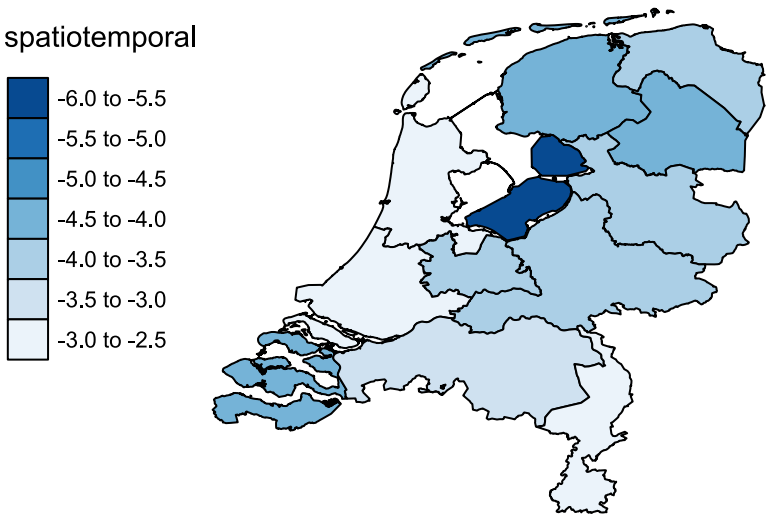

Figure 3.16: Region-specific parameters of the spatiotemporal ($\alpha_k^{(\phi)}$) term of the provinces of the Netherlands.

and improves all model selection criteria, but not considerably. More precisely, AIC improves about 113 points, BIC about 48, CAIC about 37 and HQIC about 88 points.

3.3.3 Extension 2 - Cases as covariate

In Tables A.17 - A.23, estimates of the parameters and selection criteria for each possible combination of region-specific intercepts in the second extension, are presented.

In this extension, as illustrated in Table 3.10 BIC, CAIC and HQIC suggest as well, that region-specific intercepts should be considered in the spatiotemporal term. The improvement in all those three criteria is about 200 points compared to the fixed-intercept version of the extension. AIC selects the model which has region-specific intercepts in the endemic and in the spatiotemporal term, while QAIC suggests region-specificity to all parameters. In all models, parameters in the epidemic component are higher than in the endemic.

	Endemic	Aut.	Spat.	End.&Aut.	End.&Spat.	Aut.&Spat.	All
npar:	16	16	16	27	27	27	38
LL:	-4575.325	-4580.196	-4458.638	-4545.033	-4442.097	-4451.314	**-4433.045**
AIC:	9182.65	9192.392	8949.276	9144.066	**8938.194**	8956.628	8942.09
QAIC:	6508.044	6402.231	5041.706	6069.927	4951.571	4975.298	**4845.279**
BIC:	9278.297	9288.039	**9044.923**	9305.471	9099.599	9118.033	9169.253
CAIC:	9294.297	9304.039	**9060.923**	9332.471	9126.599	9145.033	9207.253
HQIC:	9217.104	9226.846	**8983.73**	9202.207	8996.335	9014.769	9023.918

Table 3.10: Model selection criteria scores for the model with cases as covariate when region-specific parameters appear in the Endemic (End.), Autoregressive (Aut.), Spatiotemporal (Spat.) or in combinations of those three. L.L symbolizes the log-likelihood and npar is the number of parameters

Therefore, the selected model is:

$$E(Y_{kt}) = e_k v_t + \lambda_t Y_{k,t-1} + \phi_{kt} \sum_{q \neq k} w_{qt} Y_{q,t-1}$$

$$\log(v_t) = \alpha^{(v)} + \beta^{(v)} \sum_{d=1}^{5} X_{k,t-d},$$

$$\log(\lambda_t) = \alpha^{(\lambda)},$$
$$\log(\phi_{kt}) = \alpha_k^{(\phi)}$$

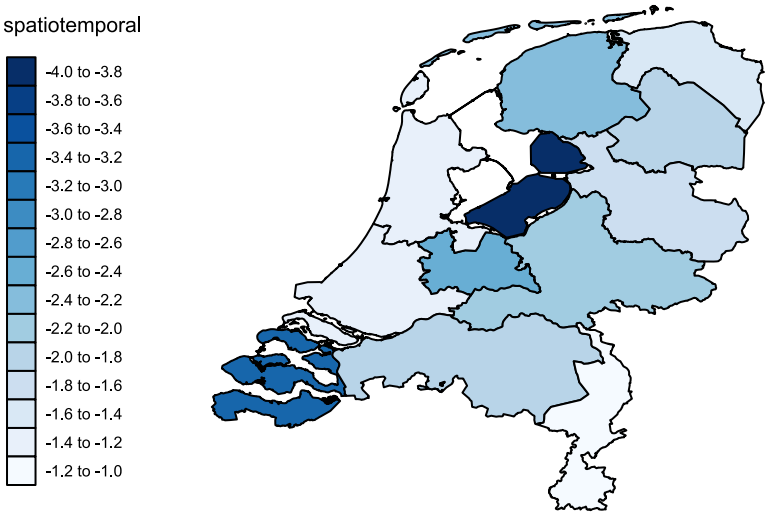

Figure 3.17: Map of the Netherlands coloured according to the region-specific spatiotemporal parameters in the model which accounts for cases five days further in the past, in the endemic part.

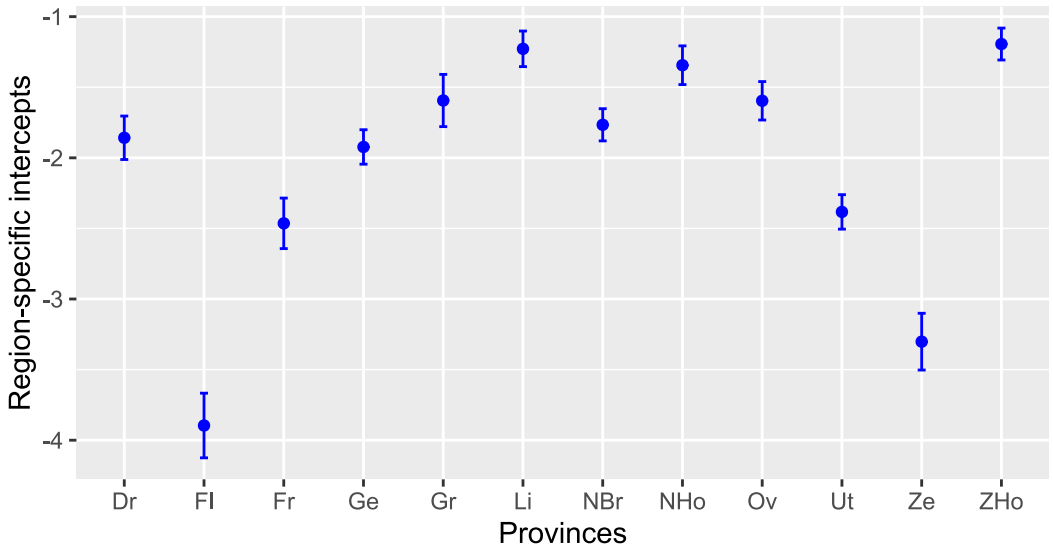

Figure 3.18: Region-specific intercepts of the spatiotemporal ($\alpha_k^{(\phi)}$) term of the provinces of the Netherlands.

As in the previous extension, we can see in Figures 3.17 and 3.18 that North-Holland, South-Holland and Limburg have the strongest effect of the spatiotemporal term in the number of deaths and this happens due to the increased mobility in those provinces. On the other hand, Flevoland and Zeeland have extremely low effect.

3.3.4 Extension 3 - Seasonality

The model with seasonality terms in the endemic and in the spatiotemporal components is examined for the adoption of region-specific intercepts. Tables A.24 - A.30 provide detailed information related to those models.

	Endemic	Aut.	Spat.	End.&Aut.	End.&Spat.	Aut.&Spat.	All
npar:	19	19	19	30	30	30	41
LL:	-4213.939	-4238.01	-4244.421	-4203.232	-4188.852	-4213.801	**-4199.591**
AIC:	8465.878	8514.02	8526.842	8466.464	**8437.704**	8487.602	8481.182
QAIC:	3128.531	3306.808	3281.73	3073.07	2988.243	3167.523	**2947.637**
BIC:	**8579.459**	8627.601	8640.423	8645.803	8617.043	8666.941	8726.279
CAIC:	**8598.459**	8646.601	8659.423	8675.803	8647.043	8696.941	8767.279
HQIC:	8506.792	8554.934	8567.756	8531.065	**8502.305**	8552.203	8569.47

Table 3.11: Model selection criteria scores for the model with seasonality when region-specific parameters appear in the Endemic (End.), Autoregressive (Aut.), Spatiotemporal (Spat.) or in combinations of those three. L.L symbolizes the log-likelihood and npar is the number of parameters

Based on Table 3.11 AIC and HQIC suggest region-specific parameters in the endemic and in the spatiotemporal term, namely, the components who already account for seasonality. QAIC selects the model which considers that all parameters have region-specific intercepts, while CAIC and BIC propose the model with region-specific intercepts in the endemic component.

Therefore, the selected model is:

$$E(Y_{kt}) = e_k v_{kt} + \lambda_t Y_{k,t-1} + \phi_{kt} \sum_{q \neq k} w_{qt} Y_{q,t-1}$$

$$\log(v_{kt}) = \alpha_k^{(v)} + \gamma^{(v)} sin(\omega t) + \delta^{(v)} cos(\omega t), \omega = \tfrac{2\pi}{365},$$
$$\log(\lambda_t) = \alpha^{(\lambda)},$$
$$\log(\phi_{kt}) = \alpha_k^{(\phi)} + \gamma^{(\phi)} sin(\omega t) + \delta^{(\phi)} cos(\omega t), \omega = \tfrac{2\pi}{365}$$

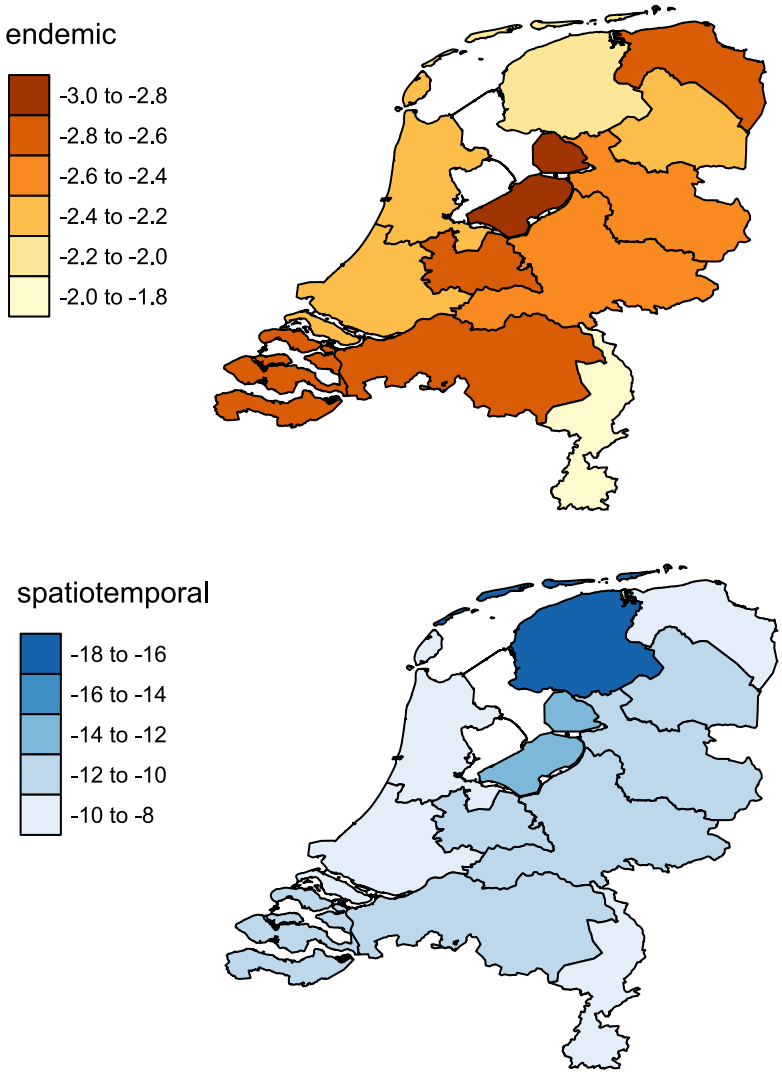

Figure 3.19: Maps of the Netherlands coloured according to the region-specific endemic (up) and spatiotemporal (down) parameters in the model that accounts for seasonality in those two terms.

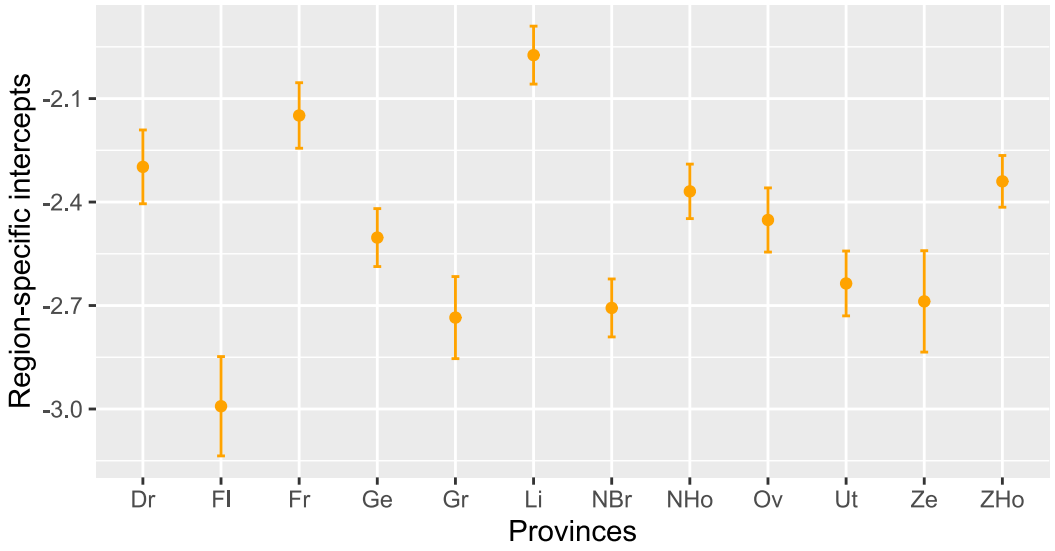

Figure 3.20: Region-specific intercepts of the endemic ($\alpha_k^{(v)}$) term of the provinces of the Netherlands.

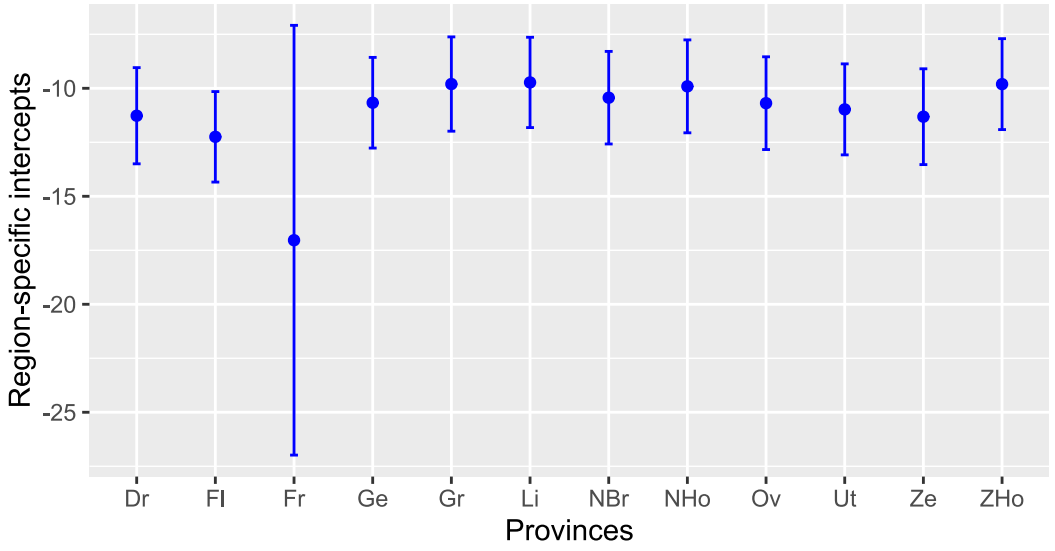

Figure 3.21: Region-specific intercepts of the spatiotemporal ($\alpha_k^{(\phi)}$) term of the provinces of the Netherlands.

The region-specific intercepts in the spatiotemporal term are very small and therefore weak dependence with the number of deaths is concluded. As it can be seen in Figure 3.19 and Figure 3.21 the weakest effect belongs to Friesland, while standard errors are relatively large for all the provinces. The endemic component presents stronger effect, without great deviations between provinces. However, the largest effect is attributed to Limburg and to Friesland and the smallest belongs to Flevoland.

3.4 Finite mixture models

Endemic-epidemic models will be now upgraded to finite mixture models by means of the EM algorithm. Using a finite mixture model, we assume that each cluster follows a different negative binomial distribution. The number of clusters will be initially selected. The algorithm will annotate each province to a cluster using mixing proportions. Finally, different parameters for each cluster will be estimated. In the following pages, the algorithm is applied to the simple version of the model. Afterwards, a model of each extension is selected to include the EM algorithm and therefore to be upgraded to a Finite Mixture Model. When we set the number of clusters G to be equal to one, then the expected log-likelihood is the same with the log-likelihood in the respective model without the algorithm, since one group is assumed in both cases. In case we ask for 12 clusters, the model coincides with the one which has region-specific intercepts to all log-linear predictors. It has to be noted, that the EM algorithm assumes different ψ parameter for each group. In general, with this form, we have to estimate the parameters in the log-linear predictors, the overdispersion parameters and the mixing proportions for each cluster. In the following subsections, various numbers of groups will be studied ranging from one to six. The number of groups that provides the best scores in information criteria, will be selected to define the partition of the Netherlands for each case.

3.4.1 Simple model

Cluster (G=1)	$\alpha_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	ψ_g	p_g
1	-3.654	-0.726	-2.541	1.318	1.0
npar:	4				
Log-likelihood:	-4626.923				
AIC:	9261.846				
BIC:	9285.758				
CAIC:	9289.758				
HQIC:	9270.459				

Table 3.12: Estimates of the parameters of the simple model which assumes one cluster, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

Cluster (G=2)	$\alpha_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	ψ_g	p_g
1	-3.638	-1.059	-3.216	0.948	0.339
2	-3.95	-1.094	-1.672	1.768	0.661
npar:	9				
Log-likelihood:	-4523.165				
AIC:	9064.33				
BIC:	9118.132				
CAIC:	9127.132				
HQIC:	9083.71				

Table 3.13: Estimates of the parameters of the simple model which assumes two clusters, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

Cluster (G=3)	$\alpha_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	ψ_g	p_g
1	-3.429	-0.972	-3.524	0.7	0.249
2	-3.766	-1.259	-1.379	1.798	0.417
3	-4.513	-1.305	-2.058	2.077	0.334
npar:	14				
Log-likelihood:	-4481.72				
AIC:	8991.441				
BIC:	9075.132				
CAIC:	9089.132				
HQIC:	9021.588				

Table 3.14: Estimates of the parameters of the simple model which assumes three clusters, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

Cluster (G=4)	$\alpha_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	ψ_g	p_g
1	-4.103	-1.542	-3.49	0.631	0.166
2	-3.877	-0.989	-1.525	2.373	0.249
3	-4.292	-1.257	-2.143	1.835	0.418
4	-3.485	-1.638	-1.264	1.085	0.167
npar:	19				
Log-likelihood:	-4466.9				
AIC:	8971.801				
BIC:	9085.382				
CAIC:	9104.382				
HQIC:	9012.715				

Table 3.15: Estimates of the parameters of the simple model which assumes four clusters, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

Cluster (G=5)	$\alpha_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	ψ_g	p_g
1	-4.102	-1.542	-3.49	0.631	0.167
2	-3.877	-0.989	-1.526	2.373	0.249
3	-4.292	-1.257	-2.143	1.837	0.416
4	-3.15	-2.249	-1.117	1.426	0.083
5	-3.75	-1.62	-1.455	0.581	0.085
npar:	24				
Log-likelihood:	-4460.653				
AIC:	8969.307				
BIC:	9112.778				
CAIC:	9136.778				
HQIC:	9020.987				

Table 3.16: Estimates of the parameters of the simple model which assumes five clusters, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

Tables 3.12 - 3.16 provide the estimations of the parameters in each group and the corresponding scores of the model selection criteria. The information criteria's scores are plotted in Figure 3.22. According to BIC and CAIC, three clusters is the appropriate assumption, while AIC suggests five and HQIC suggests four clusters. Considering three clusters, the improvement in all criteria is more than 200 points, despite the addition of 10 parameters. Compared to the model with all parameters (except overdispersion parameter) in region-specific intercepts, the model with the EM algorithm performs better. In specific, the region-specific model has a BIC score equal to 9166.935, while now BIC is equal to 9075.132. CAIC has improved more than 100 points and HQIC has a slight improvement. The selected model, in the case of region-specific intercepts was the one which adapts them in the epidemic component. That model comes across with slightly better criteria (about 100 points), but it should be mentioned that the model with the EM algorithm estimates 12 less parameters.

In Table 3.17, we can see the corresponding group of each province, for all considered scenarios. For the scenario of three clusters, which seems to be the most prominent among criteria, a map has been plotted in Figure 3.23. Flevolad, Friesland and Zeeland, being some of the less populous and less rich provinces, belong to Cluster 1 and have the highest autoregressive term. Groningen, Limburg, North-Holland, Overjissel and South-Holland belong to Cluster 2 and Gelderland, North-Brabant, Drenthe and Utrecht to Cluster 3. Cluster 2 has the highest spatiotemporal term. The inclusion of South-Holland and North-Holland, which are the most populous provinces, in Cluster 2, explains the strongest influence of the spatiotemporal part. Also, Limburg which belongs to this cluster too, has borders with two different countries, provoking higher mobility than the other provinces. The partition of the Netherlands by this way, is similar with the one that would have been proposed by taking into account region-specific epidemic parameters.

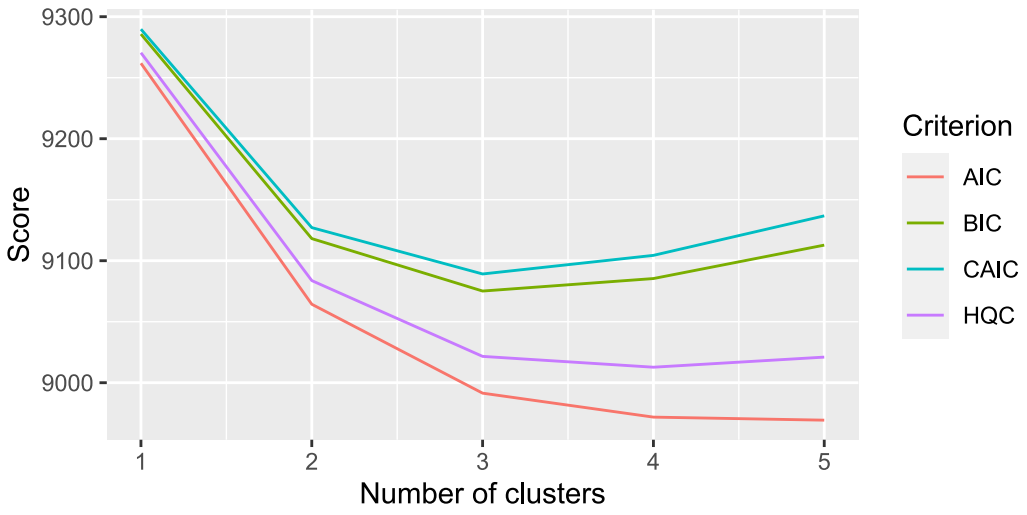

Figure 3.22: Line plot with the scores of the information criteria for each number of clusters for the simple model.

Figure 3.23: Map of the Netherlands coloured according to the cluster that each province was assigned from the EM algorithm.

Number of Clusters				
Province	Two	Three	Four	Five
Drenthe	2	3	3	3
Flevoland	1	1	1	1
Friesland	1	1	3	3
Gelderland	2	3	3	3
Groningen	2	2	4	4
Limburg	2	2	4	5
North-Brabant	2	3	3	3
North-Holland	2	2	2	2
Overjissel	2	2	2	2
Utrecht	1	3	3	3
Zeeland	1	1	1	1
South-Holland	2	2	2	2

Table 3.17: The allocation of the 12 provinces of the Netherlands for various numbers of clusters, by means of the EM algorithm in the simple model. The number of each cell denotes the cluster in which the respective province is assigned to.

3.4.2 Extension 1 - Order D

Cluster (G=1)	$\alpha_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	ψ_g	p_g
1	-4.689	-1.899	-5.057	1.73	1.0
npar:	4				
Log-likelihood:	-4450.252				
AIC:	8908.504				
BIC:	8932.416				
CAIC:	8936.416				
HQIC:	8917.118				

Table 3.18: Estimates of the parameters of the order 5 model which assumes one cluster, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

Cluster (G=2)	$\alpha_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	ψ_g	p_g
1	-4.777	-1.839	-5.252	1.94	0.833
2	-4.125	-2.471	-4.574	0.5	0.167
npar:	9				
Log-likelihood:	-4425.174				
AIC:	8868.348				
BIC:	8922.15				
CAIC:	8931.15				
HQIC:	8887.728				

Table 3.19: Estimates of the parameters of the order 5 model which assumes two clusters, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

Cluster (G=3)	$\alpha_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	ψ_g	p_g
1	-5.026	-1.982	-4.376	2.01	0.75
2	-8.352	-3.213	-2.837	0.608	0.083
3	-9.33	-3.498	-4.907	0.771	0.167
npar:	14				
Log-likelihood:	-4396.012				
AIC:	8820.024				
BIC:	8903.716				
CAIC:	8917.716				
HQIC:	8850.172				

Table 3.20: Estimates of the parameters of the order 5 model which assumes three clusters, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

Cluster (G=4)	$\alpha_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	ψ_g	p_g
1	-11.513	-3.499	-4.909	0.771	0.167
2	-5.953	-2.418	-3.763	2.436	0.324
3	-6.404	-2.536	-4.143	1.545	0.175
4	-4.737	-2.579	-3.057	1.935	0.334
npar:	19				
Log-likelihood:	-4383.509				
AIC:	8805.019				
BIC:	8918.6				
CAIC:	8937.6				
HQIC:	8845.933				

Table 3.21: Estimates of the parameters of the order 5 model which assumes four clusters, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

Cluster (G=5)	$\alpha_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	ψ_g	p_g
1	-11.513	-3.499	-4.906	0.771	0.167
2	-5.971	-2.42	-3.761	2.437	0.324
3	-5.19	-2.262	-4.283	1.351	0.093
4	-4.736	-2.579	-3.057	1.935	0.334
5	-10.942	-3.228	-3.967	1.833	0.083
npar:	24				
Log-likelihood:	-4382.29				
AIC:	8812.579				
BIC:	8956.051				
CAIC:	8980.051				
HQIC:	8864.26				

Table 3.22: Estimates of the parameters of the order 5 model which assumes five clusters, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

Number of Clusters				
Province	Two	Three	Four	Five
Drenthe	1	1	2	2
Flevoland	1	3	1	1
Friesland	1	1	3	3
Gelderland	1	1	2	2
Groningen	2	2	4	4
Limburg	1	1	4	4
North-Brabant	1	1	2	2
North-Holland	1	1	4	4
Overjissel	1	1	2	2
Utrecht	1	1	3	5
Zeeland	2	3	1	1
South-Holland	1	1	4	4

Table 3.23: The allocation of the 12 provinces of the Netherlands for various numbers of clusters, by means of the EM algorithm in the order 5 model. The number of each cell denotes the cluster in which the respective province is assigned to.

The order 5 model, with identical coefficients in lags, for each component, has been selected to be upgraded with the EM algorithm. As it can be seen in Tables 3.18 - 3.22, which is also depicted in Figure 3.24, BIC and CAIC suggest the introduction of three clusters, while AIC and HQIC perform better in the case of four clusters. For that reason, both scenarios are presented, accompanied with the corresponding map in Figure 3.25 and Figure 3.26, respectively. Whether assuming three or four clusters, the information criteria, apart from AIC, perform slightly better than the model in its full region-specific form, in previous subsection. Compared to the model with region-specific parameters only in the spatiotemporal term, the one with the EM algorithm, having just one less parameter to estimate, has very similar scores.

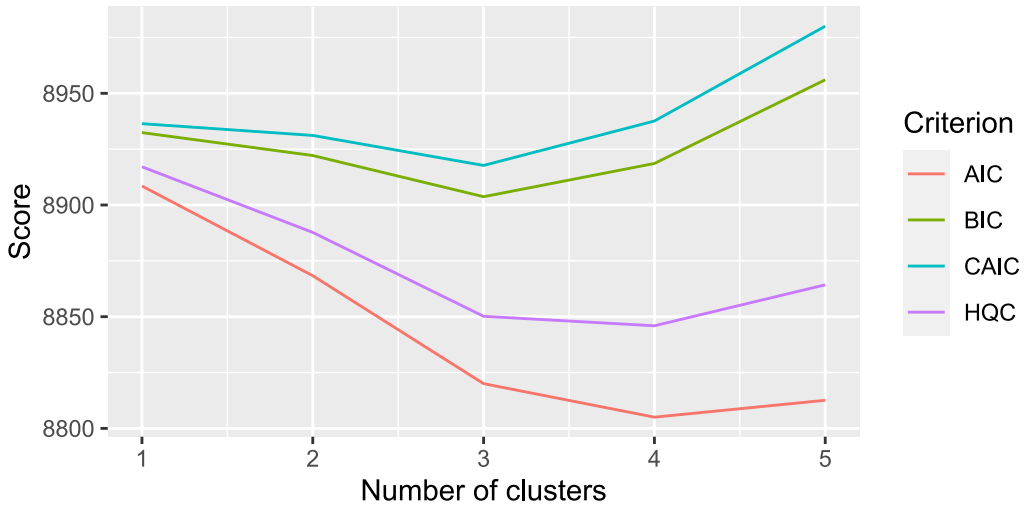

Figure 3.24: Line plot with the scores of the information criteria for each number of clusters for the order 5 model.

Table 3.23 includes information for the allocation of the provinces for each scenario but we are going to analyze the assumptions of three and four clusters, as indicated by the criteria.

Assuming three clusters, Groningen comprises Cluster 2, Flevoland and Zeeland belong to Cluster 3 and the rest provinces are part of Cluster 1. According to Table 3.20, Flevoland and Zeeland, being the less populous and most poor provinces, have the weakest dependence of all components to the number of deaths there, while the ψ parameter is very low. Groninger, in Cluster 2, has similar estimations for the ψ, the endemic and the autoregressive part, but in comparison with Cluster 3, it has larger effect of the spatiotemporal part. Cluster 1, with the inclusion of the majority of the provinces has higher effect of the endemic and the autoregressive part, while ψ parameter is appreciably higher.

According to Table 3.21, the conception of four clusters assigns Flevoland and Zeeland again the same cluster, Cluster 1, Drenthe, North-Brabant and Overjissel in Cluster 2, Friesland and Utrecht in Cluster 3 and Groningen, South-Holland, North-Holland and Limburg in Cluster 4. Cluster 4, with the two most populous and the more border-connected provinces, has the higher effect of the spatiotemporal term. Flevoland and Zeeland in Cluster 1 have the weakest impact of all components in the number of deaths, while ψ parameter is again very low. According to the region-specific selected order 5 model, depicted in Figure 3.16, Flevoland, Zeeland, Friesland and Drenthe have more similar parameters, while Limburg, North-Holland, South-Holland and Gelderland present similar effects. The partition of the Netherlands with those two different ways is similar but not identical.

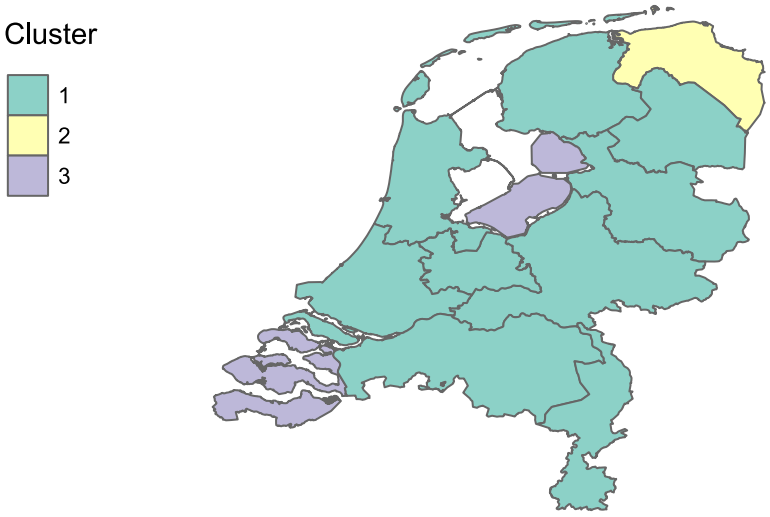

Figure 3.25: Map of the Netherlands coloured according to the cluster that each province was assigned from the EM algorithm in the order-5 extension model (according to BIC and CAIC).

Figure 3.26: Map of the Netherlands coloured according to the cluster that each province was assigned from the EM algorithm in the order-5 extension (according to AIC and HQIC).

3.4.3 Extension 2 - Cases as covariate

Cluster (G=1)	$\alpha_g^{(v)}$	$\beta_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	ψ_g	p_g
1	-3.846	-9.285	-0.778	-2.537	1.335	1.0
npar:	5					
Log-likelihood:	-4619.365					
AIC:	9248.73					
BIC:	9278.62					
CAIC:	9283.62					
HQIC:	9259.497					

Table 3.24: Estimates of the parameters of the extended with cases model which assumes one cluster, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

Cluster (G=2)	$\alpha_g^{(v)}$	$\beta_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	ψ_g	p_g
1	-3.918	-7.982	-1.063	-3.282	0.96	0.333
2	-3.95	-20.962	-1.094	-1.672	1.768	0.667
npar:	11					
Log-likelihood:	-4519.358					
AIC:	9060.715					
BIC:	9126.473					
CAIC:	9137.473					
HQIC:	9084.402					

Table 3.25: Estimates of the parameters of the extended with cases model which assumes two clusters, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

Cluster (G=3)	$\alpha_g^{(v)}$	$\beta_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	ψ_g	p_g
1	-5.093	-8.409	-1.388	-2.0	2.394	0.25
2	-3.766	-17.645	-1.259	-1.379	1.799	0.417
3	-4.937	-5.872	-1.018	-3.561	0.825	0.333
npar:	17					
Log-likelihood:	-4456.871					
AIC:	8947.741					
BIC:	9049.367					
CAIC:	9066.367					
HQIC:	8984.349					

Table 3.26: Estimates of the parameters of the extended with cases which assumes three clusters, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

Cluster (G=4)	$\alpha_g^{(v)}$	$\beta_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	ψ_g	p_g
1	-5.093	-8.409	-1.388	-2.0	2.394	0.25
2	-4.937	-5.872	-1.018	-3.561	0.825	0.333
3	-3.766	-16.662	-1.259	-1.378	1.799	0.416
4	-3.224	-3.223	-3.223	-3.223	0.04	0.0
npar:	23					
Log-likelihood:	-4456.869					
AIC:	8959.739					
BIC:	9097.232					
CAIC:	9120.232					
HQIC:	9009.266					

Table 3.27: Estimates of the parameters of the extended with cases model which assumes four clusters, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

In Tables 3.29 - 3.27, the models that take into account cases of the last five days as covariate for various numbers of clusters, utilizing the EM algorithm, are presented.

Number of Clusters			
Province	Two	Three	Four
Drenthe	2	3	2
Flevoland	1	3	2
Friesland	1	3	2
Gelderland	2	1	1
Groningen	2	2	3
Limburg	2	2	3
North-Brabant	2	1	1
North-Holland	2	2	3
Overjissel	2	2	3
Utrecht	1	1	1
Zeeland	2	3	2
South-Holland	2	2	3

Table 3.28: The allocation of the 12 provinces of the Netherlands for various numbers of clusters, by means of the EM algorithm in the model with cases as covariate. The number of each cell denotes the cluster in which the respective province is assigned to.

All model selection criteria, indicate that, when accounting for the cases in the endemic-epidemic model, the appropriate number of clusters is three. The assumption of three clusters, provides better model for the data. More precisely, AIC improves more than 300 points, while the rest criteria have an improvement more than 200 points. Information criteria, except for AIC, are also perform way better than the full region-specific model that was described in the respective section. A plot with the performance of the model selection criteria is provided in Figure 3.27. More specifically, the criteria in the selected model with region-specific intercepts only in the spatiotemporal term have very similar scores with the model examined in this section. As indicated in Table

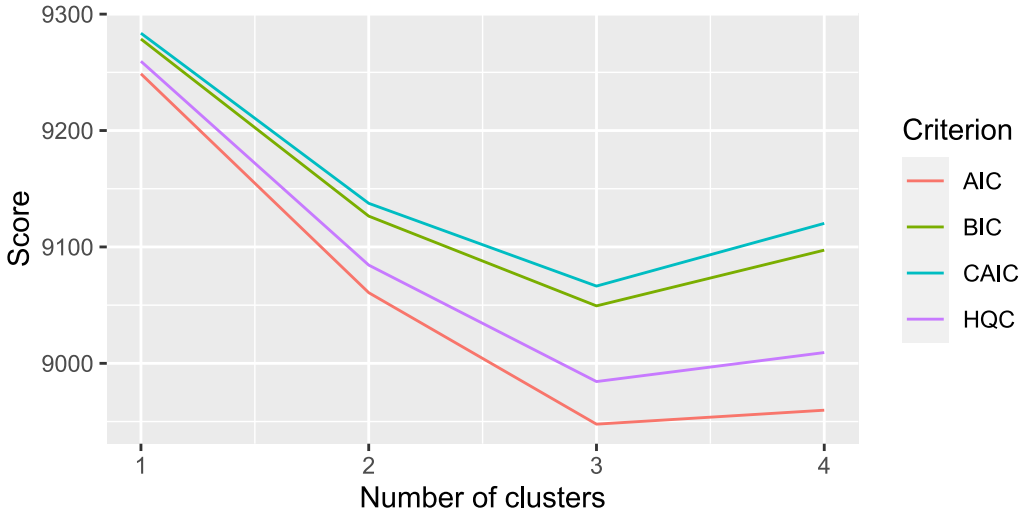

Figure 3.27: Line plot with the information criteria scores for each number of clusters for the model with COVID-19 cases as covariates.

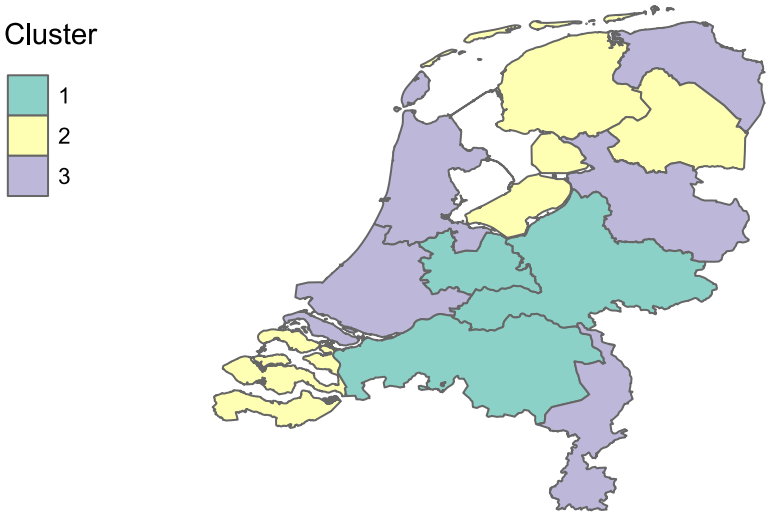

Figure 3.28: Map of the Netherlands coloured according to the cluster that each province was assigned from the EM algorithm in the extension with cases as covariate.

3.28 and depicted in Figure 3.28, Utrecht, Gelderland and North Brabant belong to Cluster 1, Drenthe, Friesland, Flevoland and Zeeland are part of Cluster 2 and Limburg, Gelderland, South- and North Holland comprise Cluster 3. This comes in great agreement with the respective selected region-specific model. Provinces in Cluster 2 share some common characteristics. They are some of the least populous and densely populated provinces, while they are the four less rich provinces in the Netherlands. In the model with the EM algorithm Cluster 1 presents the highest overdispersion, while all

components have the weakest effect on the dependent variable. Two of the provinces belonging in this cluster are two of the most large provinces and all of them have a big number of municipalities. The 3rd Cluster, with high autoregressive influence and very low overdispersion parameters consists of some of the most populous and densely populated provinces.

3.4.4 Extension 3 - Seasonality

Cluster (G=1)	$\alpha_g^{(v)}$	$\gamma_g^{(v)}$	$\delta_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	$\gamma_g^{(\phi)}$	$\delta_g^{(\phi)}$	ψ_g	p_g
1	-2.4886	0.314	1.468	-1.765	-24.861	16.833	-14.563	2.452	1
npar:	8								
Log-likelihood:	-4266.476								
AIC:	8548.952								
BIC:	8596.776								
CAIC:	8604.776								
HQIC:	8566.179								

Table 3.29: Estimates of the parameters of the model with seasonality terms, which assumes one cluster, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of parameters to be estimated.

Cluster (G=2)	$\alpha_g^{(v)}$	$\gamma_g^{(v)}$	$\delta_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	$\gamma_g^{(\phi)}$	$\delta_g^{(\phi)}$	ψ_g	p_g
1	-2.549	0.332	1.458	-1.751	-28.56	19.324	-17.232	2.745	0.917
2	-1.797	0.007	1.547	-6.016	-11.83	8.568	-6.554	2.108	0.083
npar:	17								
Log-likelihood:	-4236.09								
AIC:	8506.179								
BIC:	8607.805								
CAIC:	8624.805								
HQIC:	8542.786								

Table 3.30: Estimates of the parameters of the model with seasonality terms, which assumes two clusters, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

Tables 3.15 - 3.34 contain information for the scenarios of one to six clusters, in the model that accounts for seasonality in the endemic and in the spatiotemporal term.

Cluster (G=3)	$\alpha_g^{(v)}$	$\gamma_g^{(v)}$	$\delta_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	$\gamma_g^{(\phi)}$	$\delta_g^{(\phi)}$	ψ_g	p_g
1	-2.853	0.472	1.765	-10.794	-7.411	4.081	-3.594	12.567	0.083
2	-2.384	0.018	1.538	-1.758	-12.44	8.581	-6.178	3.575	0.337
3	-2.558	0.505	1.465	-1.749	-49.241	34.307	-31.102	1.419	0.58
npar:	26								
Log-likelihood:	-4211.096								
AIC:	8474.193								
BIC:	8629.62								
CAIC:	8655.62								
HQIC:	8530.18								

Table 3.31: Estimates of the parameters of the model with seasonality terms, which assumes three clusters, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

Cluster (G=4)	$\alpha_g^{(v)}$	$\gamma_g^{(v)}$	$\delta_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	$\gamma_g^{(\phi)}$	$\delta_g^{(\phi)}$	ψ_g	p_g
1	-2.853	0.471	1.767	-10.507	-7.335	4.026	-3.541	12.569	0.083
2	-2.326	0.205	1.379	-2.09	-13.325	9.239	-7.171	4.09	0.167
3	-2.542	0.528	1.463	-1.682	-65.477	45.9	-42.648	1.286	0.497
4	-2.487	-0.218	1.836	-2.159	-11.708	8.3	-5.038	2.994	0.252
npar:	35								
Log-likelihood:	-4193.599								
AIC:	8457.198								
BIC:	8666.427								
CAIC:	8701.427								
HQIC:	8532.566								

Table 3.32: Estimates of the parameters of the model with seasonality terms, which assumes four clusters, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

The respective model, without the incorporation of the EM algorithm, has some of the most low scores that the information criteria perform. In the case of the finite mixture model, the selected model according to AIC and HQIC is the one with five clusters. HQIC has an almost identical score in both scenarios of three and five clusters. BIC and CAIC suggest no clustering for the extension that accounts for seasonality. The latter may indicate that seasonality explains almost all of the heterogeneity between provinces. A plot with the abovementioned criteria is provided in Figure 3.29. Considering five clusters, the Netherlands are grouped as shown in Figure 3.30. Cluster 1 is composed of only one province, North-Brabant. North-Brabant seems to have a really big overdispersion in its data (ψ=12.569) and the autoregressive parameters have very weak effect. This could be followed by the fact that North-Brabant consists of many municipalities. South-Holland and North Holland belong to Cluster 2. They are the most populous, most densely populated and most rich provinces in the Netherlands. Since they are adjacent, seasonality may affect them in the same way. Cluster 3 comprises

Cluster (G=5)	$\alpha_g^{(v)}$	$\gamma_g^{(v)}$	$\delta_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	$\gamma_g^{(\phi)}$	$\delta_g^{(\phi)}$	ψ_g	p_g
1	-2.853	0.471	1.767	-10.550	-7.336	4.027	-3.542	12.569	0.083
2	-2.327	0.205	1.379	-2.09	-13.33	9.242	-7.174	4.09	0.167
3	-2.251	0.44	1.574	-1.89	-77.812	55.0523	-51.585	1.805	0.248
4	-2.426	-0.429	1.96	-2.107	-9.588	6.546	-4.192	3.664	0.168
5	-2.782	0.519	1.4	-2.231	-23.26	15.472	-12.872	1.386	0.333
npar:	44								
Log-likelihood:	-4173.958								
AIC:	8435.915								
BIC:	8698.946								
CAIC:	8742.946								
HQIC:	8530.663								

Table 3.33: Estimates of the parameters of the model with seasonality terms, which assumes five clusters, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

Cluster (G=6)	$\alpha_g^{(v)}$	$\gamma_g^{(v)}$	$\delta_g^{(v)}$	$\alpha_g^{(\lambda)}$	$\alpha_g^{(\phi)}$	$\gamma_g^{(\phi)}$	$\delta_g^{(\phi)}$	ψ_g	p_g
1	-2.853	0.471	1.767	-10.501	-7.336	4.027	-3.542	12.569	0.083
2	-2.412	0.395	1.443	-2.039	-19.65	13.754	-11.89	5.612	0.083
3	-2.783	0.517	1.401	-2.232	-22.402	14.832	-12.287	1.387	0.333
4	-2.223	-0.032	1.331	-2.405	-8.029	5.384	-3.375	3.265	0.083
5	-2.426	-0.429	1.96	-2.107	-9.586	6.545	-4.191	3.665	0.168
6	-2.251	0.44	1.574	-1.891	-75.726	53.555	-50.116	1.805	0.248
npar:	61								
Log-likelihood:	-4168.219								
AIC:	8458.437								
BIC:	8823.093								
CAIC:	8884.093								
HQIC:	8589.793								

Table 3.34: Estimates of the parameters of the model with seasonality terms, which assumes five clusters, by means of the EM algorithm. In the second part of the table the model selection criteria are provided, while npar denotes the number of the parameters.

of Limburg, Friesland and Drenthe. Provinces in this cluster have the highest influence by the autoregressive term. Gelderland and Overjiseel, which are adjacent, belong to Cluster 4, while Zeeland, Utrecht, Flevoland and Groningen are part of Cluster 5, having the lowest overdispersion parameter.

Number of Clusters					
Province	Two	Three	Four	Five	Six
Drenthe	1	3	3	3	6
Flevoland	1	3	3	5	3
Friesland	1	3	3	3	6
Gelderland	1	2	4	4	3
Groningen	1	3	3	5	5
Limburg	2	3	3	3	6
North-Brabant	1	1	1	1	1
North-Holland	1	2	2	2	4
Overjissel	1	2	4	4	5
Utrecht	1	3	4	5	3
Zeeland	1	3	3	5	3
South-Holland	1	2	2	2	2

Table 3.35: The allocation of the 12 provinces of the Netherlands for various numbers of clusters, by means of the EM algorithm in the model that accounts for seasonality. The number of each cell denotes the cluster in which the respective province is assigned to.

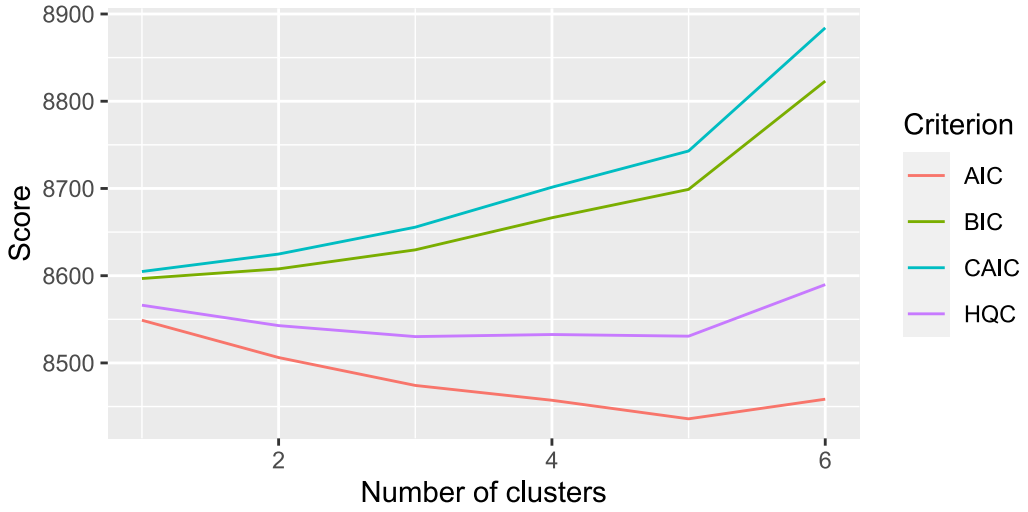

Figure 3.29: Line plot with the information criteria scores for each number of clusters for the model with seasonality terms.

All the above log-likelihoods and estimations were computed through *optim* in R. Specifically, the quasi-Newton method (also known as a variable metric algorithm) was utilized for all the models. The convergence was determined from the stopping criterion proposed by Lindsay (1995).

Figure 3.30: Map of the Netherlands coloured according to the cluster that each province was assigned from the EM algorithm in the extension with seasonality.

Chapter 4

Discussion

Although the simple form of the endemic epidemic model provides useful information, extensions for the adaption of the model in the COVID-19 framework of the pandemic should be employed. We extended the EE models to include lag in the consideration of deaths, COVID-19 cases as covariate and seasonality terms in its components. Moreover, we emphasized on region-specific intercepts in all or some of the components of the endemic-epidemic model. In our application of the basic model and its extensions to the Netherlands, the consideration of the sum of deaths of the five previous days was found to be very informative for the number of deaths in a specific day. At the same time, infections five days further in the past provided a better model, while seasonality terms were found to be pivotal. More precisely, the model with seasonality terms in the epidemic component, provided the most improved information criteria. The lag of five days in the impact of infections on deaths and the importance of seasonality terms have been also highlighted in other studies.

In this study, we have proposed a novel methodology, which can be considered an extension of the endemic-epidemic model and its existent extensions. The development of this methodology was motivated by the desire to group regions with similar characteristics. By this way the enforcement of more specific measures to some particular groups of regions and the consequent prevention of further transmission, is facilitated. The use of the finite mixture model led to significant better model in comparison with the simple one and to the one considering region-specific intercepts in all terms. In addition, the models with the EM algorithm performed similar or even better scores of the information criteria, than the model with region-specific intercepts in one or more selected components. Clustering of geographical areas is of paramount importance since heterogeneity between them is a very common characteristic and the consideration of each area as a totally different unit is not always easy. The enforcement of appropriate measures and the successful control of the contagion could be more easily facilitated when regions are considered in groups, while the information aroused from the models could guide policymakers.

In most cases, the clustering derived from the proposed endemic-epidemic finite mixture modeling was in agreement with the manual clustering of the region-specific models. Usually most populous provinces were grouped together, while the least populous

and least densely populated were part of the same cluster. This clustering, reduced the effect of the endemic component, which in our application represented the population of the provinces. Also, provinces with similar socioeconomic status were grouped together and higher status indicated weaker autoregressive term, while lower socioeconomic status inflated the autoregressive influence. Considering that richer provinces have greater vaccination coverage and that the measures are adhered more there, we suspect that this finding is quite reasonable. The spatiotemporal term was found to be more increased in groups, that contain populous areas, with borders with other countries and in general provinces that attract more people even in times of a pandemic.

From an applied perspective, the proposed model could be applied not only in other countries' provinces but also in different and even more specific levels of partitions in the Netherlands or in other countries. For example, the model could provide useful insight when applied for the clustering of the 380 municipalities of the Netherlands. In this case, the estimation of region-specific intercepts for each area, would be laborious. From a methodological viewpoint, there is also room for improvement. A natural improvement of the EM algorithm, would be to merge specific components between clusters, according to some proximity criteria, while the automated proposal of the appropriate number of clusters would be very helpful. Another promising development would be the upgrade of other existing spatiotemporal models, such as the BYM model, the Bayesian framework, the Auto-Poisson model, etc. to finite mixture models. The extensions in which the EM algorithm was incorporated can be further examined. Another consideration of weights in general, such as time-dependent weights would be beneficial to overlook some important issues. Furthermore, it would be critical to include more explanatory covariates in the endemic or in the epidemic part. Specifically, vaccination coverage and indicator functions for COVID-19 testing policies are crucial for the clustering of the provinces. In case of the extension with the cases as an explanatory variable, it is important to include a variable for the day of the week and holidays, since under-reporting during weekends and holidays is a common phenomenon. Another extension, that wasn't possible because of the limitation in the data, is age stratification as performed in Held et al. (2017). Age plays a vital role in the transmission of the virus and most importantly in the number of deaths and thus, it is crucial to be addressed. A stratified endemic-epidemic finite mixture model would provide very informative groups. Lastly, short-term predictions would be able to give significant insight for the fit of the models and provide information for the future control of the disease.

Περίληψη

Τα χωρο-χρονικά μοντέλα συνιστούν ένα βασικό εργαλείο για ένα ευρύ φάσμα επιστημονικών περιοχών. Πρόσφατα, μάλιστα, έγιναν ιδιαίτερα δημοφιλή, καθώς μπορούν να αξιοποιηθούν για τον έλεγχο της μετάδοσης της πανδημίας κορονοϊού SARS-CoV-2 (COVID-19), τόσο στον χρόνο όσο και στον χώρο. Θεωρώντας το ενδημικό-επιδημικό μοντέλο, αρχικά περιγράφουμε την γενική προσέγγιση και στη συνέχεια εξετάζουμε διάφορες επεκτάσεις του μοντέλου αυτού. Με τη βοήθεια των επεκτάσεων του ενδημικού-επιδημικού μοντέλου, πραγματοποιείται ανάλυση των ημερήσιων θανάτων της νόσου COVID-19 στις δώδεκα επαρχίες της Ολλανδίας κατά τη διάρκεια των οχτώ πρώτων μηνών του 2021. Καθώς η παρεμφερής συμπεριφορά των διαφόρων περιοχών αποτελεί ένα σύνηθες φαινομένο στα χωρο-χρονικά δεδομένα, είναι απαραίτητο να ληφθεί υπ'όψιν κατάλληλα. Στη μελέτη αυτή, προτείνουμε την ενσωμάτωση ενός αλγορίθμου που στόχο έχει την ομαδοποίηση των περιοχών με βάση τα χωρο-χρονικά τους χαρακτηριστικά. Στην εφάρμογη που πραγματοποιήσαμε, παρατηρήθηκε ότι οι επεκτάσεις του ενδημικού-επιδημικού μοντέλου που θεωρούν κάθε περιοχή ως διαφορετική οντότητα (region-specific), προσφέρουν καλύτερη προσαρμογή στα δεδομένα. Ωστόσο, όλες οι επεκτάσεις των μοντέλων, βελτιώνονται αισθητά με την ενσωμάτωση του αλγορίθμου ομαδοποίησης.

Abstract

Spatio-temporal models for count data are an important tool for a wide range of scientific fields. Recently, they have become particularly crucial since they can be employed to monitor the contagion dynamics of the COVID-19 pandemic, both in time and in space. Considering the endemic-epidemic framework, we first describe the general modelling approach and then employ various extensions. The models are exemplified through an analysis of daily COVID-19 death counts from the twelve provinces of The Netherlands during the first eight months of 2021. Since similar spatial behavior is a common feature of discrete-valued time series data, it needs to be taken into account appropriately. In this paper, we propose the incorporation of an algorithm that will cluster regions based on their spatio-temporal characteristics. In our application, we find that the region specific extensions of the endemic-epidemic model provide a better fit. However, notably, the performance of all the extensions is considerably improved by the incorporation of the clustering algorithm.

Bibliography

Aitken, A. (1926), 'Iii.—a series formula for the roots of algebraic and transcendental equations', *Proceedings of the Royal Society of Edinburgh* **45**(1), 14–22.

Alhdiri, M. A. S., Samat, N. A. & Mohamed, Z. (2017), 'Disease mapping for stomach cancer in libya based on Besag-York-Mollié (BYM) model', *Asian Pacific Journal of Cancer Prevention: APJCP* **18**(6), 1479.

Anderson, D., Burnham, K. & White, G. (1998), 'Comparison of Akaike information criterion and consistent Akaike information criterion for model selection and statistical inference from capture-recapture studies', *Journal of Applied Statistics* **25**(2), 263–282.

Ansari, M. Y., Ahmad, A., Khan, S. S., Bhushan, G. et al. (2020), 'Spatiotemporal clustering: a review', *Artificial Intelligence Review* **53**(4), 2381–2423.

Augustin, N. H., McNicol, J. & Marriott, C. A. (2006), 'Using the truncated Auto-Poisson model for spatially correlated counts of vegetation', *Journal of Agricultural, Biological, and Environmental Statistics* **11**(1), 1–23.

Behl, P., Dette, H., Frondel, M. & Tauchmann, H. (2012), 'Choice is suffering: A focused information criterion for model selection', *Economic Modelling* **29**(3), 817–822.

Berlemann, M. & Haustein, E. (2020), 'Right and yet wrong: a spatio-temporal evaluation of Germany's COVID-19 containment policy', *CESifo Working Paper* .

Blanco-Melo, D., Nilsson-Payant, B. E., Liu, W.-C., Uhl, S., Hoagland, D., Møller, R., Jordan, T. X., Oishi, K., Panis, M., Sachs, D. et al. (2020), 'Imbalanced host response to SARS-CoV-2 drives development of COVID-19', *Cell* **181**(5), 1036–1045.

Bonat, W. H. & Jørgensen, B. (2016), 'Multivariate covariance generalized linear models', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65**(5), 649–675.

Bozdogan, H. (1987), 'Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions', *Psychometrika* **52**(3), 345–370.

Bracher, J. & Held, L. (2017), 'Periodically stationary multivariate autoregressive models', *arXiv preprint arXiv:1707.04635* .

Bracher, J. & Held, L. (2020), 'Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction', *International Journal of Forecasting* **38**(3), 1221–1233.

Brewer, M. J., Butler, A. & Cooksley, S. L. (2016), 'The relative performance of AIC, AICc and BIC in the presence of unobserved heterogeneity', *Methods in Ecology and Evolution* **7**(6), 679–692.

Brugnano, L. & Iavernaro, F. (2020), 'A multi-region variant of the SIR model and its extensions', *arXiv preprint arXiv:2003.09875* .

Burnham, K. P. & Anderson, D. R. (2004), 'Multimodel inference: understanding AIC and BIC in model selection', *Sociological Methods & Research* **33**(2), 261–304.

Celani, A. & Giudici, P. (2022), 'Endemic-epidemic models to understand COVID-19 spatio-temporal evolution', *Spatial Statistics* **49**, 100528.

Celeux, G. & Govaert, G. (1992), 'A classification EM algorithm for clustering and two stochastic versions', *Computational Statistics & Data Analysis* **14**(3), 315–332.

Claeskens, G. & Hjort, N. L. (2003), 'The focused information criterion', *Journal of the American Statistical Association* **98**(464), 900–916.

Cressie, N. & Zammit-Mangion, A. (2016), 'Multivariate spatial covariance models: a conditional approach', *Biometrika* **103**(4), 915–935.

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

Diggle, A. J., Salam, M. U., Thomas, G. J., Yang, H., O'connell, M. & Sweetingham, M. (2002), 'AnthracnoseTracer: a spatiotemporal model for simulating the spread of anthracnose in a lupin field', *Phytopathology* **92**(10), 1110–1121.

Ding, J., Tarokh, V. & Yang, Y. (2017), 'Bridging AIC and BIC: a new criterion for autoregression', *IEEE Transactions on Information Theory* **64**(6), 4024–4043.

Douwes-Schultz, D., Sun, S., Schmidt, A. M. & Moodie, E. E. (2022), 'Extended Bayesian endemic-epidemic models to incorporate mobility data into COVID-19 forecasting', *Canadian Journal of Statistics* **50**(3), 713–733.

Duan, L., Zheng, Q., Zhang, H., Niu, Y., Lou, Y. & Wang, H. (2020), 'The SARS-CoV-2 spike glycoprotein biosynthesis, structure, function, and antigenicity: implications for the design of spike-based vaccine immunogens', *Frontiers in Immunology* **11**, 576622.

D'Angelo, N., Abbruzzo, A. & Adelfio, G. (2021), 'Spatio-temporal spread pattern of COVID-19 in Italy', *Mathematics* **9**(19), 2454.

Gaeta, G. (2020), 'A simple SIR model with a large set of asymptomatic infectives', *arXiv preprint arXiv:2003.08720* .

Galbadage, T., Peterson, B. M. & Gunasekera, R. S. (2020), 'Does COVID-19 spread through droplets alone?', *Frontiers in Public Health* **8**, 163.

Garg, S., Kim, L., Whitaker, M., O'Halloran, A., Cummings, C., Holstein, R., Prill, M., Chai, S. J., Kirley, P. D., Alden, N. B. et al. (2020), 'Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019—COVID-NET, 14 States, March 1-30, 2020', *Morbidity and Mortality Weekly Report* **69**(15), 458.

Glaser, S. (2017), 'A review of spatial econometric models for count data', *Hohenheim Discussion Papers in Business, Economics and Social Sciences* .

Grimée, M., Dunbar, M. B.-N., Hofmann, F., Held, L. et al. (2022), 'Modelling the effect of a border closure between Switzerland and Italy on the spatiotemporal spread of COVID-19 in Switzerland', *Spatial Statistics* **49**, 100552.

Gupta, M. R., Chen, Y. et al. (2011), 'Theory and use of the EM algorithm', *Foundations and Trends® in Signal Processing* **4**(3), 223–296.

Gusev, A. (2008), 'Temporal structure of the global sequence of volcanic eruptions: Order clustering and intermittent discharge rate', *Physics of The Earth and Planetary interiors* **166**(3-4), 203–218.

Hannan, E. J. & Quinn, B. G. (1979), 'The determination of the order of an autoregression', *Journal of the Royal Statistical Society: Series B (Methodological)* **41**(2), 190–195.

He, F., Deng, Y. & Li, W. (2020), 'Coronavirus disease 2019: What we know?', *Journal of Medical Virology* **92**(7), 719–725.

Held, L., Höhle, M. & Hofmann, M. (2005), 'A statistical framework for the analysis of multivariate infectious disease surveillance counts', *Statistical Modelling* **5**(3), 187–199.

Held, L., Meyer, S. & Bracher, J. (2017), 'Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture', *Statistics in Medicine* **36**(22), 3443–3460.

Held, L. & Paul, M. (2012), 'Modeling seasonality in space-time infectious disease surveillance data', *Biometrical Journal* **54**(6), 824–843.

Höhle, M. (2007), '*surveillance* : An R package for the monitoring of infectious diseases', *Computational Statistics* **22**(4), 571–582.

Höhle, M. (2009), 'Additive-multiplicative regression models for spatio-temporal epidemics', *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **51**(6), 961–978.

Ibañez, M. V., Martínez-Garcia, M. & Simó, A. (2021), 'A Review of Spatiotemporal Models for Count Data in R Packages. A Case Study of COVID-19 Data', *Mathematics* **9**(13), 1538.

Joe, H. & Zhu, R. (2005), 'Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution', *Biometrical Journal* **47**(2), 219–229.

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

Latouche, A., Guihenneuc-Jouyaux, C., Girard, C. & Hémon, D. (2007), 'Robustness of the BYM model in absence of spatial variation in the residuals', *International Journal of Health Geographics* **6**(1), 1–8.

Laurini, M. P. (2019), 'A spatio-temporal approach to estimate patterns of climate change', *Environmetrics* **30**(1), e2542.

Law, J., Quick, M. & Chan, P. (2014), 'Bayesian spatio-temporal modeling for analysing local patterns of crime over time at the small-area level', *Journal of Quantitative Criminology* **30**(1), 57–78.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y. et al. (2020), 'Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia', *New England Journal of Medicine* **382**, 1199–1207.

Liesenfeld, R., Richard, J.-F. & Vogler, J. (2017), 'Likelihood-Based Inference and Prediction in Spatio-Temporal Panel Count Models for Urban Crimes', *Journal of Applied Econometrics* **32**(3), 600–620.

Lindsay, B. G. (1995), Mixture models: theory, geometry, and applications, Ims.

Liu, X., Huang, J., Li, C., Zhao, Y., Wang, D., Huang, Z. & Yang, K. (2021), 'The role of seasonality in the spread of COVID-19 pandemic', *Environmental Research* **195**, 110874.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N. et al. (2020), 'Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding', *The Lancet* **395**(10224), 565–574.

Madhulatha, T. S. (2012), 'An overview on clustering methods', *arXiv preprint arXiv:1205.1117* .

Maïnassara, Y. B. & Kokonendji, C. C. (2016), 'Modified Schwarz and Hannan-Quinn information criteria for weak VARMA models', *Statistical Inference for Stochastic Processes* **19**(2), 199–217.

McNicholas, P. D., Murphy, T. B., McDaid, A. F. & Frost, D. (2010), 'Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models', *Computational Statistics & Data Analysis* **54**(3), 711–723.

Mehta, P., McAuley, D. F., Brown, M., Sanchez, E., Tattersall, R. S. & Manson, J. J. (2020), 'COVID-19: consider cytokine storm syndromes and immunosuppression', *The Lancet* **395**(10229), 1033–1034.

Meliker, J. R. & Sloan, C. D. (2011), 'Spatio-temporal epidemiology: principles and opportunities', *Spatial and Spatio-temporal Epidemiology* **2**(1), 1–9.

Meyer, S., Elias, J. & Höhle, M. (2012), 'A space–time conditional intensity model for invasive meningococcal disease occurrence', *Biometrics* **68**(2), 607–616.

Meyer, S., Held, L. & Höhle, M. (2016), 'hhh4: Endemic-epidemic modeling of areal count time series', *Jpurnal of Statistical Software* **1**, 1–55.

Meyer, S., Held, L. & Höhle, M. (2017), 'twinstim: An endemic-epidemic modeling framework for spatio-temporal point patterns', *Journal of Statistical Software* .

Müller, R., Müller, D., Schierhorn, F. & Gerold, G. (2011), 'Spatiotemporal modeling of the expansion of mechanized agriculture in the Bolivian lowland forests', *Applied Geography* **31**(2), 631–640.

Nelder, J. A. & Mead, R. (1965), 'A simplex method for function minimization', *The Computer Journal* **7**(4), 308–313.

Ng, S. K., Krishnan, T. & McLachlan, G. J. (2012), The EM algorithm, *in* 'Handbook of Computational Statistics', Springer, pp. 139–172.

Paul, M. & Held, L. (2011), 'Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts', *Statistics in Medicine* **30**(10), 1118–1136.

Paul, M. & Meyer, S. (2016), 'hhh4: An endemic-epidemic modelling framework for infectious disease counts', *Package Surveillance* .

Rajabioun, T. & Ioannou, P. A. (2015), 'On-street and off-street parking availability prediction using multivariate spatiotemporal models', *IEEE Transactions on Intelligent Transportation Systems* **16**(5), 2913–2924.

Richards, S. A. (2008), 'Dealing with overdispersed count data in applied ecology', *Journal of Applied Ecology* **45**(1), 218–227.

Rohimah, S. R., Meidianingsih, Q., Azizah, N. N. N. & Baihaqy, A. S. (2021), Analysis of factors causing the number of poor people in DKI Jakarta using spatial autoregressive Poisson model, *in* 'AIP Conference Proceedings', Vol. 2331, AIP Publishing LLC, p. 020029.

Roick, T., Karlis, D. & McNicholas, P. D. (2021), 'Clustering discrete-valued time series', *Advances in Data Analysis and Classification* **15**(1), 209–229.

Sadurní, E. & Luna-Acosta, G. (2021), 'Exactly solvable SIR models, their extensions and their application to sensitive pandemic forecasting', *Nonlinear Dynamics* **103**(3), 2955–2971.

Sajadi, M. M., Habibzadeh, P., Vintzileos, A., Shokouhi, S., Miralles-Wilhelm, F. & Amoroso, A. (2020), 'Temperature, humidity, and latitude analysis to predict potential spread and seasonality for COVID-19', *Social Science Research Network* .

Santos, B. P., Rettore, P. H., Ramos, H. S., Vieira, L. F. & Loureiro, A. A. (2018), Enriching traffic information with a spatiotemporal model based on social media, *in* '2018 IEEE Symposium on Computers and Communications (ISCC)'.

Satsuma, J., Willox, R., Ramani, A., Grammaticos, B. & Cârstea, A. S. (2004), 'Extending the SIR epidemic model', *Physica A: Statistical Mechanics and its Applications* **336**(3-4), 369–375.

Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A. & Li, F. (2020), 'Cell entry mechanisms of SARS-CoV-2', *Proceedings of the National Academy of Sciences* **117**(21), 11727–11734.

Smith, D., Moore, L. et al. (2004), 'The SIR model for spread of disease-the differential equation model', *Convergence* .

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van der Linde, A. (2014), 'The deviance information criterion: 12 years on', *Journal of the Royal Statistical Society* **76**(3), 485–493.

Telenti, A., Hodcroft, E. B. & Robertson, D. L. (2022), 'The evolution and biology of SARS-CoV-2 variants', *Cold Spring Harbor Perspectives in Medicine* **12**(5), a041390.

Van Der Linde, A. (2005), 'DIC in variable selection', *Statistica Neerlandica* **59**(1), 45–56.

Ver Hoef, J. M. & Boveng, P. L. (2007), 'Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?', *Ecology* **88**(11), 2766–2772.

Wang, X. & Brown, D. E. (2012), 'The spatio-temporal modeling for criminal incidents', *Security Informatics* **1**(1), 1–17.

Weiss, H. H. (2013), 'The SIR model and the foundations of public health', *Materials Mathematics* pp. 0001–17.

Yang, S. & Berdine, G. (2015), 'The negative binomial regression', *The Southwest Respiratory and Critical Care Chronicles* **3**(10), 50–54.

Appendix A

Tables

	Model 2.1.1.	Model 2.1.2.	Model 2.1.3.	Model 2.1.4.
	Estimate(SE)	Estimate(SE)	Estimate(SE)	Estimate(SE)
$\alpha^{(v)}$	-4.269(0.114)	-4.266(0.114)	-4.327(0.117)	-4.317(0.116)
$\alpha_1^{(\lambda)}$	-1.463(0.043)	-1.234(0.089)	-1.486(0.044)	-1.226(0.091)
$\alpha_2^{(\lambda)}$		-2.061(0.184)		-2.035 (0.176)
$\alpha_3^{(\lambda)}$		-1.296(0.092)		-1.361 (0.099)
$\alpha_1^{(\phi)}$	-4.23 (0.116)	-4.229 (0.115)	-3.983 (0.285)	-4.091 (0.319)
$\alpha_2^{(\phi)}$			-11.906 (12.707)	-12.132 (14.849)
$\alpha_3^{(\phi)}$			-3.556 (0.17)	-3.516 (0.168)
ψ	1.6 (0.006)	1.64 (0.006)	1.638 (0.006)	1.673 (0.007)
npar:	4	6	6	8
Log-likelihood:	-4515.031	-4505.076	-4497.358	**-4488.26**
AIC:	9038.062	9022.152	9006.716	**8992.52**
QAIC:	5651.789	5505.995	5496.583	**5391.162**
BIC:	9061.974	9058.02	9042.584	**9040.344**
CAIC:	9065.974	9064.02	9048.584	**9048.344**
HQIC:	9046.675	9035.072	9019.636	**9009.747**

Table A.1: Estimates and their standard error (SE) of the parameters for the order-3 models. In the second part model selection criteria and the number of the parameters (npar) are provided.

	Model 2.2.1.	Model 2.2.2.	Model 2.2.3.	Model 2.2.4.
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
$\alpha^{(v)}$	-4.537 (0.138)	-4.536 (0.137)	-4.59 (0.141)	-4.583 (0.141)
$\alpha_1^{(\lambda)}$	-1.681 (0.039)	-1.432 (0.104)	-1.693 (0.04)	-1.436 (0.108)
$\alpha_2^{(\lambda)}$		-2.357 (0.229)		-2.276 (0.21)
$\alpha_3^{(\lambda)}$		-1.594 (0.117)		-1.601 (0.12)
$\alpha_4^{(\lambda)}$		-1.572 (0.112)		-1.63 (0.119)
$\alpha_1^{(\phi)}$	-4.739 (0.132)	-4.74 (0.131)	-4.51 (0.443)	-4.549 (0.468)
$\alpha_2^{(\phi)}$			-11.873 (12.798)	-13.165 (25.779)
$\alpha_3^{(\phi)}$			-5.042 (0.831)	-4.819 (0.691)
$\alpha_4^{(\phi)}$			-3.919 (0.272)	-4.013 (0.294)
ψ	1.72 (0.007)	1.75 (0.007)	1.74 (0.007)	1.77 (0.007)
npar:	4	7	7	10
Log-likelihood:	-4460.431	-4451.707	-4449.476	**-4442.221**
AIC:	8928.862	8917.414	8912.952	**8904.442**
QAIC:	5194.548	5101.665	5128.34	**5039.459**
BIC:	**8952.774**	8959.26	8954.798	8964.222
CAIC:	**8956.774**	8966.26	8961.798	8974.222
HQIC:	8937.475	8932.488	8928.026	**8925.976**

Table A.2: Estimates and their standard error (SE) for the parameters of the order-4 models. In the second part model selection criteria and the number of the parameters (npar) are provided.

Province	$\alpha_k^{(v)}(SE)$	$\alpha^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	ψ
Drenthe	-3.662 (0.309)	-0.822 (0.069)	-2.52 (0.077)	1.398
Flevoland	-5.376 (0.697)			
Friesland	-3.459 (0.254)			
Gelderland	-4.095 (0.274)			
Groningen	-3.154 (0.197)			
Limburg	-2.591 (0.153)			
North-Brabant	-4.215 (0.241)			
North-Holland	-3.198 (0.148)			
Overijssel	-3.604 (0.21)			
Utrecht	-4.693 (0.456)			
Zeeland	-4.828 (0.573)			
South-Holland	-3.478 (0.168)			
npar:	15			
LL:	-4584.434			
AIC:	9198.868			
QAIC:	6588.561			
BIC:	9288.538			
CAIC:	9303.538			
HQIC:	9231.169			

Table A.3: Estimates and their standard error in parenthesis, for the parameters of the simple model with region-specific parameters in the endemic part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters.

Province	$\alpha^{(v)}(SE)$	$\alpha_k^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	ψ
Drenthe	-3.739 (0.082)	-0.787 (0.218)	-2.48 (0.074)	1.432
Flevoland	-11.0.53 (52.759)			
Friesland	-1.086 (0.246)			
Gelderland	0.94 (0.195)			
Groningen	-1.32 (0.269)			
Limburg	-0.309 (0.13)			
North-Brabant	-0.871 (0.175)			
North-Holland	-0.42 (0.117)			
Overijssel	-0.647 (0.16)			
Utrecht	-2.187 (0.587)			
Zeeland	-2.823 (1.384)			
South-Holland	-0.419 (0.109)			
npar:	15			
LL:	-4583.505			
AIC:	9197.01			
QAIC:	6431.543			
BIC:	9286.68			
CAIC:	9301.68			
HQIC:	9229.311			

Table A.4: Estimates and their standard error (SE) for the parameters of the simple model with region-specific parameters in the autoregressive part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters.

Province	$\alpha^{(v)}(SE)$	$\alpha^{(\lambda)}(SE)$	$\alpha_k^{(\phi)}(SE)$	ψ
Drenthe	-3.988 (0.087)	-1.381 (0.104)	-1.871 (0.148)	1.78
Flevoland			-3.911 (0.189)	
Friesland			-2.323 (0.14)	
Gelderland			-2.061 (0.114)	
Groningen			-1.514 (0.157)	
Limburg			-1.094 (0.102)	
North-Brabant			-1.907 (0.119)	
North-Holland			-1.229 (0.113)	
Overijssel			-1.625 (0.131)	
Utrecht			-2.517 (0.121)	
Zeeland			-3.293 (0.164)	
South-Holland			-1.201 (0.106)	
npar:	15			
LL:	-4460.234			
AIC:	8950.468			
QAIC:	5047.136			
BIC:	9040.138			
CAIC:	9055.138			
HQIC:	8982.769			

Table A.5: Estimates and their standard error (SE) for the parameters of the simple model with region-specific parameters in the spatiotemporal part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters.

Province	$\alpha_k^{(v)}(SE)$	$\alpha_k^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	ψ
Drenthe	-3.682 (0.319)	-0.785 (0.22)	-2.507 (0.074)	1.5
Flevoland	-5.335 (0.693)	-12.992 (146.83)		
Friesland	-3.411 (0.253)	-1.21 (0.255)		
Gelderland	-4.107 (0.285)	-0.82 (0.181)		
Groningen	-3.041 (0.193)	-1.602 (0.357)		
Limburg	-2.629 (0.189)	-0.752 (0.218)		
North-Brabant	-4.27 (0.251)	-0.724 (0.157)		
North-Holland	-3.386 (0.181)	-0.532 (0.143)		
Overijssel	-3.673 (0.22)	-0.645 (0.161)		
Utrecht	-4.401 (0.405)	-1.777 (0.418)		
Zeeland	-4.687 (0.571)	-2.298 (0.844)		
South-Holland	-3.821 (0.201)	-0.39 (0.114)		
npar:	26			
LL:	-4551.702			
AIC:	9155.404			
QAIC:	6129.039			
BIC:	9310.831			
CAIC:	9336.831			
HQIC:	9211.392			

Table A.6: Estimates and their standard error (SE) for the parameters of the simple model with region-specific parameters in the endemic and in autoregressive part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters.

Province	$\alpha_k^{(v)}(SE)$	$\alpha^{(\lambda)}(SE)$	$\alpha_k^{(\phi)}(SE)$	ψ
Drenthe	-3.977(0.365)	-1.45(0.111)	-1.854(0.154)	1.809
Flevoland	-4.026(0.492)		-3.892(0.227)	
Friesland	-3.382(0.277)		-2.448(0.176)	
Gelderland	-4.766(0.504)		-1.908(0.118)	
Groningen	-3.655(0.292)		-1.592(0.185)	
Limburg	-3.292(0.243)		-1.205(0.123)	
North-Brabant	-4.805(0.347)		-1.746(0.11)	
North-Holland	-3.527(0.187)		-1.339(0.135)	
Overijssel	-4.065(0.28)		-1.589(0.135)	
Utrecht	-4.773(0.521)		-2.377(0.121)	
Zeeland	-3.936(0.54)		-3.293(0.197)	
South-Holland	-4.042(0.216)		-1.166(0.109)	
npar:	26			
LL:	-4444.643			
AIC:	8941.286			
QAIC:	4965.923			
BIC:	9096.713			
CAIC:	9122.713			
HQIC:	8997.274			

Table A.7: Estimates and their standard error (SE) for the parameters of the simple model with region-specific parameters in the endemic and in spatiotemporal part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters.

Province	$\alpha^{(v)}(SE)$	$\alpha_k^{(\lambda)}(SE)$	$\alpha_k^{(\phi)}(SE)$	ψ
Drenthe	-4.0 (0.089)	-1.135 (0.322)	-1.94 (0.18)	1.808
Flevoland		-3.219 (2.041)	-3.746 (0.184)	
Friesland		-1.097 (0.28)	-2.421 (0.18)	
Gelderland		-1.385 (0.38)	-2.056 (0.162)	
Groningen		-1.672 (0.358)	-1.488 (0.156)	
Limburg		-1.511 (0.344)	-1.069 (0.114)	
North-Brabant		-1.514 (0.421)	-1.862 (0.167)	
North-Holland		-0.921 (0.211)	-1.423 (0.166)	
Overijssel		-1.152 (0.307)	-1.708 (0.186)	
Utrecht		-2.479 (0.869)	-2.34 (0.131)	
Zeeland		-1.33 (0.354)	-3.302 (0.175)	
South-Holland		-1.157 (0.31)	-1.292 (0.18)	
npar:	26			
LL:	-4452.499			
AIC:	8956.998			
QAIC:	4933.331			
BIC:	8936.91			
CAIC:	8940.91			
HQIC:	8921.611			

Table A.8: Estimates and their standard error (SE) for the parameters of the simple model with region-specific parameters in the autoregressive and in spatiotemporal part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters.

Province	$\alpha_k^{(v)}(SE)$	$\alpha^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	ψ
Drenthe	-3.998 (0.366)	-1.137 (0.321)	-1.94 (0.183)	1.852
Flevoland	-3.97 (0.483)	-3.256 (2.157)	-3.752 (0.211)	
Friesland	-3.324 (0.269)	-1.001 (0.265)	-2.67 (0.258)	
Gelderland	-4.76 (0.502)	-1.428 (0.369)	-1.915 (0.152)	
Groningen	-3.59 (0.291)	-1.801 (0.42)	-1.578 (0.183)	
Limburg	-3.115 (0.241)	-2.422 (0.991)	-1.114 (0.125)	
North-Brabant	-4.808 (0.347)	-1.524 (0.387)	-1.726 (0.144)	
North-Holland	-3.614 (0.201)	-1.018 (0.242)	-1.497 (0.185)	
Overijssel	-4.065 (0.281)	-1.153 (0.303)	-1.694 (0.19)	
Utrecht	-4.767 (0.51)	-2.408 (0.763)	-2.247 (0.123)	
Zeeland	-3.945 (0.541)	-1.334 (0.354)	-3.314 (0.206)	
South-Holland	-4.074 (0.221)	-1.143 (0.304)	-1.282 (0.178)	
npar:	37			
LL:	-4435.875			
AIC:	8945.75			
QAIC:	4864.362			
BIC:	9166.935			
CAIC:	9203.935			
HQIC:	9025.425			

Table A.9: Estimates and their standard error (SE) for the parameters of the simple model with region-specific parameters in all components. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters.

Province	$\alpha_k^{(v)}(SE)$	$\alpha^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	ψ
Drenthe	-13.685 (48.049)	-1.945 (0.045)	-4.904 (0.143)	1.734
Flevoland	-16.447 (195.56)			
Friesland	-4.943 (0.713)			
Gelderland	-4.852 (0.441)			
Groningen	-4.403 (0.517)			
Limburg	-3.807 (0.343)			
North-Brabant	-4.926 (0.365)			
North-Holland	-4.109 (0.277)			
Overijssel	-4.615 (0.427)			
Utrecht	-5.077 (0.577)			
Zeeland	-3.762 (0.401)			
South-Holland	-4.396 (0.296)			
npar:	15			
LL:	-4439.857			
AIC:	8909.714			
QAIC:	5150.942			
BIC:	8999.384			
CAIC:	9014.384			
HQIC:	8942.015			

Table A.10: Estimates and their standard error (SE) for the parameters of the order 5 model with region-specific parameters in the endemic part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters

Province	$\alpha^{(v)}(SE)$	$\alpha_k^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	ψ
Drenthe	-4.927 (0.186)	-2.133 (0.176)	-4.587 (0.107)	1.806
Flevoland	-12.269 (42.906)			
Friesland	-2.304 (0.19)			
Gelderland	-1.996 (0.107)			
Groningen	-2.225 (0.186)			
Limburg	-1.739 (0.083)			
North-Brabant	-1.937 (0.095)			
North-Holland	-1.801 (0.081)			
Overijssel	-2.095 (0.127)			
Utrecht	-2.147 (0.143)			
Zeeland	-2.634 (0.276)			
South-Holland	-1.812 (0.076)			
npar:	15			
LL:	-4421.404			
AIC:	8872.808			
QAIC:	4926.35			
BIC:	8962.478			
CAIC:	8977.478			
HQIC:	8905.109			

Table A.11: Estimates and their standard error (SE) for the parameters of the order 5 model with region-specific parameters in the autoregressive part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters

Province	$\alpha^{(v)}(SE)$	$\alpha^{(\lambda)}(SE)$	$\alpha_k^{(\phi)}(SE)$	ψ
Drenthe	-5.378 (0.263)	-2.726 (0.141)	-4.165 (0.171)	1.924
Flevoland			-5.533 (0.197)	
Friesland			-4.263 (0.147)	
Gelderland			-3.599 (0.132)	
Groningen			-3.751 (0.159)	
Limburg			-2.519 (0.132)	
North-Brabant			-3.266 (0.133)	
North-Holland			-2.825 (0.127)	
Overijssel			-3.8 (0.156)	
Utrecht			-3.864 (0.136)	
Zeeland			-4.148 (0.162)	
South-Holland			-2.745 (0.122)	
npar:	15			
LL:	-4382.646			
AIC:	8795.292			
QAIC:	4585.765			
BIC:	8884.962			
CAIC:	8899.962			
HQIC:	8827.593			

Table A.12: Estimates and their standard error (SE) for the parameters of the order 5 model with region-specific parameters in the spatiotemporal part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters

Province	$\alpha_k^{(v)}(SE)$	$\alpha_k^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	ψ
Drenthe	-12.43 (25.185)	-1.644 (0.124)	-5.287 (0.149)	1.807
Flevoland	-5.129 (0.731)	-2.368 (0.274)		
Friesland	-3.688 (0.279)	-1.851 (0.131)		
Gelderland	-4.072 (0.258)	-1.83 (0.101)		
Groningen	-3.187 (0.222)	-2.011 (0.176)		
Limburg	-3.547 (0.275)	-1.713 (0.095)		
North-Brabant	-4.306 (0.236)	-1.787 (0.091)		
North-Holland	-3.686 (0.22)	-1.846 (0.096)		
Overijssel	-3.7 (0.226)	-1.92 (0.118)		
Utrecht	-3.915 (0.252)	-1.885 (0.121)		
Zeeland	-2.72 (0.199)	-2.199 (0.21)		
South-Holland	-4.014 (0.227)	-1.792 (0.081)		
npar:	26			
LL:	-4460.851			
AIC:	8973.702			
QAIC:	4989.3			
BIC:	9129.129			
CAIC:	9155.129			
HQIC:	9029.69			

Table A.13: Estimates and their standard error (SE) for the parameters of the order 5 model with region-specific parameters in the endemic and in autoregressive part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters

Province	$\alpha_k^{(v)}(SE)$	$\alpha^{(\lambda)}(SE)$	$\alpha_k^{(\phi)}(SE)$	ψ
Drenthe	-12.337 (23.565)	-2.894 (0.173)	-4.044 (0.154)	1.936
Flevoland	-6.177 (2.01)		-5.385 (0.186)	
Friesland	-5.196 (0.882)		-4.183 (0.147)	
Gelderland	-5.054 (0.563)		-3.55 (0.137)	
Groningen	-5.09 (0.722)		-3.686 (0.16)	
Limburg	-4.061 (0.443)		-2.549 (0.146)	
North-Brabant	-5.664 (0.594)		-3.171 (0.126)	
North-Holland	-4.201 (0.327)		-2.913 (0.147)	
Overijssel	-4.772 (0.525)		-3.759 (0.166)	
Utrecht	-6.238 (1.289)		-3.737 (0.127)	
Zeeland	-3.785 (0.5)		-4.306 (0.226)	
South-Holland	-4.707 (0.386)		-2.731 (0.127)	
npar:	26			
LL:	-4376.51			
AIC:	8805.02			
QAIC:	4573.188			
BIC:	8960.447			
CAIC:	8986.447			
HQIC:	8861.008			

Table A.14: Estimates and their standard error (SE) for the parameters of the order 5 model with region-specific parameters in the endemic and in spatiotemporal part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters

Province	$\alpha^{(v)}(SE)$	$\alpha_k^{(\lambda)}(SE)$	$\alpha_k^{(\phi)}(SE)$	ψ
Drenthe	-5.365 (0.263)	-1.979 (0.234)	-4.845 (0.459)	1.957
Flevoland		-3.802 (1.23)	-5.263 (0.21)	
Friesland		-2.382 (0.328)	-4.455 (0.285)	
Gelderland		-2.533 (0.406)	-3.71 (0.299)	
Groningen		-3.341 (0.686)	-3.582 (0.188)	
Limburg		-2.334 (0.287)	-2.767 (0.282)	
North-Brabant		-3.368 (1.003)	-3.032 (0.243)	
North-Holland		-2.466 (0.411)	-2.985 (0.328)	
Overijssel		-2.271 (0.231)	-4.139 (0.295)	
Utrecht		-4.576 (3.349)	-3.507 (0.212)	
Zeeland		-3.406 (0.638)	-3.987 (0.176)	
South-Holland		-3.097 (0.819)	-2.589 (0.262)	
npar:	26			
LL:	-4373.644			
AIC:	8799.288			
QAIC:	4521.743			
BIC:	8954.715			
CAIC:	8980.715			
HQIC:	8855.276			

Table A.15: Estimates and their standard error (SE) for the parameters of the order 5 model with region-specific parameters in the autoregressive and in spatiotemporal part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters

Province	$\alpha_k^{(v)}(SE)$	$\alpha^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	ψ
Drenthe	-13.505 (43.465)	-1.967 (0.237)	-4.838 (0.464)	1.961
Flevoland	-10.391 (23.699)	-4.457 (2.892)	-5.163 (0.222)	
Friesland	-5.57 (1.614)	-2.437 (0.397)	-4.415 (0.352)	
Gelderland	-5.505 (0.86)	-2.776 (0.55)	-3.562 (0.288)	
Groningen	-5.492 (0.953)	-3.725 (1.132)	-3.515 (0.2)	
Limburg	-4.506 (0.648)	-2.408 (0.343)	-2.794 (0.311)	
North-Brabant	-6.296 (0.893)	-3.522 (1.285)	-2.97 (0.252)	
North-Holland	-4.313 (0.383)	-2.93 (0.722)	-2.874 (0.312)	
Overijssel	-4.897 (0.65)	-2.239 (0.235)	-4.268 (0.39)	
Utrecht	-11.793 (31.19)	-4.238 (2.266)	-3.491 (0.196)	
Zeeland	-4.15 (0.759)	-3.526 (0.785)	-4.096 (0.247)	
South-Holland	-4.698 (0.419)	-3.976 (3.054)	-2.477 (0.348)	
npar:	37			
LL:	-4365.005			
AIC:	8804.01			
QAIC:	4525.815			
BIC:	9025.195			
CAIC:	9062.195			
HQIC:	8883.685			

Table A.16: Estimates and their standard error (SE) for the parameters of the order 5 model with region-specific parameters in all components. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters

Province	$\alpha_k^{(v)}(SE)$	$\beta^{(v)}$	$\alpha^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	ψ
Drenthe	-3.661 (0.301)	-9.01 (0.229)	-0.844 (0.069)	-2.561 (0.079)	1.413
Flevoland	-5.417 (0.696)				
Friesland	-3.488 (0.246)				
Gelderland	-4.267 (0.259)				
Groningen	-3.211 (0.194)				
Limburg	-2.732 (0.15)				
North-Brabant	-4.528 (0.24)				
North-Holland	-3.646 (0.182)				
Overijssel	-3.717 (0.205)				
Utrecht	-4.784 (0.429)				
Zeeland	-4.851 (0.572)				
South-Holland	-4.055 (0.218)				
npar:	16				
LL:	-4575.325				
AIC:	9182.65				
QAIC:	6508.044				
BIC:	9278.297				
CAIC:	9294.297				
HQIC:	9217.104				

Table A.17: Estimates and their standard error (SE) for the parameters of the extension with cases as covariate with lag 5 days. The endemic part consists of province-specific parameters. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters.

Province	$\alpha^{(v)}(SE)$	$\beta^{(v)}$	$\alpha_k^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	ψ
Drenthe	-3.876 (0.1)	-9.609 (0.372)	-0.772 (0.215)	-2.481 (0.073)	1.438
Flevoland			-7.817 (10.498)		
Friesland			-1.074 (0.243)		
Gelderland			0.958 (0.198)		
Groningen			-1.296 (0.264)		
Limburg			-0.307 (0.13)		
North-Brabant			-0.927 (0.184)		
North-Holland			-0.467 (0.123)		
Overijssel			-0.642 (0.159)		
Utrecht			-2.172 (0.578)		
Zeeland			-2.733 (1.128)		
South-Holland			-0.496 (0.119)		
npar:	16				
LL:	-4580.196				
AIC:	9192.392				
QAIC:	6402.231				
BIC:	9288.039				
CAIC:	9304.039				
HQIC:	9226.846				

Table A.18: Estimates and their standard error (SE) for the parameters of the model with cases considered in lag 5. The parameters in the autoregressive part are region-specific. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters.

Province	$\alpha^{(v)}(SE)$	$\beta^{(v)}$	$\alpha^{(\lambda)}(SE)$	$\alpha_k^{(\phi)}(SE)$	ψ
Drenthe	-4.096 (0.108)	-9.872 (0.533)	-1.386 (0.104)	-1.86 (0.147)	1.78
Flevoland				-3.888 (0.186)	
Friesland				-2.31 (0.137)	
Gelderland				-2.069 (0.115)	
Groningen				-1.5 (0.156)	
Limburg				-1.092 (0.102)	
North-Brabant				-1.936 (0.123)	
North-Holland				-1.252 (0.116)	
Overijssel				-1.619 (0.131)	
Utrecht				-2.514 (0.121)	
Zeeland				-3.278 (0.162)	
South-Holland				-1.245 (0.113)	

npar:	16
LL:	-4458.638
AIC:	8949.276
QAIC:	5041.706
BIC:	9044.923
CAIC:	9060.923
HQIC:	8983.73

Table A.19: Estimates and their standard error (SE) for the parameters of the model with cases as covariate and region-specific parameters in the spatiotemporal part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters.

Province	$\alpha_k^{(v)}(SE)$	$\beta^{(v)}$	$\alpha_k^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	ψ
Drenthe	-3.691 (0.313)	-9.093 (0.23)	-0.779 (0.22)	-2.542 (0.076)	1.511
Flevoland	-5.365 (0.691)		-10.4 (40.39)		
Friesland	-3.443 (0.247)		-1.106 (0.251)		
Gelderland	-4.269 (0.274)		-0.83 (0.183)		
Groningen	-3.096 (0.192)		-1.602 (0.357)		
Limburg	-2.734 (0.183)		-0.817 (0.23)		
North-Brabant	-4.553 (0.254)		-0.749 (0.161)		
North-Holland	-3.804 (0.217)		-0.546 (0.146)		
Overijssel	-3.787 (0.217)		-0.643 (0.161)		
Utrecht	-4.508 (0.387)		-1.773 (0.4)		
Zeeland	-4.706 (0.57)		-2.218 (0.786)		
South-Holland	-4.3 (0.243)		-0.437 (0.118)		
npar:	27				
LL:	-4545.033				
AIC:	9144.066				
QAIC:	6069.927				
BIC:	9305.471				
CAIC:	9332.471				
HQIC:	9202.207				

Table A.20: Estimates and their standard error (SE) for the parameters of the model with cases as covariate in a 5 lag and with region-specific parameters in the endemic and in autoregressive part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters.

Province	$\alpha_k^{(v)}(SE)$	$\beta^{(v)}$	$\alpha^{(\lambda)}(SE)$	$\alpha_k^{(\phi)}(SE)$	ψ
Drenthe	-3.997 (0.364)	-9.483 (0.43)	-1.452 (0.111)	-1.858 (0.154)	1.814
Flevoland	-4.05 (0.49)			-3.896 (0.229)	
Friesland	-3.403 (0.274)			-2.464 (0.179)	
Gelderland	-4.858 (0.493)			-1.923 (0.122)	
Groningen	-3.701 (0.292)			-1.594 (0.185)	
Limburg	-3.347 (0.238)			-1.228 (0.126)	
North-Brabant	-4.972 (0.348)			-1.766 (0.114)	
North-Holland	-3.852 (0.237)			-1.344 (0.137)	
Overijssel	-4.141 (0.28)			-1.596 (0.136)	
Utrecht	-4.858 (0.513)			-2.383 (0.122)	
Zeeland	-3.937 (0.54)			-3.302 (0.201)	
South-Holland	-4.365 (0.265)			-1.194 (0.113)	

npar:	27
LL:	-4442.097
AIC:	8938.194
QAIC:	4951.571
BIC:	9099.599
CAIC:	9126.599
HQIC:	8996.335

Table A.21: Estimates and their standard error (SE) for the parameters of the extended model with 5 days lagged cases as covariate and with region-specific parameters in the endemic and in spatiotemporal part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters

Province	$\alpha^{(v)}(SE)$	$\beta^{(v)}$	$\alpha_k^{(\lambda)}(SE)$	$\alpha_k^{(\phi)}(SE)$	ψ
Drenthe	-4.094 (0.109)	9.996 (0.622)	-1.132 (0.321)	-1.932 (0.179)	1.809
Flevoland			-3.186 (1.967)	-3.731 (0.182)	
Friesland			-1.103 (0.281)	-2.408 (0.178)	
Gelderland			-1.378 (0.377)	-2.068 (0.163)	
Groningen			-1.657 (0.353)	-1.478 (0.155)	
Limburg			-1.5 (0.339)	-1.071 (0.114)	
North-Brabant			-1.512 (0.424)	-1.888 (0.173)	
North-Holland			-0.946 (0.217)	-1.435 (0.169)	
Overijssel			-1.153 (0.306)	-1.703 (0.185)	
Utrecht			-2.462 (0.854)	-2.341 (0.131)	
Zeeland			-1.326 (0.353)	-3.29 (0.174)	
South-Holland			-1.178 (0.316)	-1.325 (0.187)	
npar:	27				
LL:	-4451.314				
AIC:	8956.628				
QAIC:	4975.298				
BIC:	9118.033				
CAIC:	9145.033				
HQIC:	9014.769				

Table A.22: Estimates and their standard error (SE) for the parameters of the model with cases as covariate in lag of 5 days and with region-specific parameters in the autoregressive and in spatiotemporal part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters

Province	$\alpha_k^{(v)}(SE)$	$\beta^{(v)}$	$\alpha_k^{(\lambda)}(SE)$	$\alpha_k^{(\phi)}(SE)$	ψ
Drenthe	-4.019 (0.365)	-9.412 (0.406)	-1.138 (0.322)	-1.944 (0.185)	1.859
Flevoland	-3.993 (0.48)		-3.276 (2.196)	-3.756 (0.212)	
Friesland	-3.345 (0.265)		-0.991 (0.262)	-2.701 (0.266)	
Gelderland	-4.857 (0.489)		-1.417 (0.366)	-1.935 (0.157)	
Groningen	-3.639 (0.291)		-1.806 (0.422)	-1.58 (0.183)	
Limburg	-3.16 (0.232)		-2.643 (1.235)	-1.128 (0.126)	
North-Brabant	-4.988 (0.348)		-1.518 (0.387)	-1.749 (0.149)	
North-Holland	-3.939 (0.254)		-1.027 (0.242)	-1.495 (0.185)	
Overijssel	-4.147 (0.282)		-1.151 (0.304)	-1.706 (0.193)	
Utrecht	-4.865 (0.504)		-2.392 (0.751)	-2.254 (0.124)	
Zeeland	-3.945 (0.538)		-1.336 (0.355)	-3.324 (0.21)	
South-Holland	-4.423 (0.272)		-1.146 (0.297)	-1.313 (0.18)	
npar:	38				
LL:	-4433.045				
AIC:	8942.09				
QAIC:	4845.279				
BIC:	9169.253				
CAIC:	9207.253				
HQIC:	9023.918				

Table A.23: Estimates and their standard error (SE) for the parameters of the model with the five days lagged covariate of cases and with region-specific parameters in all components. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters

	$\alpha_k^{(v)}(SE)$	$\gamma^{(v)}(SE)$	$\alpha^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	$\gamma^{(\phi)}(SE)$	ψ
Drenthe	-2.342 (0.107)	0.285 (0.057)	-18.207 (5.578)	12.096 (4.294)	-9.881 (3.767)	2.727
Flevoland	-3.057 (0.144)					
Friesland	-2.24 (0.098)	$\delta^{(v)}(SE)$			$\delta^{(\phi)}(SE)$	
Gelderland	-2.504 (0.081)	1.497 (0.041)			-2.076 (0.171)	
Groningen	-2.705 (0.113)					
Limburg	-1.925 (0.08)					
North-Brabant	-2.694 (0.08)					
North-Holland	-2.347 (0.075)					
Overjissel	-2.486 (0.092)					
Utrecht	-2.647 (0.091)					
Zeeland	-2.682 (0.138)					
South-Holland	-2.311 (0.072)					
npar:	19					
LL:	-4213.939					
AIC:	8465.878					
QAIC:	3128.531					
BIC:	8579.459					
CAIC:	8598.459					
HQIC:	8506.792					

Table A.24: Estimates and their standard error (SE) for the parameters of the extension with seasonality terms in the endemic and spatiotemporal part. The endemic parameters are region-specific. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters

	$\alpha_k^{(v)}(SE)$	$\gamma^{(v)}(SE)$	$\alpha^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	$\gamma^{(\phi)}(SE)$	ψ
Drenthe	-2.504 (0.042)	0.317 (0.06)	-1.615 (0.37)	-21.725 (7.661)	14.772 (5.846)	2.593
Flevoland			-18.819 (1482)			
Friesland		$\delta^{(v)}(SE)$	-1.379 (0.255)		$\delta^{(\phi)}(SE)$	
Gelderland		1.483 (0.044)	-2.019 (0.393)		-12.14 (5.174)	
Groningen			-13.533			
Limburg			-1.044 (0.176)			
North-Brabant			-14.237 (217.44)			
North-Holland			-1.55 (0.247)			
Overjissel			-1.497 (0.277)			
Utrecht			-11.223			
Zeeland			-1.806 (0.452)			
South-Holland			-1.391 (0.205)			

npar:	19
LL:	-4238.01
AIC:	8514.02
QAIC:	3306.808
BIC:	8627.601
CAIC:	8646.601
HQIC:	8554.934

Table A.25: Estimates and their standard error (SE) for the parameters of the model with seasonality terms in the endemic and spatiotemporal components. The parameters in the autoregressive part are region-specific. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters

	$\alpha_k^{(v)}(SE)$	$\gamma^{(v)}(SE)$	$\alpha^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	$\gamma^{(\phi)}(SE)$	ψ
Drenthe	-2.732 (0.076)	0.642 (0.125)	-2.057 (0.176)	-2.8 (0.551)	-0.896 (0.444)	2.617
Flevoland				-12.4 (30.061)		
Friesland		$\delta^{(v)}(SE)$		-3.584 (0.451)	$\delta^{(\phi)}(SE)$	
Gelderland		1.187 (0.092)		-3.071 (0.335)	0.689 (5.174)	
Groningen				-5.077 (4.167)		
Limburg				-1.714 (0.27)		
North-Brabant				-4.042 (0.853)		
North-Holland				-2.376 (0.385)		
Overjissel				-2.441 (0.278)		
Utrecht				-4.043 (0.497)		
Zeeland				-17.585 (437.289)		
South-Holland				-2.168 (0.322)		

npar:	19
LL:	-4244.421
AIC:	8526.842
QAIC:	3281.73
BIC:	8640.423
CAIC:	8659.423
HQIC:	8567.756

Table A.26: Estimates and their standard error (SE) for the parameters of the model with seasonality terms in the endemic and autoregressive part and region-specific parameters in the spatiotemporal one. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters

	$\alpha_k^{(v)}(SE)$	$\gamma^{(v)}(SE)$	$\alpha^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	$\gamma^{(\phi)}(SE)$	ψ
Drenthe	-2.355 (0.139)	0.297 (0.058)	-2.006 (0.685)	-18.697 (5.427)	12.565 (4.227)	2.79
Flevoland	-2.96 (0.131)		-35.988			
Friesland	-2.297 (0.122)	$\delta^{(v)}(SE)$	-1.735 (0.426)		$\delta^{(\phi)}(SE)$	
Gelderland	-2.523 (0.109)	1.491 (0.043)	-1.976 (0.503)		-10.11 (3.605)	
Groningen	-2.664 (0.121)		-2.684 (0.894)			
Limburg	-1.878 (0.094)		-2.712 (0.89)			
North-Brabant	-2.578 (0.07)		-25.068			
North-Holland	-2.381 (0.101)		-1.884 (0.439)			
Overjissel	-2.727 (0.139)		-1.141 (0.259)			
Utrecht	-2.591 (0.114)		-2.729 (1.123)			
Zeeland	-2.788 (0.158)		-1.457 (0.363)			
South-Holland	-2.387 (0.106)		-1.669 (0.389)			
npar:	30					
LL:	-4203.232					
AIC:	8466.464					
QAIC:	3073.07					
BIC:	8645.803					
CAIC:	8675.803					
HQIC:	8531.065					

Table A.27: Estimates and their standard error (SE) for the parameters of the model with seasonality considered in the endemic and spatiotemporal part. Parameters in the endemic and in autoregressive part are region-specific. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters

	$\alpha_k^{(v)}(SE)$	$\gamma^{(v)}(SE)$	$\alpha^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	$\gamma^{(\phi)}(SE)$	ψ
Drenthe	-2.298 (0.107)	0.155 (0.062)	-2.296 (0.205)	-11.27 (2.227)	6.832 (1.731)	2.861
Flevoland	-2.992 (0.144)			-12.247 (2.095)		
Friesland	-2.149 (0.095)	$\delta^{(v)}(SE)$		-17.034 (9.949)	$\delta^{(\phi)}(SE)$	
Gelderland	-2.503 (0.084)	1.606 (0.044)		-10.668 (2.099)	-4.999 (1.329)	
Groningen	-2.735 (0.119)			-9.803 (2.183)		
Limburg	-1.974 (0.084)			-9.728 (2.091)		
North-Brabant	-2.707 (0.084)			-10.435 (2.143)		
North-Holland	-2.369 (0.079)			-9.910 (2.151)		
Overjissel	-2.452 (0.093)			-10.689 (2.147)		
Utrecht	-2.636 (0.094)			-10.976 (2.107)		
Zeeland	-2.688 (0.147)			-11.315 (2.218)		
South-Holland	-2.34 (0.075)			-9.807 (2.105)		

npar:	30
LL:	-4188.852
AIC:	8437.704
QAIC:	2988.243
BIC:	8617.043
CAIC:	8647.043
HQIC:	8502.305

Table A.28: Estimates and their standard error (SE) for the parameters of the extended model with seasonality terms in the endemic and spatiotemporal term and with region-specific parameters in the endemic and in spatiotemporal part. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters

	$\alpha_k^{(v)}(SE)$	$\gamma^{(v)}(SE)$	$\alpha^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	$\gamma^{(\phi)}(SE)$	ψ
Drenthe	-2.5 (0.044)	0.196 (0.064)	-1.68 (0.391)	-12.203 (2.484)	7.703 (1.989)	2.712
Flevoland			-14.348 (280.25)	-13.689 (2.412)		
Friesland		$\delta^{(v)}(SE)$	-1.351 (0.245)	-21.704 (106.77)	$\delta^{(\phi)}(SE)$	
Gelderland		1.581 (0.046)	-2.287 (0.514)	-11.789 (2.401)	-5.688 (1.493)	
Groningen			-3.587 (1.899)	-11.165 (2.522)		
Limburg			-1.385 (0.243)	-10.643 (2.401)		
North-Brabant			-14.931 (277.77)	-11.678 (2.436)		
North-Holland			-1.732 (0.302)	-11.052 (2.453)		
Overjissel			-1.567 (0.297)	-11.938 (2.454)		
Utrecht			-4.585 (5.454)	-12.059 (2.396)		
Zeeland			-1.756 (0.431)	-12.675 (2.489)		
South-Holland			-1.661 (0.282)	-10.952 (2.401)		

npar:	30
LL:	-4213.801
AIC:	8487.602
QAIC:	3167.523
BIC:	8666.941
CAIC:	8696.941
HQIC:	8552.203

Table A.29: Estimates and their standard error (SE) for the parameters of the model with seasonality terms in the endemic and spatiotemporal part and region-specific parameters in the epidemic component. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters

	$\alpha_k^{(v)}(SE)$	$\gamma^{(v)}(SE)$	$\alpha^{(\lambda)}(SE)$	$\alpha^{(\phi)}(SE)$	$\gamma^{(\phi)}(SE)$	ψ
Drenthe	-2.62 (0.158)	0.687 (0.087)	-3.007 (1.781)	-3.584 (0.604)	-3.059 (1.989)	2.931
Flevoland	-3.018 (0.166)		-9.433 (68.4)	-6.412 (1.042)		
Friesland	-2.461 (0.13)	$\delta^{(v)}(SE)$	-1.43 (0.327)	-13.252 (95.193)	$\delta^{(\phi)}(SE)$	
Gelderland	-2.793 (0.136)	1.155 (0.052)	-1.895 (0.45)	-3.815 (2.401)	2.335 (0.661)	
Groningen	-2.81 (0.142)		-3.035 (1.334)	-4.241 (1.014)		
Limburg	-2.088 (0.118)		-4.496 (5.799)	-3.244 (0.543)		
North-Brabant	-2.756 (0.094)		-11.676 (100.025)	-4.18 (0.649)		
North-Holland	-2.571 (0.126)		-1.953 (0.484)	-3.754 (0.897)		
Overjissel	-2.982 (0.152)		-1.99 (0.58)	-2.873 (0.471)		
Utrecht	-2.796 (0.131)		-2.308 (0.695)	-4.728 (0.674)		
Zeeland	-2.639 (0.137)		-1.916 (0.529)	-20.901 (1211.25)		
South-Holland	-2.592 (0.124)		-1.654 (0.396)	-3.9 (0.842)		
npar:	41					
LL:	-4199.591					
AIC:	8481.182					
QAIC:	2947.637					
BIC:	8726.279					
CAIC:	8767.279					
HQIC:	8569.47					

Table A.30: Estimates and their standard error (SE) for the parameters of the model with seasonality terms in the endemic and spatiotemporal component and region-specific parameters in all terms. In the second part of the table the model selection criteria are provided. L.L symbolizes the log-likelihood and npar represents the number of parameters