Investigation of the Effect of Meteorological Data on Spatial Surveillance of Disease Outbreaks

Untersuchung des Einflusses von Wetterdaten auf die räumliche Überwachung von Krankheitsausbrüchen

Bachelor thesis in Computer Science by Robert Heimbach
Date of submission: 02.12.2019

1. Review: Prof. Dr. Johannes Fürnkranz
2. Review: Dr. Eneldo Loza Mencía
Darmstadt

TECHNISCHE UNIVERSITÄT DARMSTADT

Computer Science Department
Knowledge Engineering Group

Erklärung zur Abschlussarbeit gemäß §22 Abs. 7 und §23 Abs. 7 APB der TU Darmstadt

Hiermit versichere ich, Robert Heimbach, die vorliegende Bachelorarbeit ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die Quellen entnommen wurden, sind als solche kenntlich gemacht worden. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Mir ist bekannt, dass im Fall eines Plagiats (§38 Abs. 2 APB) ein Täuschungsversuch vorliegt, der dazu führt, dass die Arbeit mit 5,0 bewertet und damit ein Prüfungsversuch verbraucht wird. Abschlussarbeiten dürfen nur einmal wiederholt werden.

Bei der abgegebenen Thesis stimmen die schriftliche und die zur Archivierung eingereichte elektronische Fassung gemäß §23 Abs. 7 APB überein.

Bei einer Thesis des Fachbereichs Architektur entspricht die eingereichte elektronische Fassung dem vorgestellten Modell und den vorgelegten Plänen.

Darmstadt, den 02.12.2019

R. Heimbach

Contents

List of Tables

List of Figures

1 Introduction

Every day people get in contact with bacteria, viruses or fungi living around us. Some of these organisms are harmless or even helpful and often nothing happens. But sometimes people fall ill and catch a disease.

The group of infectious diseases is large and diverse, differing among other things in the way of transmission. Tuberculosis and influenza are transmitted by air. Measles by person-to-person contact. Campylobacteriosis by eating contaminated food. Tetanus through wounds in the skin. Malaria by mosquito bites. For some, but not for all, of the diseases vaccinations exist that make people immune, or at least reduce the severity of the symptoms (influenza).

The burden infectious diseases pose on humans is large. Most often, especially in industrialized countries, the costs amount to just production losses or additional doctor visits. However some diseases, e.g. pneumonia, HIV or dengue fever, cause several hundred thousands of deaths each year worldwide [45].

The effort of countries to keep infectious diseases in check includes public surveillance, which is conducted by the Robert Koch Institute for Germany. Its tasks include research, advising and monitoring. To have a good monitoring system is crucial for fighting an outbreak. The earlier unusual patterns are detected, the earlier preventive measures can be carried out. One integral part of the detection mechanism, is to have a good forecast. Whether an observed number of cases is normal or exceptional, can only be judged by a comparison of the observed data with a forecast.

A disease which is monitored because of its inherent risk is influenza. Its mutation ability can cause a very serious and huge worldwide outbreak at any time. Public institutes monitor the number of reported cases and analyze samples to see how the virus changes. However, reporting of cases is usually not instantly. Several days to weeks pass by until a person feels so ill to see a doctor and the laboratory confirms a diagnosis. The FluTrends monitoring system was created by researchers to earlier detect influenza outbreaks by

using Google search data [13].[1] People who feel ill use Google to search for their symptoms or medicine before even going to a doctor.

Another monitored disease is campylobacteriosis, a gastrointestinal infection resulting in vomiting, fever and diarrhea. It is caused by Campylobacter bacteria transmitted from animals to humans via consumption of contaminated food. In Europe, Campylobacter is the pathogen most often found by laboratories analyzing fecal samples for every year since 2005 [12].

Despite many differences between influenza and campylobacteriosis, the two diseases share a common feature. When plotting the number of observed cases over time, distinctive seasonal variation can be observed. One factor that often comes to mind when thinking about seasonality is temperature. Meteorological factors surely play some role in the process of infection. For influenza, it is well understood how humidity and temperature affect the virus' ability to spread from person to person via air [23]. The role of weather is much less understood for Campylobacter, and research results are ambivalent [19, 10].

Two additional factors make meteorological data enticing to use in models forecasting case data. First, weather data is readily available. For Germany it is provided by Deutsche Wetterdienst. Second, the employed weather stations are spread all over the country providing variation across regions and thus being very well suited for use in models that have a space and time component.

The thesis is structured as follows: In chapter 2 some background information is given about the two infectious diseases influenza and campylobacteriosis. Literature that focuses on the influence of weather is discussed as well. The employed statistical model, the Endemic-Epidemic Model, is introduced in chapter 3. The various used data sources are presented in chapter 4. These include case data from the Robert Koch Institute, weather data from the Deutsche Wetterdienst and population data from the Statistische Bundesamt. A special section is reserved for explaining the employed procedures to align the data, i.e. to make the sources compatible. Results are presented in chapter 5 and discussed in chapter 6.

[1] However, the system has already closed by now.

2 Background

This chapter describes the two diseases for which case data is used in this study: Campylobacter[1] in section 2.1 and influenza in section 2.2. A special focus is made on discussing literature about the influence of weather on the incidence of the diseases and on its way of transmission.

2.1 Campylobacter

Campylobacteriosis is a bacterial infection, which is typically transmitted from animal to humans via the consumption of contaminated food or water. It causes a gastrointestinal infection, similar to a Norovirus infection, which can result in vomiting, cramps, fever, severe abdominal pain and (bloody) diarrhea [11, 43].

According to the European Food Safety Authority and European Centre for Disease Prevention and Control [12], Campylobacter was the bacterial pathogen most often found in humans having a gastroenteritis for each year since 2005. The number of cases reported by 37 European countries in 2017 was 246,158, which is an incidence rate of 64.8 per 100,000 population. The reported fatality was low (0.04%), though.

The largest source of infection in the European Union is via eating contaminated poultry, even though other animals we eat, like swine or cattle, are potential hosts for the bacteria too, as well as cats or dogs, with whom we live [8]. The prevalence of Campylobacter in raw poultry meat varies strongly across European countries. On the lower end are Finland (11%) and Denmark (12%), while on the higher end are Spain (70%) and Austria (71%); Germany is in the middle with 38% (year 2013). Therefore a large portion ($>$ 50%) of Campylobacter cases in Nordic countries, can be attributed to imported meat or traveling [12].

[1] The term Campylobacter is not only used to refer to the bacteria name, but also to the caused disease.

Given the purely monitoring nature of the study, no reasons are discussed that could explain those large differences in the prevalence numbers. It is known that Campylobacter bacteria are (usually) not transmitted by air and not vertically transmitted, unlike salmonellae, from the hen to the (contaminated) egg and forward to the hatched chicken [26]. S. J. Evans et al. [36] report that no Campylobacter was found in the environment of broiler houses after adequate cleansing and disinfection between flocks - as it is usually done. Several succeeding flocks were nevertheless later infected by Campylobacter despite hygiene barriers in place. The ability of Campylobacter to spread within an environment of a broiler house is very high, given that either no or almost all of the sampled birds testes positive.

Other, rather unexpected, ways of transmission have been described in the literature too. As early as 1990, which is 11 years before Campylobacter was included in the public surveillance in Germany, the Public Health Laboratory Service in the United Kingdom was already able to spot a fourfold increase in Campylobacter cases in the Ogwr District in Wales. Conducting interviews with the infected people and a comparable control group[2], they found out that 80% of the infected reported the drinking of milk bottles delivered to the door and attacked by birds, compared to only 8% of the control group [39]. The relevance of this transmission path for Germany should be rather low, though.

A striking feature of a time series plotting Campylobacter cases is its seasonality, characterized by increasing case numbers in each spring, a peak during the late spring or early summer and decreasing numbers in autumn. The weekly case numbers from 2001:01 to 2019:39 for Germany are shown in figure 2.1

This general pattern is very consistent over the years and across most countries in temperate climate [27, 19]. However, the timing of the peak can vary considerably across European countries: Early peaks (week 21 to 23) can be observed for example in Wales and late peaks (week 31 to 34) in Sweden [27]. The peak timing was found to be weakly associated with the temperature during the winter months, suggesting that higher number of bacteria are able to survive in the environment during milder winters and can start to replicate earlier [19].

The shape and timing of the seasonality does not only differ across countries, but also between regions. Louis et al. [22] documented sharp peaks in early June for Wales, but far less pronounced peaks in late June for the Southeast of England. In a very simple regression model, they established a correlation between the number of cases and the

[2]For each infected person, two non-infected persons were interviewed who were comparable according to age, sex and area of residence.

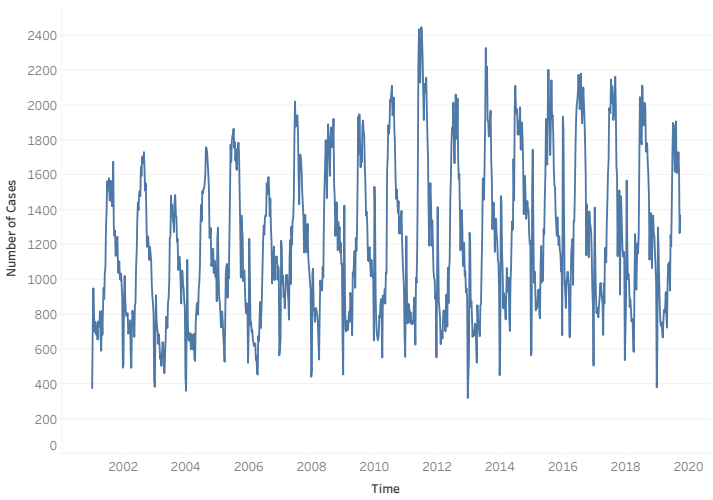


Figure 2.1: Weekly Campylobacter cases for Germany, from 2001:01 till 2019:39.

meteorological variables temperature, precipitation and hours of sunshine, as well as variables measuring the degree of urbanization (population density, total number of cattle, pigs, sheep and poultry).

While no causal relationship can be established in the literature, two groups of reasons for the observed seasonality are discussed: Seasonal variation in human behavior that changes the exposure to Campylobacter bacteria (e.g. barbecuing in the summer), and seasonal variation of the prevalence of the bacteria in its reservoirs. Several authors focused on investigating whether meteorological factors can explain the variation in the Campylobacter cases, given the availability of weather data and the prominent role of seasonality in variables like temperature.

Focusing on the main source of infection, [42] showed that the number of bacteria found on raw meat of a poultry processing plant in Lancashire, UK, was correlated with the minimum and maximum temperatures as well as the hours of sunshine. A much weaker correlation of the prevalence of Campylobacter with meteorological factors was reported by [30] for raw poultry meat in Denmark, tested during a national monitoring program from 1998 to 2001. The correlation was higher, though, for the reported number of cases in humans, with the maximum temperature 4 weeks prior being the best single predictor.

Ambivalent results are reported with regard to the role of weather affecting the number of infections in humans. [19] used time series data from 14 different countries and

found almost no effect of short-term variations in temperature on the Campylobacter incidence. In contrast, [10] conducted an analysis on the regional level, using weather data (temperature, rainfall) for Wales and England from 2005 to 2009, as well as postcode information about the laboratory that conducted each Campylobacter proof. They claim that the relationship between temperature and Campylobacter infections is non-linear and can account for up to 33.3% of the expected cases.

2.2 Influenza

Influenza, also called flu, is a highly contagious respiratory infection caused by a group of influenza viruses. The symptoms, usually including high fever, coughing, severe malaise, muscle pain and a running nose, occur suddenly and can be mild to severe. For people at high risk, however, like young children and old people, influenza can be deadly. The World Health Organization (WHO) estimates that worldwide 290,000 to 650,000 people, mostly in developing countries, die each year from an influenza infection [44].

The virus is monitored by the WHO and countries around the world because of its inherent risk. Transmission occurs very easily because every infected person disperses droplets containing the virus into the air and onto surfaces when coughing or sneezing. Additionally high mutation rates reduces the body's ability to detect and fight the virus, which can lead to global and very severe outbreaks (called pandemics) [46].

Similar to Campylobacter infections, influenza features a seasonality in temperate climates. However, it is much more pronounced, the cases occur mainly during winter months and the total number of infected people can vary sharply from year to year. The weekly case numbers for Germany from 2001:01 till 2019:39 are shown in figure 2.2

Biological, social and environmental factors contributing to the seasonality are discussed in the literature: Lower melatonin levels in the darker winter months reduces the immune systems of humans and animals. Additionally, people tend to gather indoors (e.g. Christmas shopping) enhancing the ability of the virus to spread from person to person. Heating is turned up during colder months as well, which lead to lower air humidity, which increases the survival of viral particles in the air [21].

The role of meteorological factors in the transmission of influenza is much deeper understood than for Campylobacter. In an experiment using guinea pigs under varying temperature and relative humidity conditions, Lowen et al. were able to provide direct evidence that cold and dry conditions favor the transmission via aerosols together with

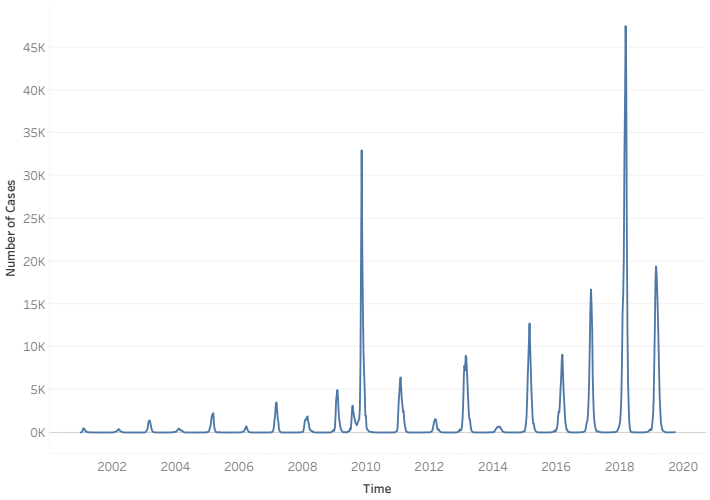


Figure 2.2: Weekly influenza cases for Germany, from 2001:01 till 2019:39.

results suggesting that cold temperatures (5°C) did not lower the responses of the immune system. Transmission was almost completely blocked when high temperatures (30°C) were combined with very high relative humidity (80%) [23]. Measuring the temperature, humidity and CO2 levels in classrooms at two Minnesota grade schools, Tyler H. Koep et al. estimate that running an air humidifier for four hours could lower the amount of virus particles in the air, measured 1 hour afterwards, by 30% [41].

Despite the findings in the guinea pig experiment [23], influenza infections occur throughout the whole year with almost no seasonality in tropical regions. In those wet and warm conditions, the virus probably behaves somewhat differently. While the aerosol transmission is indeed reduced, much more virus particles can be found for a longer duration on surfaces, increasing the risk of contact transmission [33].

The link between (absolute) humidity and influenza cases was confirmed in studies using population data, for example [37] use 30 years of health and climate data for the United States. Inserting meteorological variables like rainfall, temperature and atmospheric pressure into a time series model proved valuable for warmer regions (Arizona and Hong Kong) as well [38].

3 The Endemic-Epidemic Model

This chapter introduces the chosen statistical modeling approach, the Endemic-Epidemic Model. It was created by Held et al. [15, 18] to handle the type of data that typically arises from public health surveillance. Given needs for privacy protection, the individual case data is aggregated by health departments such that only the number of reported cases in a given region during a given time period remains. Statistically, those are event counts with a time and space dimension leading to the class of multivariate Poisson time series models.

The used model assumes that the number of reported cases Y_{it} in region i at week t, conditioned on the known past number of cases $\boldsymbol{Y_{t-1}}$, follows a **negative binomial distribution** with a region- and time-specific mean μ_{it} and a time-invariant overdispersion parameter $\psi_i > 0$ controlling the variance.[1]

$$Y_{it}|\boldsymbol{Y_{t-1}} \sim \mathbf{NB}(\mu_{it}, \psi_i) \tag{3.1}$$

Given the large number of counties used in this study, it is further assumed that the overdispersion parameter is not only constant over time but also constant across regions, leading to $\psi_i \equiv \psi$. Replacing the negative binomial distribution with a more restricting Poisson distribution ($\psi_i = 0$) is never done, because the overdispersion parameter is consistently and significantly above zero in all estimations.[2]

[1] The variance is given by $\text{Var}(y_{it}) = \mu_{it}(1 + \psi_i\mu_{it})$ and exceeds the conditional mean μ_{it} as long as $\psi_i > 0$.

[2] Overdispersion can for example arise from under-reporting and reporting delays, or any other unobserved covariates affecting how many and at what time cases are reported. [31].

The characteristic feature of the Endemic-Epidemic Model is the additive decomposition of the conditional mean μ_{it} into an endemic and an epidemic component.

$$\mu_{it} = \underbrace{\lambda_{it}^{\text{AR}} Y_{i,t-1} + \lambda_{it}^{\text{NE}} \sum_{j \neq i} w_{ji} Y_{j,t-1}}_{\text{Epidemic}} + \underbrace{\phi_{it} \lambda_{it}^{\text{EN}}}_{\text{Endemic}} \tag{3.2}$$

The distinction is based on the literature on models for infectious disease counts, in which the endemic component is described as persistent and stable over-time, in contrast to the epidemic component, which comprises occasional outbreaks that eventually, and usually, burn out as time progresses [15].

Mathematically, the central distinction is the epidemic component's dependency on observed cases from the previous week - either from the same region $Y_{i,t-1}$ or from different regions $Y_{j,t-1}$ with $j \neq i$.

The λ_{it}^{τ} with $\tau \in \{\text{AR}, \text{NE}, \text{EN}\}$ describe three distinctive rates of infection of healthy people. In the **autoregressive rate (AR)** the healthy are infected by ill people living in the same region. In contrast, its the ill people from different regions $j \neq i$ who are infecting the healthy in the **neighborhood rate (NE)**. How fast and how far a disease can spread, is determined by the weights w_{ji}. These can either be set or estimated within the model, for example assuming a power-law distance decay [24].[3]

The third rate, the **endemic (EN)**, comprises all other causes of infection, which are independent of disease counts from the previous week. The endemic λ_{it}^{EN} is typically scaled by some population measure ϕ_{it}, as it is a useful approximation to the theoretical quantity of the number of healthy people who are at risk of becoming ill [25].

The above mentioned characteristic equation 3.2 can be enriched by modeling each of the three rates λ_{it}^{τ} with $\tau \in \{\text{AR}, \text{NE}, \text{EN}\}$ in a log-linear way with an intercept α_i^{τ} and up to K possibly time- and region-varying covariates.

$$log(\lambda_{it}^{\tau}) = \alpha_i^{\tau} + \sum_{k=1}^{K} \beta_k^{\tau} \boldsymbol{x}_{it}^{\tau} \tag{3.3}$$

[3] The power-law distance decay assumes the functional form $w_{ji} = o_{ji}^{-d}$ with o_{ji} being the order in a neighborhood graph of the regions, and d the so-called decay parameter, which is to be estimated.

While fixed-effects estimations (FE) or random-effects estimations (RE) of the region-varying intercept(s) α_i^τ is possible [31], it is assumed in this study that the intercept in each component is constant across regions, i.e. $\alpha_i^\tau \equiv \alpha^\tau$ for each $\tau \in \{\text{AR}, \text{NE}, \text{EN}\}$. The reason is that the cross-sectional dimension (401 counties) is very large compared to the time dimension (104 time points) and thus the estimation, especially the RE case, too difficult and time consuming.[4]

The regressors $\boldsymbol{x}_{it}$, also called covariates or features, can be used like in any linear regression frameworks. For example, to account for seasonality using the sum of sine and cosine terms [17].[5] Or to incorporate other recurring effects like notification gaps as done by Held et al. modeling cases of Norovirus gastroenteritis in Berlin [16], or school closures as done by Bauer et al. modeling cases of hand, foot and mouth disease in the central north region of China. All the meteorological variables used in this study, like temperature, rain, hours of sunshine and humidity, will enter the model via this path.

Specification and estimation of the model was done in R, using the surveillance package [25], in particular the *hhh4* method.[6] The various used formulations of equation 3.3 are presented in the (sub-)sections of chapter 5 (Results).

While the development of and the extensions to the Endemic-Epidemic Model were mainly driven by statistical and practical concerns, the model itself can be viewed as the statistical equivalent of a theoretical susceptible-infectious-removed (SIR) model allowing for immigration [1]. In SIR models the population is divided into three distinct groups: healthy people facing the risk of becoming ill (the susceptible), ill people who possibly infect healthy people (the infectious), and people who were ill but are now healthy again or dead (the removed). The transit from one group to another is modeled using deterministic differential equations.

[4] A random effects estimation assumes $\alpha_i \sim N(0, \sigma_\alpha^2)$. That is each region's intercept is a random draw from a normal distribution with mean zero and variance σ_α^2, which needs to be estimated.

[5] One sine-cosine pair with a periodicity of 52 weeks has the formula: $\beta_1 \sin(\frac{t}{52} 2\pi) + \beta_2 \cos(\frac{t}{52} 2\pi)$.

[6] The package obtains parameter estimates for the model by maximizing its log-likelihood using the quasi-Newton algorithm, together with the Fisher information, and the model's analytical score function.

4 Data

This chapter describes all data sources used in the estimation and documents the various performed preprocessing procedures. Recalling the formula of the model equation 3.2 gives a good idea what data sources are needed.

$$\mu_{it} = \lambda_{it}^{AR} Y_{i,t-1} + \lambda_{it}^{NE} \sum_{j \neq i} w_{ji} Y_{j,t-1} + \phi_{it} \lambda_{it}^{EN}$$

The mean of the negative binomial counts distribution is explained by the number of cases $Y_{i,t-1}$ in region i at week $t-1$ (see section 4.1), a spillover effect from nearby counties (see section 4.2) and an endemic component, which is scaled by the county's population ϕ_{it} (see section 4.3).

The meteorological data which is incorporated into the above model equation will be presented in section 4.4. The main challenge encountered during preprocessing, the aligning of the data from the various sources, will be extensively described in section 4.5.

4.1 Count Data

Campylobacter and influenza are both infectious diseases which require a notification to state health departments in case of a positive laboratory diagnosis. The requirement is stated in the Act on the Prevention and Control of Infectious Diseases in Man, *Infektionsschutzgesetz – IfSG* §7(1) [6].

The notifications are passed on to the **Robert Koch-Institut** (RKI), which provides access to an aggregated version of the individual notification data via the SurvStat@RKI 2.0 web tool [35].

The number of times a disease was reported in county i at week t can also be called the cases or the count for county i in week t. At the time of the thesis, these counts were available starting from week 1 in 2001 to week 40 in 2019. Despite the availability of the counts, all estimations with data about influenza were restricted to the period 2008 to 2019, because broader notification requirements [35], starting March 2007, lead to a structural break in the time series by considerably increasing the number of counts.

Three data preparation steps were needed. First, cases assigned to the county called *unknown* were deleted. Second, the counts for the 12 districts of Berlin were merged together, because population and polygon map data for each district was not available (see sections 4.2 and 4.3). Third, all data registered for some week 53 were removed because the employed statistical model works on the assumption that every year has exactly 52 weeks. The years 2004, 2009 and 2015 are affected by this.

4.2 Neighborhood Data

For the neighborhood component of the model, $\lambda_{it}^{NE} \sum_{j \neq i} w_{ji} Y_{j,t-1}$, knowledge about the location of each county in Germany and its neighbors is needed, in particular, the weights w_{ij} need to be specified.

The needed data is freely available for academic and non-commercial use from **GADM**, the Database of Global Administrative Areas.[1] It is basically a map of Germany, made up of a collection of polygons, one for each county. The R library **sp** [34, 3] was used to adjust the map data, as it provides functionality for a *SpatialPolygonsDataFrame*, which is used for storing the data.

Three changes were applied to the downloaded data. First, the polygon *Bodensee* was deleted. It was included as a water body, despite not being a separate county. Second, encoding errors were eliminated. Some counties with parentheses as part of the name, like *Fürth (Kreisfreie Stadt)*, had the Unicode replacement character U+FFFD instead of the *ü*. Third, the two polygons *Osterode am Harz* and *Göttingen* were merged, because their fusion, which happened on the 1st of November in 2016 [20], was not incorporated.

The polygon map was subsequently used to specify the weights w_{ji}. The rather simple assumption that disease transmission is only allowed from direct neighbors is used

[1] The data was downloaded as version 3.6 on the 15th of May in 2019 using the URL: `https://biogeo.ucdavis.edu/data/gadm3.6/Rsp/gadm36_DEU_2_sp.rds`

throughout all estimations. This restricts the weights to $w_{ji} = 1$ if county j and county i are directly adjacent, and $w_{ji} = 0$ otherwise. Note that $w_{ji} = w_{ij}$, given the symmetry of the neighbor relation.

4.3 Population Data

Population data for each county i and each week t is needed in the endemic component of the model for scaling purposes - it is the ϕ_{it} in $\phi_{it}\lambda_{it}^{EN}$.

I chose to use data provided by the Federal Statistical Office (*Statistische Bundesamt*) as part of the **census 2011** [40]. Several pieces of unnecessary information needed to be filtered out: Population information of larger, aggregated areas like *Regierungsbezirke* or *Bundesländer*, information about the gender split in each region, and the comparison with past population forecasts. The needed data entries were identified using the *Amtliche Gemeindeschlüssel*: Counties possess a five-digit code, while larger administrative areas have only two or three digits.

Each county's population is further divided by the total population of Germany. That is, the population fraction is used in the estimations. It is also assumed that this fraction is constant over time, resulting in $\phi_{it} \equiv \phi_i$.

4.4 Weather Data

A broad variety of meteorological measurements is available from the Climate Data Center (CDC) [9] run by the **Deutsche Wetterdienst** (DWD). The measurements of their weather stations are available to the public free of charge, according to the laws GeoNutzV §2 [7] and DWD-Gesetz §6 (2a) and §6 (6) [5].

The measurements of most weather features consist of a set of time series, one for each selected weather station, and are available for download in various versions: high-frequency data providing hourly measurements or aggregated versions providing daily, monthly or yearly data. Of those, monthly or yearly data could not be used because of the weekly frequency of the count data.[2]

[2] Monthly and yearly data could technically be used. However, a lot of variation and thus information would be needlessly thrown away.

The selected set of measures is shown in table 4.1. For each measure, the DWD already aggregated the hourly measurements to daily measurements by using the average, sum, minimum or maximum functions. Other available but not selected weather measures include: cloud amount, wind direction and wind strength, as well as amount and type of fresh-fallen snow.

For each measure a different number of weather stations is initially available, as documented in column 3. For example, precipitation is measured by 4278 stations, while sunlight measurements are only offered by 383 stations. Humidity and the four temperature measures are provided by around 700 different stations. However, many stations do not cover the whole selected data range from November 1, 2000, to January 7, 2019. This happens, among other things, when stations need to get repaired, are moved to an entirely different site or are not yet deployed on the 1st of November in 2000. After filtering out all those stations, only 34% of the initially available rain stations are available. For humidity and temperature about 57% remain and for sunlight 34%. The absolute numbers are shown in column 4 of table 4.1. Note that for some measures the number of available stations is already below 401, the number of different counties in Germany, without even accounting for the geographical spread of the weather stations.

Table 4.1: Used weather measures, their CDC file name and the number of available weather stations before and after filtering.

		Number of stations	
Weather measure and its measurement unit	File name	Before	After
Average humidity in %	UPM_MN004	714	414
Sum of precipitation (rain and snow) in mm	RS_MN006	4278	1458
Sum of sunlight in hours	SDK_MN004	383	271
Maximum temperature in °C	TXK_MN004	683	392
Minimum temperature in °C	TNK_MN004	683	392
Average temperature in °C	TMK_MN004	714	413
Minimum temperature in °C at ground level	CTGK_MN004	697	406

Every meteorological variable is available as a daily measure. All temperature variables, except the minimum ground temperature, are measured at 2 meter height. The columns Before and After refer to the filtering of stations which do not cover the whole period 2000/11/01 - 2019/01/07.

4.5 Aligning the Data

Several data preprocessing procedures needed to be conducted before being able to plug in the various kinds of data, described in the sections above, into the same statistical estimation model. In particular, the count data, weather data and population data needs to be provided in matrices of the same size, one row for each week and one column for each county.

Preparing the population matrix was easy given the constant-over-time assumption. The population of each county just needed to stay the same in each week, resulting in a matrix with identical rows.

4.5.1 Aggregating Weather Measures

The daily weather data needed to be aggregated to be aligned to the weekly count data. I mostly used the same set of aggregation functions the DWD already used for the hourly-to-daily transformation. Potentially, every aggregation function could be used on any measure, however, I restricted the combinations to the somewhat more sensible ones. That is, I refrained from summing up temperature variables. The made choices are documented in table 4.2. For the precipitation measurements, henceforth called rain, a new aggregation function was used: The number of days with positive, i.e. existing, rainfall were counted. This variable can range from 0 (days) to 7 (days) and is called *rain ndays*.

4.5.2 Mapping Weather Stations to Counties

Given the incorporation of weather as a feature into a spatial-temporal estimation model, a mapping between the weather measurements and the counties needs to be established.

The data supplied by the Deutsche Wetterdienst mainly consists of two parts: The raw data and metadata. Each row of the raw data stores one measurement, its timestamp and the ID of the weather station conducting the measurement. The metadata is organized as a set of files, one for each weather station. It has technical information about the used instruments as well as information about the location, given as a triple of latitude, longitude and elevation.

To translate that triple to an address, I used the Nominatim (`https://nominatim.openstreetmap.org/`) search engine for **OpenStreetMap** (OSM) data [28]. However,

Table 4.2: Names of the weather variables used in the estimations, and their descriptions.

№	Name of the variable	Description
1	Humidity	Average humidity
2	Rain mean	Average fallen rain
3	Rain ndays	Number of days in week some rain has fallen
4	Sun sum	Weekly sum of hours of sunlight
5	Mean($°C^{\text{Max}}$)	Average of the maximum temperature of each day
6	Max($°C^{\text{Max}}$)	Maximum temperature of the week
7	Mean($°C^{\text{Mean}}$)	Average of the average temperature of each day
8	Mean($°C^{\text{Min}}$)	Average of the minimum temperature of each day
9	Min($°C^{\text{Min}}$)	Minimum temperature of the week
10	Mean($°C^{\text{Min, 5cm}}$)	Average of the minimum temperature of each day
11	Min($°C^{\text{Min, 5cm}}$)	Minimum temperature of the week

The measure *rain* is officially called precipitation by the DWD and includes snowfall as well. Variables 5 to 9 are measured at 2m height. Variables 10 and 11 at 5cm height.

querying the county field from the returned address was often not successful. For some OSM addresses, especially in cities, the county information was empty. For others it returned an administration area of a higher, more aggregated level than needed, e.g. *Regierungsbezirk* instead of *Stadt-* or *Landkreis*.

Therefore, I decided to extract the zip codes (*Postleitzahl*) from the OSM addresses and link those to the counties. The zip code information was hardly ever missing and OpenStreetMap data, compiled by `https://www.suche-postleitzahl.org/`, provided the needed information [29]. 8,173 zip code-county pairs for Germany were documented in the file.[3] However, some zip code areas overlap with county borders, resulting in 117 zip codes that are shared between two counties and 5 zip codes that are shared between three counties. In case of sharing, I decided that a city always had precedence over the region around the city (39 times) and if no city was involved, the zip code was given to the preceding county in alphabetical order.

In a given county, there can either be none, one, or more than one weather station be located (after filtering). If one weather station is available, it will, of course, be selected.

[3] Actually, 13,108 places in Germany, like villages or cities, were listed together with its zip code and county information. The zip code-county number is lower because several villages from the same county also share a zip code.

If more than one is available, the station with the lowest ID gets selected, and if none are available, the station closest to the county is selected. For computing that distance, the locations of counties missing a station is defined as the longitude and latitude of county's central node in the OSM data, which was again queried using Nominatim [28].

How many different stations remain and are used for each weather measure is documented in table 4.3. For comparison, recall that (count) data for 401 counties is used. Thus, in the case of humidity, each station needs to represent about 8.5 counties on average. In the case of precipitation, each station only needs to represent 1.4 counties on average. It is to be expected, that the resulting regional variation of the weather feature in the estimation will be higher the more stations are used.

Table 4.3: The number of different weather stations used for each weather measure.

Weather measure	Stations used
Average humidity in %	47
Sum of precipitation (rain and snow) in mm	281
Sum of sunlight in hours	79
Maximum temperature in °C	162
Minimum temperature in °C	161
Average temperature in °C	160
Minimum temperature in °C at ground level	122

For comparison: There are 401 counties in Germany.

4.5.3 Lagging Weather Variables

When forecasting the expected number of counts for week $t+1$, at some time in week t, the weather features for that week t cannot be used because of two reasons. First, the measurements of the current week are not yet available. Second it takes time for people to show symptoms, see a doctor, submit samples to a laboratory and notify the RKI. This 'reporting lag' leads to a situation that the cases reported in week t are mostly people falling ill during week $t-1$. Thus the weather of week $t-1$ (or earlier) is more relevant and should be used in the estimation. This usage of weather data from past periods, is called: using lagged weather variables.

When lagging the weather variables, attention to detail is needed. As mentioned in section 4.1, some years have a week 53 and the statistical estimation procedure is not able to

cope with a varying number of weeks per year. Thus, the counts for that extra week were deleted. However, the week 53 weather measurements cannot be deleted likewise. As illustrated in table 4.4, the weather data from x weeks prior to the week 53 needs to be deleted, where x is the lag order of the weather variable used. For example, when forecasting the counts for week 2 in 2010, also written 2010:02, the count data from the previous week 2010:01 is used in the estimation, together with weather data from some earlier week. If a one-week lag is used, the weather measure needs to be from 2009:53, and if a two-week lag is used, the weather measure needs to be from 2009:52 instead.

Table 4.4: Illustration of the difficulties arising from years with 53 weeks and the lagging of weather variables.

Actual ISO week	2010:02	2010:01	2009:53	2009:52	2009:51
Count data used	2010:02	2010:01	-	2009:52	2009:51
Weather lag(1)	2010:01	2009:53	-	2009:51	2009:50
Weather lag(2)	2009:53	2009:52	-	2009:50	2009:49

Dates are in the YYYY-WW format. Lag(x) means that the weather variable of x weeks prior is used.

4.5.4 Ensuring Identical Column Order

All the used data need to be provided with an identical column order $[c_1, c_2, ..., c_{401}]$, with c_i being column i. Expressed differently, if a given county finds its data to be in column i in the population matrix, the same county's data needs to be in column i in the count matrix, the weather matrix and the neighborhood matrix as well.[4]

Ensuring identical column orders comes with obstacles, though. The data hails from four different sources and each source uses slightly different names for the counties and no unique identifier shared by all sources exists to easily merge the data. To provide an idea how the county names vary across data sources, three counties were selected and its various spellings are shown in table 4.5.

In general, the spelling variations could be grouped in systematic and unsystematic disparities. An example for an unsystematic disparity can be found in column 2. The sources 2 and 4 have shortened parts of the county name and use *i.d.OPf.* instead of *in*

[4]In fact, the county's data need to be in row i of the neighborhood matrix too, given its symmetry.

der Oberpfalz. What makes this unsystematic, is the lack of a straightforward rule to not only identify counties that have their names shortened but to also provide the exact way of transforming a county's name from one version to the other. To still be able to find the corresponding, slightly changed spellings for each county, an algorithm that provides close matches instead of exact matches was used. It did well most of the time and the seldom cases for which it failed were spotted by manual checking the logs.

Systematic differences in the spelling across the data sources, are based on the way how the two types of counties, the Landkreis (rural district) and the Stadtkreis (urban district), are identified and distinguished: Source 1 uses the *(Stadtkreis)* suffix for urban districts and no prefix or suffix for rural districts. Source 2 uses the prefixes *LK* and *SK* for Landkreis and Stadtkreis, respectively. Source 3 and 4, on the other hand, use an additional variable to distinguish between the two types. This information needs to stay attached for identification purposes in the reordering process.

Table 4.5: The spellings used for three different counties in the four different data source.

Source	Example 1	Example 2	Example 3
1	Neumarkt in der Oberpfalz	Karlsruhe (Stadtkreis)	Karlsruhe
2	LK Neumarkt i.d.OPf.	SK Karlsruhe	LK Karlsruhe
3	Landkreis Neumarkt in der Oberpfalz	Karlsruhe	Karlsruhe
4	Neumarkt i.d.OPf	Karlsruhe	Karlsruhe

The sources are: [1] the Germany map (neighborhood), [2] the SurvStat-RKI data (counts), [3] the zip code to county data (weather) and [4] census data (population). The bottom two data sources use an additional identifier for differentiating between the urban and rural district of Karlsruhe.

5 Results

This chapter presents the results of several forecast experiments, each estimating the model equation $\mu_{it} = \lambda_{it}^{AR} Y_{i,t-1} + \lambda_{it}^{NE} \sum_{j \neq i} w_{ji} Y_{j,t-1} + \phi_{it} \lambda_{it}^{EN}$ with varying configurations of the endemic and autoregressive component, including various weather measures. The usefulness for forecasting is evaluated for each configuration using rolling one-week-ahead forecasts, $\hat{\mu}_{i,t+1}$, for the identical period: week 1 through week 52 of 2018. The forecasts are compared to the true counts y_{it} and averaged over time (52 weeks) and counties (401).

The evaluation measures used for comparison are the Mean Absolute Error (MAE)

$$MAE = \frac{1}{52} \sum_{t=1}^{52} \frac{1}{401} \sum_{i=1}^{401} |\hat{\mu}_{i,t} - y_{i,t}| \tag{5.1}$$

and the Mean Squared Error (MSE), which places a larger penalty on higher deviations.

$$MSE = \frac{1}{52} \sum_{t=1}^{52} \frac{1}{401} \sum_{i=1}^{401} (\hat{\mu}_{i,t} - y_{i,t})^2 \tag{5.2}$$

For each week t, the set of time points used to fit the model is adjusted by adding the newest time point and removing the oldest. Thus each one-week-ahead forecast is based on the same number of time points. Models computing the forecast for week $t > 1$ use the parameter estimates of the previous week's model as the starting point of the optimization procedure. The model for week 1 uses estimates of a corresponding basic model, which differs only by using less or no weather features.

5.1 Calibration Period

While the evaluation period is fixed to 2018:01 to 2018:52, it is unclear how many weeks of data should be used to fit the model when conducting a one-week-ahead forecast.

To answer that question, the following very basic form of the model equation 3.2 will be used, with ϕ_i being the population fraction in county i.

$$\mu_{it} = \lambda^{AR} Y_{i,t-1} + \lambda^{NE} \sum_{j \neq i} w_{ji} Y_{j,t-1} + \phi_i \lambda_t^{EN} \tag{5.3}$$

$$log(\lambda^{AR}) = \alpha^{AR} \tag{5.4}$$

$$log(\lambda^{NE}) = \alpha^{NE} \tag{5.5}$$

$$log(\lambda_t^{EN}) = \alpha^{EN} + \beta_t t + \beta_1 \sin(\frac{t}{52} 2\pi) + \beta_2 \cos(\frac{t}{52} 2\pi) \tag{5.6}$$

$$w_{ji} = 1 \text{ if distance between } j \text{ and } i \text{ is 1 and 0 otherwise.} \tag{5.7}$$

Each component's intercept, α^τ with $\tau \in \{\text{AR, EN, NE}\}$, is assumed to be time-invariant and identical across regions. Seasonality is only specified in the endemic component, modeled as one sine-cosine pair and a time trend t (see equation 5.6). Transmission in the neighborhood component is only allowed from directly adjacent counties (see equation 5.7). Note that the population fraction ϕ_i provides variation across counties for the endemic component, despite λ_t^{EN} being identical for all counties.[1]

The various tested lengths of calibration periods (called sample periods) and the corresponding MSE and MAE figures are shown in table 5.1. Only multiples of full years were considered, with a minimum of 2 years.[2] For influenza, the sample was restricted to data starting January of 2008, because an expansion of the notification requirements in March of 2007 resulted in a structural break of the time series.

It can be seen that, for both diseases, all forecast error measures in table 5.1 are declining from top to bottom. That is, the lowest (best) value in each column is achieved with the sample period 2016:01 till 2017:52. The dominance of shorter sample periods over longer

[1] Note that multiplying with ϕ_i in equation 5.3 is the same as using $log(\phi_i)$ in the log-linear form of equation 5.6.

[2] Strictly speaking, the number of weeks of count data which needs to be provided for a sample period of x years is $x * 52 + 1$. The additional week is needed because of the $Y_{i,t-1}$ in the AR- and NE-component.

Table 5.1: Various sample period lengths and the resulting MSE and MAE.

		Campylobacter		Influenza	
Sample Period	Length	MSE	MAE	MSE	MAE
2002:01 - 2017:52	16	9.69	1.979	-	-
2003:01 - 2017:52	15	9.64	1.972	-	-
2004:01 - 2017:52	14	9.58	1.964	-	-
2005:01 - 2017:52	13	9.57	1.964	-	-
2006:01 - 2017:52	12	9.49	1.953	-	-
2007:01 - 2017:52	11	9.48	1.951	-	-
2008:01 - 2017:52	10	9.42	1.947	632.5	6.517
2009:01 - 2017:52	9	9.34	1.944	635.5	6.566
2010:01 - 2017:52	8	9.26	1.939	575.1	6.059
2011:01 - 2017:52	7	9.22	1.938	580.3	6.046
2012:01 - 2017:52	6	9.11	1.933	563.1	5.987
2013:01 - 2017:52	5	8.90	1.911	562.0	5.952
2014:01 - 2017:52	4	8.72	1.896	548.4	5.928
2015:01 - 2017:52	3	8.47	1.883	547.3	5.908
2016:01 - 2017:52	2	**8.31**	**1.882**	**531.8**	**5.870**

The sample period is given in the YYYY:WW format. Column 2, Length, denotes the length of the sample period in years. MSE stands for Mean Squared Error and MAE for Mean Absolute Error; see equations 5.1 and 5.2, respectively.

ones, might be the result of a seasonality that changed over time (e.g. shape or timing of the peak). In that case recent years are more representative for forecasting than years further in the past.

All following experiments use data from the 2-year period 2016:01 to 2017:52 for fitting the week 1 (2018) forecast model. For week 2, the adjusted, same-length period 2016:02 - 2018:01 is used. For week 3, 2018, the period 2016:03 - 2018:02 is used, and so on.

5.2 Christmas and New Year

In this section, an idea from Held et al. is examined, who added an indicator variable for calendar weeks 52 and 1 to the endemic component when modeling norovirus and rotavirus incidence in Berlin [16]. The idea is "to capture changes in reporting behaviour or social contact patterns during the Christmas break" [4, p. 10].

Thus, equation 5.8 presents the new formulation of the endemic component that replaces equation 5.6. The only change is the added indicator variable d_t. Its effect is measured by β_d.

$$log(\lambda_t^{EN}) = \alpha^{EN} + \beta_t t + \beta_1 \sin(\frac{t}{52}2\pi) + \beta_2 \cos(\frac{t}{52}2\pi) + \beta_d d_t \\ \text{with } d_t = 1 \text{ in certain weeks and } 0 \text{ otherwise.} \tag{5.8}$$

Three different types of the indicator variable d_t are considered. First, the suggested indicator for calendar weeks 52 and 1, called *Last Week & First Week*. Second, a slightly adjusted version, called *Christmas & New Year*, which is 1 for weeks including the 25th of December or the 1st of January. Differences are seldom, but the adjusted version for example does not have a New Year in 2016, because of its removal as a part of calendar week 53, 2015.[3] As a third version, an indicator for calendar week 2 is tested. Instead of modeling a (potential) dip in reported cases over Christmas and New Year, the idea is to model a potential surge afterwards.

Results for the two diseases, Campylobacter and influenza, are shown in table 5.2. For comparison, the row *None* presents the MSE and MAE obtained when no indicator was used; those results were already shown in table 5.1.

[3] Calendar week 53 in 2015 ranges from December 28, 2015, to January 3, 2016.

Even though the coefficients for the two types *Last Week & First Week* and *Christmas & New Year* were always significantly different from zero, including an indicator variable of those types is not helpful for forecasting - the MSE is about the same as the one obtained by *None*. The only improvement was obtained by using the *Week 02* indicator for Campylobacter: The MSE decreased from 8.31 to 8.20. No improvement was detected for influenza. Perhaps, the increase of cases in week 2 for Campylobacter, picked up by the indicator variable, is a result of people getting infected preparing poultry meat (duck, goose, ...) for Christmas or New Year.[4]

Table 5.2: MSE and MAE for the Christmas - New Year experiment.

	Campylobacter		Influenza	
Type	MSE	MAE	MSE	MAE
None	8.31	1.882	**531.8**	**5.870**
Last Week & First Week	8.32	1.884	532.0	5.874
Christmas & New Year	8.32	1.881	531.9	5.878
Week 02	**8.20**	**1.871**	533.5	5.882

The column 'Type' specifies which weeks the dummy variable d_t marks with a 1. None of the 95%-confidence intervals for the dummy's coefficient contained the 0.

Therefore, the Week 02 indicator variable will stay in the endemic component for the following experiments - but only for Campylobacter. For influenza, the formulation from equation 5.6, which excludes any indicator variable, is used.

5.3 Weather Features

In this section, several model formulations are examined that introduce meteorological variables into the model. Simple extensions of the endemic component are considered first, before slightly more complex formulations, and extensions to the autoregressive component are tested.

[4] All coefficients for the Week 02 indicator variable were positive, meaning that there are some additional cases in week 2 which are not picked up by the sine-cosine pair modeling the seasonality.

5.3.1 Single Variable

The first formulation of the model allowing for weather effects, extends the endemic component by including one past week of one single weather variable. The weather variable is used as originally measured and not differenced, also called its level form.

The new endemic component's equation is shown in 5.9. Note that the weather variable $z_{i,t-j}$ brings new between-county variation that was not present before, leading to λ_{it} instead of λ_t. As for the weather variable, data from 1 week ago up to 5 weeks ago is tested to find the best lag $j^* \geq 1$ for $z_{i,t-j}$ providing the lowest MSE and MAE. Results are shown in table 5.3 for Campylobacter, and table 5.4 for influenza.

$$log(\lambda_{it}^{EN}) = \alpha^{EN} + \beta_t t + \beta_1 \sin(\frac{t}{52}2\pi) + \beta_2 \cos(\frac{t}{52}2\pi) + \beta_d d_t^{(\text{Campyl.})} + \beta_z z_{i,t-j} \quad (5.9)$$

with $d_t = 1$ in calendar weeks 2, otherwise 0. Only used for Campylobacter.

Table 5.3: MSE and MAE when adding one weather feature in the endemic component for Campylobacter.

	MSE					MAE				
Measure	L^1	L^2	L^3	L^4	L^5	L^1	L^2	L^3	L^4	L^5
Humidity	8.23	8.20	**8.19**	8.20	8.21	1.875	1.870	**1.869**	**1.869**	1.870
Rain mean	8.19	8.20	8.20	**8.18**	8.20	1.869	1.869	1.870	**1.868**	1.870
Rain ndays	**8.18**	8.19	8.19	8.20	8.20	**1.868**	1.869	1.869	1.870	1.870
Sunshine sum	8.23	8.20	**8.19**	**8.19**	8.21	1.876	1.871	**1.869**	**1.869**	1.871
Mean($°C^{\text{Max}}$)	8.21	8.20	**8.17**	8.21	8.22	1.873	1.872	**1.868**	1.872	1.873
Max($°C^{\text{Max}}$)	8.21	8.20	**8.18**	8.21	8.21	1.873	1.871	**1.868**	1.873	1.872
Mean($°C^{\text{Mean}}$)	8.20	8.22	**8.19**	8.22	8.21	1.872	1.874	**1.870**	1.873	1.873
Mean($°C^{\text{Min}}$)	**8.19**	8.23	8.21	8.22	8.21	**1.871**	1.873	1.872	1.872	1.872
Min($°C^{\text{Min}}$)	**8.18**	8.21	8.22	8.21	8.21	**1.869**	1.871	1.872	1.871	1.871
Mean($°C^{\text{Min, 5cm}}$)	**8.19**	8.24	8.22	8.23	8.21	**1.870**	1.873	1.873	1.872	1.872
Min($°C^{\text{Min, 5cm}}$)	**8.18**	8.21	8.23	8.21	8.21	**1.869**	1.870	1.872	1.870	1.871

L^x denotes the lag operator, meaning that the weather measure of x weeks prior was used. Rain ndays = The number of days with precipitation > 0.

Note that each column's title indicates the lag length of the weather feature using the lag operator notation, as defined e.g. by [14, ch. 2]. The lag operator L denotes that a measure one time point prior is used, that is $Lz_t = z_{t-1}$. If the operator is applied j times, it becomes $L^j = L^{j-1}(Lz_t) = L^{j-1}z_{t-1} = ... = z_{t-j}$.

For each weather measure (row) the best forecast error value is presented in bold. Overall best results are additionally underlined. Judging by the MSE values, the overall best results for Campylobacter are obtained with Mean($°C^{\text{Max}}$) at lag length 3. The achieved 8.17 is a tiny bit better than the 8.20 from the model without a weather variable included (see table 5.2). For MAE, there is a four-way tie for the best forecast model: Mean($°C^{\text{Max}}$) at lag length 3, Max($°C^{\text{Max}}$) at lag length 3, Rain ndays at lag length 1 and Rain mean at lag length 4 all achieve an error of 1.868. For each weather variable, MSE and MAE consider the same lag lengths to be the best.

Table 5.4: MSE and MAE when adding one weather feature in the endemic component for Influenza.

	MSE					MAE				
Measure	L^1	L^2	L^3	L^4	L^5	L^1	L^2	L^3	L^4	L^5
Humidity	533.3	532.1	532.0	532.2	**531.8**	5.874	5.870	5.874	5.872	**5.869**
Rain mean	531.4	**530.9**	531.1	532.0	531.9	**5.870**	5.871	5.873	5.873	5.873
Rain ndays	530.0	**529.9**	531.4	532.0	531.7	**5.865**	5.866	5.874	5.875	5.869
Sunshine sum	**529.8**	531.4	532.1	532.7	531.9	**5.863**	5.869	5.878	5.876	5.872
Mean($°C^{\text{Max}}$)	529.4	**529.3**	532.1	532.1	532.9	**5.854**	5.860	5.874	5.872	5.887
Max($°C^{\text{Max}}$)	531.5	**530.7**	531.9	532.0	532.6	5.867	**5.864**	5.871	5.869	5.880
Mean($°C^{\text{Mean}}$)	**528.6**	**528.6**	532.2	532.0	532.9	**5.853**	5.860	5.876	5.870	5.887
Mean($°C^{\text{Min}}$)	533.3	532.1	532.0	532.2	**531.8**	5.874	5.870	5.874	5.872	**5.869**
Min($°C^{\text{Min}}$)	529.8	**529.1**	532.0	532.1	532.7	**5.863**	5.869	5.876	5.872	5.884
Mean($°C^{\text{Min, 5cm}}$)	530.0	**529.1**	531.7	532.1	532.9	**5.864**	5.866	5.878	5.872	5.888
Min($°C^{\text{Min, 5cm}}$)	528.9	**528.6**	532.1	532.1	532.7	**5.856**	5.861	5.878	5.870	5.883

L^x denotes the lag operator, meaning that the weather measure of x weeks prior was used. Rain ndays = The number of days with precipitation > 0.

For influenza, presented in table 5.4, both forecast error measures agree that Mean($°C^{\text{Mean}}$) at lag length 1 is the best. When judging by MSE, two same-scoring alternatives exist: Mean($°C^{\text{Mean}}$) at lag length 2 instead of 1 and Min($°C^{\text{Min, 5cm}}$) using L^2. The obtained MSE of 528.6 is hardly an improvement over 531.8, which was achieved without weather

variables.

Overall, both result tables show a remarkable low level of variation between the measured forecast errors. It hardly matters what weather measure or lag length is used.

In the following experiments, reporting of MAE results is omitted given the agreement between the two error measures. The MAE scores were nevertheless computed and checked to detect a situation of disagreement (none found). Additionally, only lag lengths of 1 to 4 are considered, given the very low variation and the slight preference for lower lag lengths.

5.3.2 Polynomial

In this section, the equation of the endemic component is changed to allow (more) non-linearity by incorporating a third-degree polynomial in the weather feature $z_{i,t-j}$.[5] The third-degree was chosen to balance the need for flexibility - a cubic function can have one inflection point - while limiting the possibility of over-fitting and keeping estimation times low.

The new equation is presented in 5.10. As before, the indicator for calendar week 2 d_t is only used for Campylobacter and the best lag $j^* \geq 1$ for the weather variable needs to be determined. Note that the data of all three parts of the polynomial is from the same week, i.e. $z_{i,t-j}$, $z^2_{i,t-j}$ and $z^3_{i,t-j}$ have the identical j.

$$log(\lambda^{EN}_{it}) = \alpha^{EN} + \beta_t t + \beta_1 \sin(\frac{t}{52}2\pi) + \beta_2 \cos(\frac{t}{52}2\pi) + \beta_d d_t^{(\text{Campyl.})} + \beta_{z_1} z_{i,t-j} + \beta_{z_2} z^2_{i,t-j} + \beta_{z_3} z^3_{i,t-j} \quad (5.10)$$

The results for Campylobacter and influenza are presented together in table 5.5. No MSE values are shown for humidity as the estimation procedure had problems to converge.

For Campylobacter, the MSE is not better than using only one single variable. In fact, for each weather measure there is a lag length configuration which provides a better or the same MSE by omitting the quadratic and cubic terms $z^2_{i,t-j}$ and $z^3_{i,t-j}$. For Influenza, the MSE values are a little lower. For example, the temperature measure Min($°C^{\text{Min, 5cm}}$) achieves an MSE of 521.9, which is equivalent to an improvement of 7.2 or 1.36%.

[5] The relationship between the counts and weather was already non-linear because the mean of an underlying negative binomial distribution is modeled, instead of the counts itself.

Table 5.5: MSE, separately for Campylobacter and Influenza, when using a 3rd-degree polynomial in the weather measure.

	Campylobacter				Influenza			
Measure	L^1	L^2	L^3	L^4	L^1	L^2	L^3	L^4
Humidity	-	-	-	-	-	-	-	-
Rain mean	8.19	8.19	8.20	**8.18**	530.3	**528.5**	530.8	532.1
Rain ndays	8.21	**8.19**	**8.19**	8.21	529.7	**529.5**	531.2	532.4
Sun sum	8.24	**8.20**	**8.20**	**8.20**	**530.7**	531.4	532.7	532.2
Mean($°C^{\text{Max}}$)	8.31	8.25	**8.19**	8.22	**525.8**	528.6	536.2	534.6
Max($°C^{\text{Max}}$)	8.29	8.24	**8.19**	8.23	532.3	**529.8**	533.9	533.8
Mean($°C^{\text{Mean}}$)	8.30	8.27	**8.20**	8.22	**524.1**	527.3	535.5	534.3
Mean($°C^{\text{Min}}$)	8.25	8.25	**8.22**	**8.22**	**523.4**	526.3	534.0	533.6
Min($°C^{\text{Min}}$)	8.23	**8.22**	**8.22**	**8.22**	**525.7**	526.2	533.8	533.6
Mean($°C^{\text{Min, 5cm}}$)	8.25	8.26	8.23	**8.22**	**521.9**	525.4	533.7	533.3
Min($°C^{\text{Min, 5cm}}$)	8.22	8.22	8.23	**8.21**	**526.3**	527.5	533.3	533.7

Lx denotes the lag operator, meaning that the weather measure of x weeks prior was used. Rain ndays = The number of days with precipitation > 0. There are no results for humidity as the estimation procedure had problems to converge.

5.3.3 Pairs

In this section, the equation of the endemic component is changed to incorporate two weather features instead of only one. As can be seen in the new formulation in equation 5.11, there are two weather features $z^{(1)}$ and $z^{(2)}$ as well as the interaction $z^{(1)}z^{(2)}$. There are no quadratic or cubic terms.

$$log(\lambda_{it}^{EN}) = \alpha^{EN} + \beta_t t + \beta_1 \sin(\frac{t}{52}2\pi) + \beta_2 \cos(\frac{t}{52}2\pi) + \beta_d d_t^{(\text{Campyl.})} + \beta_{z_1} z_{i,t-j}^{(1)} + \beta_{z_2} z_{i,t-j}^{(2)} + \beta_{z_1,z_2} z_{i,t-j}^{(1)} z_{i,t-j}^{(2)} \quad (5.11)$$

As before, d_t is an indicator for calendar week 2 and only used for Campylobacter. To determine the MSE minimizing $j^* \geq 1$, lag lengths of 1 to 4 are considered. The two weather features and the interaction term need to be from the same past week. To lower the amount of pairings, only one rain and one temperature measure is selected, based on the single variable results for Campylobacter (see table 5.3).

Table 5.6: MSE, separately for Campylobacter and Influenza, when using pairs of weather features.

	Campylobacter				Influenza			
Measure	L^1	L^2	L^3	L^4	L^1	L^2	L^3	L^4
Humidity & Rain mean	8.22	**8.19**	**8.19**	**8.19**	532.9	**531.3**	531.4	532.5
Humidity & Sun sum	8.23	**8.19**	8.20	8.20	**530.9**	531.4	533.3	533.4
Humidity & Mean(°C^{Max})	8.26	8.23	**8.19**	8.21	536.0	534.3	**532.7**	533.0
Rain mean & Sun sum	8.24	8.20	8.20	**8.19**	**529.8**	531.0	531.8	533.0
Rain mean & Mean(°C^{Max})	8.21	**8.19**	**8.19**	**8.19**	**525.4**	525.6	531.9	532.2
Sun sum & Mean(°C^{Max})	8.26	8.25	**8.20**	8.22	537.6	534.2	534.2	**533.3**

L^x denotes the lag operator, meaning that the weather measure of x weeks prior was used.

For Campylobacter, the results from the pair experiment shown in table 5.6 are in line with the 8.18 achieved by the polynomial formulation (see table 5.5) and the 8.17 achieved by using a single weather measure (see table 5.3). All pairs except for the last one reached 8.19.

For Influenza, most of the scores are even worse than the ones achieved in table 5.4 using only one single variable, except for the Rain mean & Mean($°C^{\text{Max}}$) pairing. That pair achieved a MSE of 525.4, which is better than the best single weather measure score of 528.6 but not as good as the polynomial results (521.9).

Browsing the logs, which document the coefficient estimates of each one-week-ahead forecast, reinforces the insight that using pairs is not helpful. Several of the weather variables and/or the interaction term were insignificant. Thus, the pairing of weather measures does not seem to provide any extra information or exploitable variation. It could be that the weather features are too heavily correlated, or that the functional form assumption - the additive individual effects in level form and the multiplicative interaction term - is way too restricting or just wrong.

5.3.4 First Differences

In this section, a new formulation of the endemic component is proposed and tested, in which the weather variable is used in its differenced form instead of its level form. Differencing a variable with a time dimension like z_t means computing the difference between the value at time t and the value of the same variable at some earlier time $t-j$, with $j > 1$. The variable is said to be first-differenced if $j = 1$ and thus the difference between the value at t and $t-1$ is computed: $\Delta z_t = z_t - z_{t-1}$, which is just the one-period change.

Adding first differences in the weather variable to the endemic component will give us equation 5.12.

$$log(\lambda_{it}^{EN}) = \alpha^{EN} + \beta_t t + \beta_1 \sin(\frac{t}{52}2\pi) + \beta_2 \cos(\frac{t}{52}2\pi) + \beta_d d_t^{(\text{Campyl.})} + \beta_z \Delta z_{i,t-j} \quad (5.12)$$

As before, d_t is an indicator for calendar week 2 and only used for Campylobacter, and lag lengths of 1 to 4 are considered to determine the MSE minimizing $j^* \geq 1$.

The results, presented in table 5.7, only slightly differ from the other results. The best scores are between 8.19 and 8.21 for Campylobacter and between 529.9 and 531.8 for influenza - a tiny bit worse compared to the single variable in its level form formulation (see tables 5.3 and 5.4). Several coefficients were insignificant.

Table 5.7: MSE, separately for Campylobacter and Influenza, when using a single weather features in first difference form.

	Campylobacter				Influenza			
Measure	$j=1$	$j=2$	$j=3$	$j=4$	$j=1$	$j=2$	$j=3$	$j=4$
Humidity	**8.21**	**8.21**	**8.21**	**8.21**	532.2	533.0	**531.6**	532.2
Rain mean	8.21	8.21	8.21	**8.20**	532.0	532.0	**531.8**	532.0
Rain ndays	**8.20**	**8.20**	8.21	**8.20**	532.0	531.9	**531.7**	**531.7**
Sun sum	8.21	**8.20**	8.21	**8.20**	531.6	533.5	**531.2**	531.9
Mean($°C^{\text{Max}}$)	8.20	8.20	**8.19**	8.24	532.3	**530.6**	531.8	532.2
Max($°C^{\text{Max}}$)	**8.19**	8.21	8.20	8.24	532.0	**531.5**	531.6	532.5
Mean($°C^{\text{Mean}}$)	8.21	**8.19**	**8.19**	8.24	532.5	**529.9**	531.6	532.1
Mean($°C^{\text{Min}}$)	8.23	**8.19**	**8.19**	8.23	532.5	**530.4**	531.5	532.0
Min($°C^{\text{Min}}$)	8.22	**8.20**	8.21	8.21	532.6	**530.7**	531.5	531.9
Mean($°C^{\text{Min, 5cm}}$)	8.24	**8.19**	**8.19**	8.22	532.5	**530.8**	531.3	531.9
Min($°C^{\text{Min, 5cm}}$)	8.23	**8.20**	8.21	8.21	532.5	531.4	**531.0**	531.8

$j = x$ refers to the value for j in equation 5.12. Rain ndays = Number of days with precipitation > 0.

5.3.5 Three Lagged First Differences

Instead of using information of only one past weather measure change, three consecutive weekly changes are considered in this section. The resulting formulation of the endemic component is given in equation 5.13.

$$log(\lambda_{it}^{EN}) = \alpha^{EN} + \beta_t t + \beta_1 \sin(\frac{t}{52}2\pi) + \beta_2 \cos(\frac{t}{52}2\pi) + \beta_d d_t^{(\text{Campyl.})} + \beta_{z_1}\Delta z_{i,t-j} + \beta_{z_2}\Delta z_{i,t-j-1} + \beta_{z_3}\Delta z_{i,t-j-2} \quad (5.13)$$

As before, d_t is an indicator for calendar week 2 and only used for Campylobacter, and values 1 to 4 are considered for j. Results are shown in table 5.8. For Campylobacter, the MSE scores are almost equal to the results in 5.7 when using only a single past change. The MSE score is at most 0.01 lower when using three weekly changes. For Influenza, the MSE score tends to be a bit lower when three differenced terms are used instead of only one. The improvement is especially noticeable for the temperature measures Mean($°C^{\text{Mean}}$)

and Mean($°C^{\text{Min}}$). However, the results are not as good as in the formulation using a 3-rd degree polynomial (521.9).

Table 5.8: MSE, separately for Campylobacter and Influenza, when using the changes in the weather variable of three consecutive past weeks.

	Campylobacter				Influenza			
Measure	$j=1$	$j=2$	$j=3$	$j=4$	$j=1$	$j=2$	$j=3$	$j=4$
Humidity	8.22	**8.21**	8.22	8.22	534.8	533.2	**532.0**	532.5
Rain mean	8.21	**8.20**	8.21	8.21	532.0	531.7	531.8	**530.7**
Rain ndays	**8.20**	**8.20**	8.21	8.21	531.0	**530.1**	531.2	531.8
Sun sum	8.21	**8.20**	8.21	8.22	532.6	532.5	**531.2**	532.2
Mean($°C^{\text{Max}}$)	**8.17**	8.21	8.22	8.25	**529.3**	529.9	532.5	532.6
Max($°C^{\text{Max}}$)	**8.18**	8.23	8.24	8.25	**530.9**	531.4	533.1	533.0
Mean($°C^{\text{Mean}}$)	**8.18**	8.21	8.23	8.24	**527.6**	528.4	532.3	532.6
Mean($°C^{\text{Min}}$)	**8.20**	8.21	8.22	8.22	**527.5**	528.4	532.0	532.6
Min($°C^{\text{Min}}$)	**8.21**	**8.21**	8.22	**8.21**	**529.2**	529.4	531.9	532.4
Mean($°C^{\text{Min, 5cm}}$)	8.22	**8.20**	8.21	8.21	**528.0**	528.5	531.9	532.6
Min($°C^{\text{Min, 5cm}}$)	8.22	**8.21**	8.22	**8.21**	**529.5**	529.6	530.9	531.8

$j = x$ refers to the value for j in equation 5.13. Rain ndays = Number of days with precipitation > 0.

Overall none of the formulations provided much - if any - success. One feature all formulations had in common, is that the endemic component was used to incorporate the weather variables. The next section will explore how and if results differ when the autoregressive component is used instead.

5.3.6 Autoregressive Component

This section will present the results from two model formulations that incorporate the weather features into the autoregressive component instead of the endemic component. Therefore, the endemic component will be the one shown in equation 5.14. It still captures seasonality with the sine-cosine pair and has the calendar week 2 indicator d_t for Campylobacter.

$$log(\lambda_t^{EN}) = \alpha^{EN} + \beta_t t + \beta_1 \sin(\frac{t}{52}2\pi) + \beta_2 \cos(\frac{t}{52}2\pi) + \beta_d d_t^{(\text{Campyl.})} \tag{5.14}$$

For the autoregressive component, two different formulations are considered. The first, shown in equation 5.15, adds a single weather feature in its level form.

$$log(\lambda_{it}^{AR}) = \alpha^{AR} + \beta_{z_1} z_{i,t-j} \tag{5.15}$$

The second formulation is presented in equation 5.16. It adds three past weeks of the differenced weather feature.

$$log(\lambda_{it}^{AR}) = \alpha^{AR} + \beta_{z_1} \Delta z_{i,t-j} + \beta_{z_2} \Delta z_{i,t-j-1} + \beta_{z_3} \Delta z_{i,t-j-2} \tag{5.16}$$

Again, values from 1 to 4 were considered for j. When incorporating the weather variable into the autoregressive component, one has to keep in mind that the λ_{it}^{AR} is multiplied by $Y_{i,t-1}$ in the model equation $\mu_{it} = \lambda_{it}^{AR} Y_{i,t-1} + \lambda_{it}^{NE} \sum_{j \neq i} w_{ji} Y_{j,t-1} + \phi_{it} \lambda_{it}^{EN}$. While it does not make much sense at first glance to use weather features from any other week than $t-1$, effects like reporting delays and incubation times of the diseases potentially lead to a misalignment of the count and weather data. Cases reported for week t fell ill before, probably because of weather effects in $t-1$ or earlier.

Table 5.9: MSE, separately for Campylobacter and Influenza, when using a single weather variable in the autoregressive component.

	Campylobacter				Influenza			
Measure	$j=1$	$j=2$	$j=3$	$j=4$	$j=1$	$j=2$	$j=3$	$j=4$
Humidity	**8.17**	8.18	8.18	8.18	796.7	796.4	621.4	**533.8**
Rain mean	8.26	8.24	8.23	**8.22**	**510.0**	529.3	540.2	533.6
Rain ndays	8.23	8.24	8.23	**8.22**	525.9	556.0	559.7	**508.9**
Sun sum	**8.14**	8.15	8.18	8.17	776.9	823.5	625.1	**503.5**
Mean($°C^{\text{Max}}$)	**8.09**	8.13	8.16	8.12	917.2	1661.5	1329.1	**834.2**
Max($°C^{\text{Max}}$)	**8.10**	8.14	8.16	8.13	**548.6**	558.3	743.5	727.7
Mean($°C^{\text{Mean}}$)	**8.07**	8.10	8.15	8.12	1170.3	2072.2	1335.1	**775.9**
Mean($°C^{\text{Min}}$)	**8.07**	8.11	8.16	8.13	1188.0	1805.9	1061.3	**650.3**
Min($°C^{\text{Min}}$)	**8.08**	8.13	8.17	8.15	830.5	1098.6	906.5	**652.2**
Mean($°C^{\text{Min, 5cm}}$)	**8.08**	8.13	8.17	8.14	886.2	1320.9	878.9	**613.0**
Min($°C^{\text{Min, 5cm}}$)	**8.09**	8.13	8.17	8.15	681.2	872.5	778.2	**626.4**

$j = x$ refers to the value for j in equation 5.15. Rain ndays = Number of days with precipitation > 0.

The results for the first, single variable formulation are shown in table 5.9. For Campylobacter the MSE scores are lower for most measures than in any other formulation tested before. The overall best score of 8.07 is 0.1 points better than the 8.17 scored in the single variable formulation using the endemic component. The best j is $j = 1$, i.e. the change from the beginning to the end of past week ($t - 2$ to $t - 1$).

For influenza, the variations in MSE are huge. There is variation between measures, with rain measures being better than temperature measures, and variation across the different j. Unlike Campylobacter, influenza favors a high j of 4. The lowest MSE for influenza is 503.5, when the weekly sum of sunlight from 4 weeks prior is used. This is lower than the lowest MSE achieved when using the endemic component (521.9 with the polynomial formulation). However, some measure - lag length combinations are multitudes worse than any other, formerly used formulation, including models without any weather variables. It is concerning that very similar weather measures like Mean($°C^{\text{Max}}$) and Max($°C^{\text{Max}}$) can exhibit so different MSE scores.

Table 5.10: MSE, separately for Campylobacter and Influenza, when using three lags of the differenced weather feature in the autoregressive component.

	Campylobacter				Influenza			
Measure	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 1$	$j = 2$	$j = 3$	$j = 4$
Humidity	8.23	8.22	**8.21**	8.22	556.4	567.1	540.1	**508.4**
Rain mean	8.22	**8.21**	**8.21**	**8.21**	**483.7**	490.8	545.2	573.4
Rain ndays	**8.20**	**8.20**	**8.20**	8.22	452.7	**428.4**	542.3	572.0
Sun sum	**8.21**	**8.21**	**8.21**	8.23	**507.1**	555.7	550.3	531.4
Mean($°C^{\text{Max}}$)	**8.18**	8.23	8.24	8.25	**341.4**	533.1	634.8	661.1
Max($°C^{\text{Max}}$)	**8.20**	8.24	8.24	8.24	450.6	**362.8**	478.6	599.8
Mean($°C^{\text{Mean}}$)	**8.18**	8.22	8.24	8.23	**322.2**	613.1	734.9	692.1
Mean($°C^{\text{Min}}$)	**8.19**	8.21	8.22	**8.19**	**321.5**	619.7	734.8	652.6
Min($°C^{\text{Min}}$)	**8.19**	8.20	8.21	**8.19**	**330.3**	467.4	631.1	637.6
Mean($°C^{\text{Min, 5cm}}$)	8.19	8.20	8.21	**8.17**	**324.8**	503.5	644.0	630.0
Min($°C^{\text{Min, 5cm}}$)	8.20	8.20	8.21	**8.19**	**361.9**	421.7	567.6	627.4

$j = x$ refers to the value for j in equation 5.16. Rain ndays = Number of days with precipitation > 0.

The results for the second formulation of the autoregressive component, equation 5.16, with the 3 lagged differences in the weather features are shown in table 5.10. For Campylobacter, the obtained MSE are in line with the results from all the formulations using

the endemic component. The lower MSE value from the first autoregressive component experiment (8.07), table 5.9, could not be obtained again.

For influenza, the MSE values are the best values obtained so far, by a huge margin. The weather measure Mean($°C^{\text{Min}}$) achieved a score of 321.5 and several other measures scored well below 400. Even the 508.4 from humidity and 507.1 from sunlight are lower than any other MSE value. The variation across different j values is still noticeable, but not as large as in the first experiment using a single weather variable in the autoregressive component.

In the next section, the two different formulations used in this section will be joined by their corresponding endemic component versions. Desirable results are an even lower forecast error or less variation across different configurations.

5.3.7 Both Components

In this section, the meteorological features will simultaneously enter the model through the endemic and autoregressive component. The same two formulations from last section's autoregressive component tests are used, together with their endemic equivalent.

The two equations shown in 5.17 summarize the endemic and autoregressive component of the first model formulation. Each component features only a single weather variable used in its level form. Both components were already used separately before. The endemic component in section 5.3.1, and the autoregressive component in section 5.3.6.

$$\begin{aligned} log(\lambda_{it}^{EN}) &= \alpha^{EN} + \beta_t t + \beta_1 \sin(\frac{t}{52}2\pi) + \beta_2 \cos(\frac{t}{52}2\pi) + \beta_d d_t^{(\text{Campyl.})} + \beta_{z_{EN}} z_{i,t-j} \\ log(\lambda_{it}^{AR}) &= \alpha^{AR} + \beta_{z_{AR}} z_{i,t-j} \end{aligned} \tag{5.17}$$

The second model formulation, shown in equation 5.18, is equivalently composed of an endemic and an autoregressive component, which were already used separately. The former in section 5.3.5 and the latter in section 5.3.6.

$$
\begin{aligned}
log(\lambda_{it}^{EN}) = \alpha^{EN} &+ \beta_t t + \beta_1 \sin(\frac{t}{52}2\pi) + \beta_2 \cos(\frac{t}{52}2\pi) + \beta_d d_t^{\text{(Campyl.)}} \\
&+ \beta_{z_1}\Delta z_{i,t-j} + \beta_{z_2}\Delta z_{i,t-j-1} + \beta_{z_3}\Delta z_{i,t-j-2} \\
log(\lambda_{it}^{AR}) = \alpha^{AR} &+ \beta_{z_4}\Delta z_{i,t-j} + \beta_{z_5}\Delta z_{i,t-j-1} + \beta_{z_6}\Delta z_{i,t-j-2}
\end{aligned} \tag{5.18}
$$

As usual, d_t is an indicator for calendar week 2 and only used in the Campylobacter estimation. Both model formulations restrict the index variable j to be the same across components. That is, the weather feature in the endemic component needs to use data from the same prior week as the (same) weather feature in the autoregressive component - and vice versa.

Table 5.11: MSE, separately for Campylobacter and Influenza, when a single weather feature is used in the endemic and autoregressive component.

	Campylobacter				Influenza			
Measure	$j=1$	$j=2$	$j=3$	$j=4$	$j=1$	$j=2$	$j=3$	$j=4$
Humidity	8.15	8.11	**8.09**	8.11	802.2	807.5	640.9	**549.5**
Rain mean	8.25	8.24	8.24	**8.20**	**510.0**	533.5	542.1	533.5
Rain ndays	8.26	8.22	8.23	**8.20**	538.2	572.9	568.4	**511.5**
Sun sum	8.16	**8.10**	8.12	8.11	798.0	838.7	652.7	**521.7**
Mean($°C^{\text{Max}}$)	**8.09**	8.10	**8.09**	8.12	919.3	1621.1	1330.3	**840.4**
Max($°C^{\text{Max}}$)	8.09	8.10	**8.08**	8.12	**555.5**	562.8	749.5	735.8
Mean($°C^{\text{Mean}}$)	**8.07**	8.11	8.10	8.13	1184.4	2013.4	1344.0	**788.9**
Mean($°C^{\text{Min}}$)	**8.07**	8.13	8.14	8.15	1216.1	1756.2	1060.7	**659.0**
Min($°C^{\text{Min}}$)	**8.08**	8.13	8.17	8.15	863.4	1086.9	911.7	**661.4**
Mean($°C^{\text{Min, 5cm}}$)	**8.08**	8.15	8.16	8.17	928.8	1297.3	878.5	**622.6**
Min($°C^{\text{Min, 5cm}}$)	**8.09**	8.14	8.18	8.16	716.9	874.8	775.9	**637.1**

$j = x$ refers to the value for j in equation 5.17. Rain ndays = Number of days with precipitation > 0.

The results for the formulation using a single weather variable are shown in table 5.11. For Campylobacter, the MSE values are very similar to the ones obtained when the single weather feature was only used in the autoregressive component (see table 5.9). Almost all temperature variables achieve the exact same MSE value (between 8.07 and 8.09), with $j = 1$ being the preferred lag length. The measures humidity, sunlight, rain mean and

rain ndays show slight improvement. The former advantage of the temperature measures over those 4 partly vanished.

For influenza, the overall picture very much resembles the one from the autoregressive-only formulation. The best MSE values are obtained with the rain measures or sunlight and these 3 are the only measures for which the obtained MSE is better than in the endemic-only formulation. Including the weather feature additionally in the endemic component, as it is done in this both-component formulation, is useless though. All MSE values are worse than in the autoregressive-only formulation.

Table 5.12: MSE, separately for Campylobacter and Influenza, when using three lags of the differenced weather feature in the endemic and autoregressive component.

	Campylobacter				Influenza			
Measure	$j=1$	$j=2$	$j=3$	$j=4$	$j=1$	$j=2$	$j=3$	$j=4$
Humidity	8.24	**8.23**	**8.23**	**8.23**	530.5	557.0	542.2	**509.7**
Rain mean	8.22	**8.21**	**8.21**	**8.21**	**483.9**	490.8	548.3	577.1
Rain ndays	**8.20**	**8.20**	8.21	8.22	454.2	**431.9**	549.3	573.2
Sun sum	**8.22**	**8.22**	**8.22**	8.24	**501.1**	552.5	556.6	533.4
Mean($°C^{\text{Max}}$)	**8.17**	8.22	8.23	8.26	**337.0**	515.8	645.7	674.9
Max($°C^{\text{Max}}$)	**8.19**	8.25	8.24	8.26	453.8	**362.6**	481.7	611.6
Mean($°C^{\text{Mean}}$)	**8.17**	8.22	8.23	8.24	**319.8**	594.1	756.8	720.1
Mean($°C^{\text{Min}}$)	**8.19**	8.21	8.21	8.20	**324.8**	611.4	761.4	684.7
Min($°C^{\text{Min}}$)	8.21	8.21	8.22	**8.19**	**335.8**	462.8	642.9	654.5
Mean($°C^{\text{Min, 5cm}}$)	8.21	8.20	8.20	**8.18**	**329.5**	506.4	670.2	664.9
Min($°C^{\text{Min, 5cm}}$)	8.22	8.21	8.22	**8.20**	**366.0**	421.0	575.2	646.1

$j = x$ refers to the value for j in equation 5.18. Rain ndays = Number of days with precipitation > 0.

The results for the both-component version using the 3 lags of the first differenced weather variable (see equation 5.18) are presented in table 5.12. The MSE values obtained for both diseases do not really differ from those obtained in the autoregressive-only formulation, shown in table 5.10. The both-component formulation very slightly improved upon the autoregressive-only for some measures like the overall-best Mean($°C^{\text{Mean}}$). The endemic-only formulation showed about the same Campylobacter results but a lot worse MSE values for influenza (see table 5.8).

Concluding, using both components together does not show much promise. The driv-

ing force behind the results had been the autoregressive-only formulation in each case. Whenever the autoregressive-only formulation had an advantage over the endemic-only formulation, that advantage could be retained. But the both-component formulation was incapable of exploiting any advantage of the endemic-only formulation.

5.4 Robustness Check

While the incorporation of meteorological variables did not show much promise for most of the tested models, especially the ones using the endemic component, there were two exceptions. Using a single weather feature in the autoregressive component was useful for Campylobacter, and using 3 lags of a first-differenced temperature variable was useful for influenza. The best feature for both diseases had been Mean($°C^{\text{Min}}$). It lead to an MSE improvement from 8.20 to 8.07 for Campylobacter and 531.8 to 321.5 for influenza.

However, those two model formulations were only evaluated on data from 2018. While the whole evaluation period comprises 52 weeks, it is not much when a seasonality with a period length of 52-weeks is present - each time point (week) of the seasonality was evaluated only once. This section will therefore try to examine if the two model formulations are robust, that is if the model formulations are still good when other evaluation periods are used.

5.4.1 Different Evaluation Periods

To recapitulate, the two best model formulations are both using the autoregressive component to incorporate the meteorological variables into the model. Thus, the endemic component is without any weather features and just models the seasonality with a sine-cosine pair and, for Campylobacter only, the calendar week 2 effect d_t. The formulation from section 5.3.6 is repeated in equation 5.19.

$$log(\lambda_t^{EN}) = \alpha^{EN} + \beta_t t + \beta_1 \sin(\frac{t}{52}2\pi) + \beta_2 \cos(\frac{t}{52}2\pi) + \beta_d d_t^{(\text{Campyl.})} \tag{5.19}$$

The two formulations for the autoregressive component are equation 5.20 for Campylobacter and 5.21 for influenza. Both use the weather measure Mean($°C^{\text{Min}}$) as $z_{i,t}$. The

usually present lag length j was already replaced with the value 1, as it was determined to be the best.

$$log(\lambda_{it}^{AR}) = \alpha^{AR} + \beta_{z_1} z_{i,t-1} \tag{5.20}$$

$$log(\lambda_{it}^{AR}) = \alpha^{AR} + \beta_{z_1} \Delta z_{i,t-1} + \beta_{z_2} \Delta z_{i,t-2} + \beta_{z_3} \Delta z_{i,t-3} \tag{5.21}$$

Table 5.13: Comparison over various evaluation periods of a model for Campylobacter with the weather feature Mean($°C^{\text{Min}}$) in the autoregressive component against the same model without weather features.

Evaluation	MSE		MAE	
Year	With	Without	With	Without
2018	**8.07**	8.20	**1.87**	**1.87**
2017	**9.45**	9.62	**1.95**	1.96
2016	**7.57**	7.69	**1.76**	1.77
2015	**6.47**	6.52	**1.69**	**1.69**
2014	**6.89**	6.94	**1.73**	**1.73**
2013	**6.54**	6.64	**1.60**	**1.60**
2012	**5.78**	5.85	**1.58**	1.59
2011	**8.27**	8.52	**1.72**	1.73
2010	**7.04**	7.11	**1.64**	1.65
2009	**5.79**	**5.79**	**1.59**	**1.59**
2008	**6.36**	6.46	**1.63**	1.64
2007	**6.72**	6.80	**1.68**	1.69
2006	**4.91**	4.99	**1.44**	**1.44**
2005	**6.26**	6.31	**1.60**	1.60
2004	**5.41**	5.48	**1.48**	**1.48**

Each model was fit using 2 years prior to the evaluation year. The autoregressive component's formulation is equation 5.20.

The resulting MSE values for the first model formulation for Campylobacter from equation 5.20 is shown in table 5.13 together with the results from the corresponding model without any weather features. One can quickly detect that the model with the weather feature

was never worse than the model without. Judging by the MSE values, the model with was better for each evaluation period, except for 2009, for which it was equally good as the model without. The two models are more often tied when MAE is used instead of MSE. The size of the MSE differences are comparable across the different evaluation periods. Thus, 2018 was not exceptionally different from any other year before.

Table 5.14: Comparison over various evaluation periods of a model for influenza with three lags of the differenced weather feature Mean($°C^{\text{Min}}$) in the autoregressive component against the same model without weather features.

Evaluation	MSE		MAE	
Year	With	Without	With	Without
2018	**321.5**	531.8	**4.60**	5.87
2017	**74.1**	80.7	**2.24**	2.37
2016	**43.0**	45.7	**1.64**	1.70
2015	**75.7**	77.8	**2.00**	2.04
2014	**1.5**	**1.5**	**0.35**	**0.35**
2013	**45.5**	46.9	**1.78**	**1.78**
2012	**3.9**	**3.9**	**0.49**	**0.49**
2011	36.8	**32.4**	1.41	**1.36**
2010	4.8	**3.4**	0.60	**0.57**

Each model was fit using 2 years prior to the evaluation year. The autoregressive component's formulation is equation 5.21.

The same type of forecast error comparison over various evaluation years is repeated for the best influenza model, equation 5.21. The results are presented in table 5.14. Judging by MSE, the model 'with' was worse in 2010 and 2011, equally good in 2012 and 2014 and better for every other evaluation period. The size of the MSE values itself vary considerably over the years. Some years like 2010, 2012 and 2014 have extremely small MSE values, while it is enormous for 2018. Potentially, because the 2018 influenza season extraordinarily changed its form compared to the past two years or an extraordinary number of people got infected. In either case, both models made very large forecast errors in 2018, but the model with the weather measure did considerably better than the model without.

Given that the MSE for 2018 is at least four times as large as in any other years, there is a need to analyze in more detail why the weather model for influenza was so much better

in that particular year. The following section tries to shed light onto that question.

5.5 Detailed Examination of the Influenza Model for 2018

As the first inspection step, the obtained weekly MSE values were plotted for both influenza models in figure 5.1.

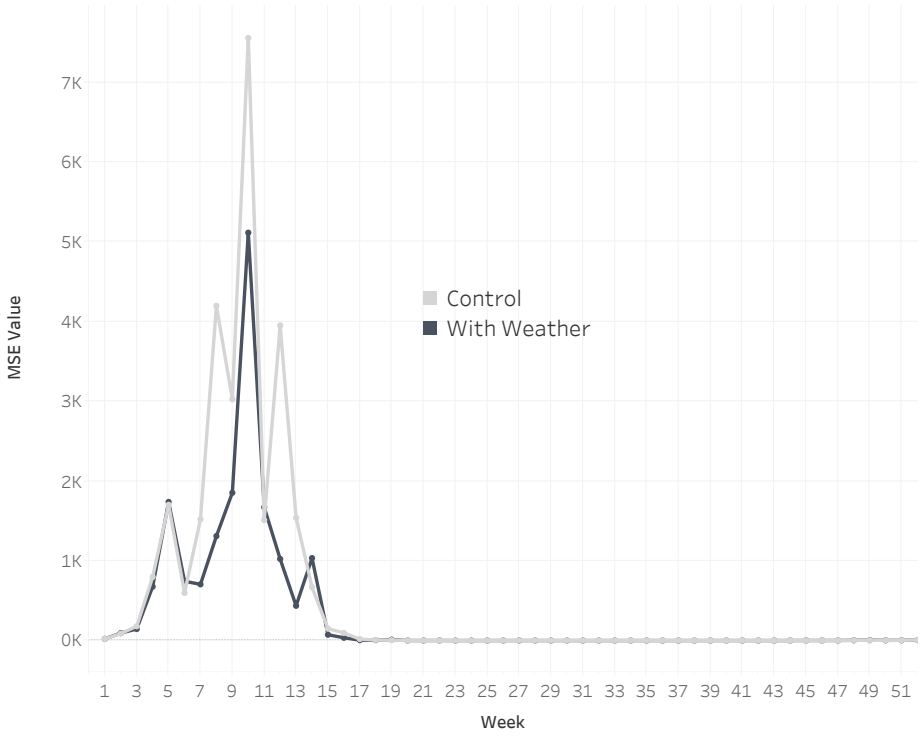


Figure 5.1: Weekly MSE values when evaluated for 2018, for the model including weather features and a control model.

The values obtained by the model with the weather feature are presented in dark, while the values obtained by the control model (without weather features) are presented in light gray. A huge variation of the MSE values across weeks can be seen. For weeks 16 to 52, almost no errors were made by both models. Between weeks 1 and 15, both lines start near 0, increase very fast until reaching a peak in week 10 and then fall as fast as

before towards zero. There are some weeks in the ascent when the MSE shortly decreases and likewise some weeks in the descent when the MSE shortly increases before the lines follow their original trajectory.

The dark line is either very close to the light gray line or considerably below. Thus, the weather model is equally good as the control model in most weeks and fares considerably better in some weeks. There are three weeks with an especially large disparity, that drive the overall advantage of the weather model: Week 12, week 8 and week 10. The MSE disparity between the control model and the model with weather are 3952 to 1023, 4197 to 1310 and 7555 to 5114, respectively.

5.5.1 Counties Driving the Result

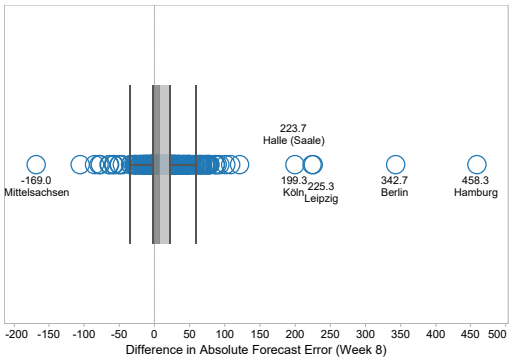
The next step of the inspection aims to detect which counties are most important in establishing the advantage of the weather model over the control model. For the three most important weeks, week 8, 10, 12, the difference between the absolute forecast errors of the two models is computed. The equation is given by 5.22.

$$\Delta\text{AE}_{it} = |\hat{\mu}_{i,t}^{\text{weather}} - y_{i,t}^{\text{weather}}| - |\hat{\mu}_{i,t}^{\text{control}} - y_{i,t}^{\text{control}}| \tag{5.22}$$

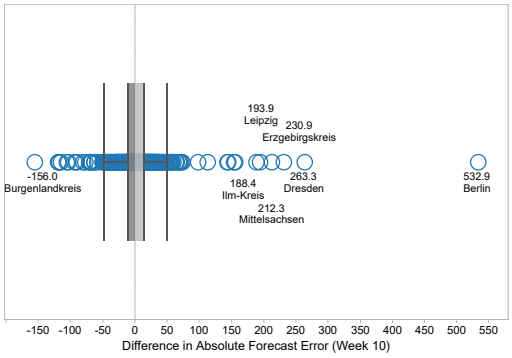
Two things are worthy to point out. First, ΔAE_{it} lacks the averaging over space and time and, thus, varies across counties and weeks. Second, it can be positive and negative. If $\Delta\text{AE}_{it} > 0$, then the weather model has a smaller absolute error for that specific county and week combination and if $\Delta\text{AE}_{it} < 0$, then the control model has the smaller absolute error.

Which ΔAE_{it} values the 401 counties obtain in week 8, 10 and 12 is plotted in figure 5.2. Each week is shown separately, week 8 in 5.2a at the top, week 10 in 5.2b in the middle and week 12 in 5.2c at the bottom. Outlier detection is supported by the box plot overlay in each graph.

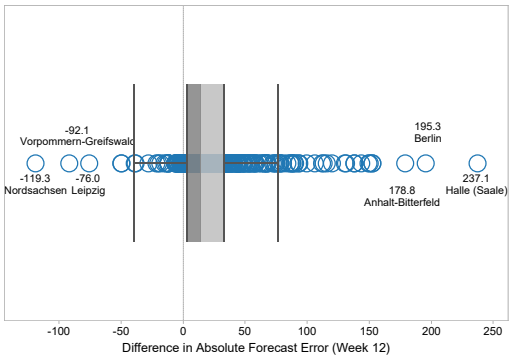
The median of the distribution is the value which splits the dark and light area in the box. The lower (left) end of the box is the 25%-percentile, also called lower quartile, and the upper (right) end is the 75%-percentile, or upper quartile. The whiskers extend to 1.5 IQR of the 25%-percentile to the left and 1.5 IQR of the 75%-percentile to the right. IQR is the interquartile range, the difference between the upper and the lower quartile, i.e. the width of the box.



(a) Week 8



(b) Week 10



(c) Week 12

Figure 5.2: Distribution of the county-specific absolute error differences between the weather and the control model, computed for week 8, 10 and 12 (2018).

The median of the ΔAE_{it} values is clearly positive (to the right of zero) for weeks 8 and 12. For week 10, the median is almost zero - its value is 0.2. The reason why week 10 still contributes so heavily to the advantage of the weather model over the control model, are large and positive outliers, especially Berlin. The ΔAE value for Berlin in week 10 is 532.9, which is very large given that this measures works with the absolute values instead of any squared error values. While there are other big and positive ΔAE as well, none comes close to the value of Berlin. It almost makes the average AE for week 10 positive by itself. The largest outliers for the other weeks are Hamburg (458.3) for week 8, with Berlin following on the second place, and Halle (Saale) with a value of 237.1 for week 12. The outliers for week 8, though, are much closer to the box plot than for the other two weeks.

5.5.2 Observed Cases Over Time

After figuring out which weeks and counties are the primary forces behind the advantage of the influenza weather model, it is about time to look at the observed cases, instead of the forecast error measures. Figure 5.3 plots the observed cases over time, from week 1, 2016, until week 52, 2018, for the three largest outlier counties in terms of ΔAE: Hamburg as the 'winner' of week 8, Berlin for week 10 and Halle (Saale) for week 12.

The plot offers three findings. First, the general shape of the time series plotting the observed cases pretty much resembles the time series when weekly MSE values are plotted. Thus, the more cases are observed, the higher the risk to make wrong forecasts and the higher the actual forecast error. The models have no problem with the summer and autumn weeks that have almost zero reported cases.

The second striking feature is that this general shape is quite similar for the three counties. The absolute number of cases is different, but so is the population in each county, which is used in the estimation. Thus, the made assumption of a seasonality component that is the same across counties, looks quite reasonable.

Third, the influenza season of 2018 had about 3 times as many reported cases compared to the two years before. The peak of each time series is a lot higher in 2018, which results in a relative importance of the autoregressive component of the model. Even if the endemic component's seasonality were to estimate a peak value of about 600 observed cases for Berlin based on the years 2016 and 2017, this would be 1200 cases too low in 2018.

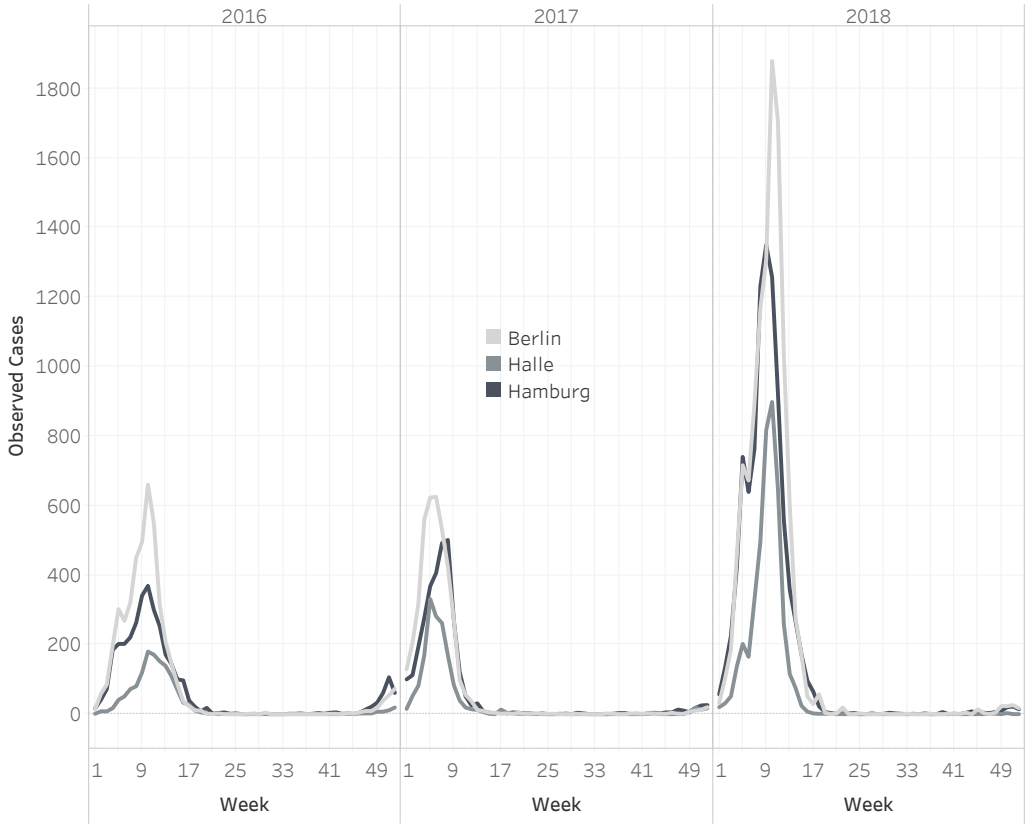


Figure 5.3: The weekly number of observed cases for three selected cities from 2016-01 to 2018-52.

5.5.3 Seasonality

The three figures 5.4, 5.5 and 5.6 were created to get an idea if and how the seasonal pattern changes over the year. The number of observed cases are plotted separately for each of the three counties, Berlin, Hamburg and Halle (Saale), but this time each year's time series share the same x-axis. Thus, for each x-value (week of year), there are 3 y-values (observed cases): One for 2016, one for 2017 and one for 2018.

Several similarities can be observed. For each county, there are (similar) differences in the height of the peak and the timing of the peak. The middle gray (2017) line has the

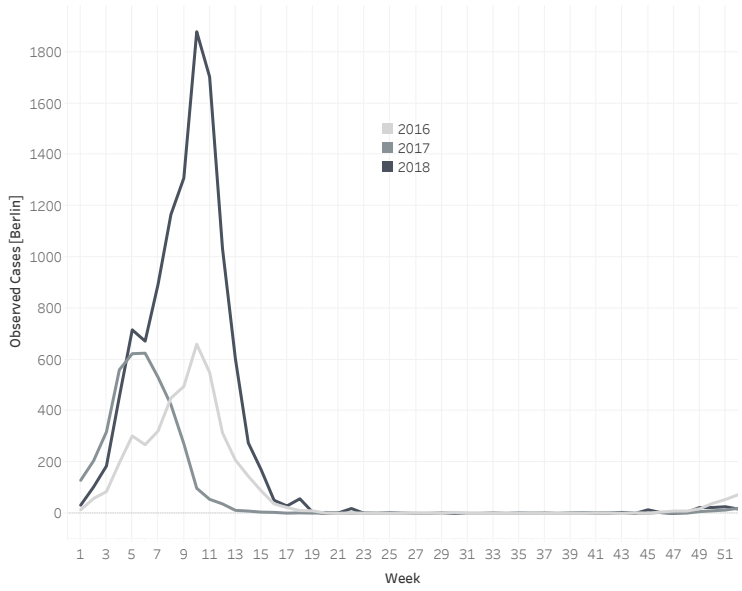


Figure 5.4: Weekly observed cases in Berlin for 2016, 2017 and 2018.

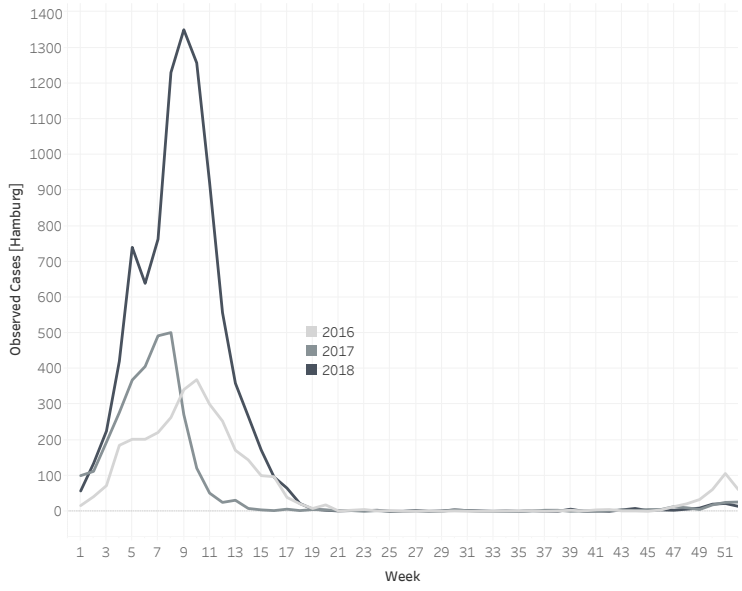


Figure 5.5: Weekly observed cases in Hamburg for 2016, 2017 and 2018.

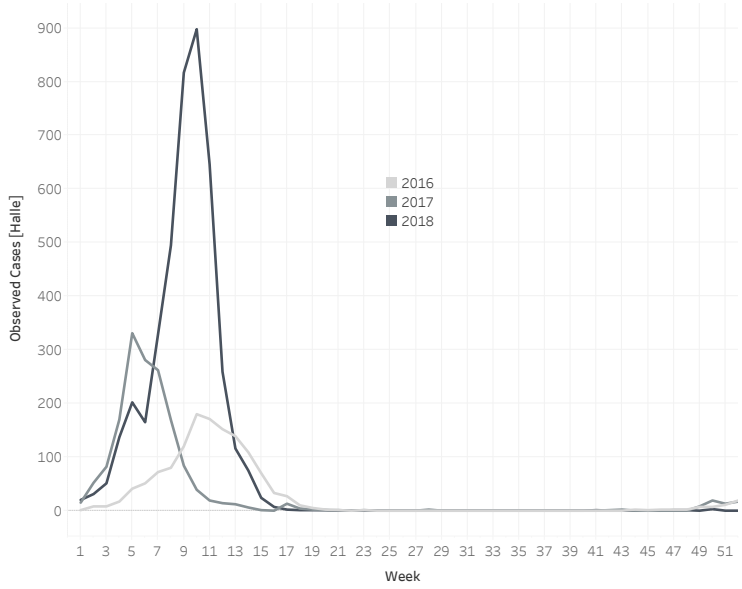


Figure 5.6: Weekly observed cases in Halle (Saale) for 2016, 2017 and 2018.

earliest peak, and the light gray (2016) line has the latest peak. The highest peak, by far, has the dark line (2018).

Differences in the general pattern between the three counties are rather small but do exist. For example Hamburg, see figure 5.5, experienced their influenza season peak in 2017 a bit later: week 8 instead of week 6 (Berlin) or week 5 (Halle). For Berlin, see figure 5.4, the peaks of the 2016 and 2017 influenza season had the same height, unlike the peaks in Hamburg or Halle, which both experienced a less severe influenza season in 2016 compared to 2017.

To sum up, there is a lot more variation of the observed number of cases over time than across counties for influenza. The variation over time not only concerns the absolute number of cases but also the timing, which makes the endemic component's seasonality very ill-equipped to forecast the correct number of cases. Seasonality is seen as something regularly occurring, with the same strength and timing. The observed shifts do not fit into that picture or the way seasonality is formulated in the model equation. The autoregressive component, on the other hand, has the necessary ignorance of calendar time to help forecasting too-early, too-strong or too-late increases or decreases.

Thus, it might come as little surprise that the two most (or only) promising models had incorporated the weather features via the autoregressive component.

5.5.4 Corresponding Weather

We have seen in section 5.3.6, together with the box plots in figure 5.2, that the weather measure Mean($°C^{\text{Min}}$) is helpful for forecasting the observed number of cases when plugged into the autoregressive component - especially for week 8, 10 and 12 of the influenza season in 2018 (see figure 5.1).

Weather data from that period, together with the observed number of cases and the predictions of the model incorporating the weather variable and the control model excluding it, are shown in the figures 5.7 (Berlin), 5.8 (Hamburg) and 5.9 (Halle).

Each upper figure is a time series plot of the observed counts and the predictions of the two models. Only the first 12 weeks of 2018 are shown to focus on the important weeks 8 to 12, which drive the results. The weather model predictions are in dark gray. The observed cases in middle gray and the predictions from the control model in light gray. Note that the y-axis scaling is different for each county, depending on the number of cases. Thus peaks that look the same at first glance, can be quite different when the y-axis values are taken into account - as seen in figure 5.3 when plotted together.

When looking at the upper figures, it stands out that the control time series (light gray) looks like a shifted-to-the-right version of the observed series (middle gray). Every kink of the observed series is present in the control series. However, the kinks and all other movements (accelerations, deceleration etc.) are one week too late. Well visible around the peak week, which is week 10 in Berlin and Halle, and week 9 in Hamburg. The observed cases decrease in the week after, while the control series is still increasing.

The weekly predictions of the weather model (dark gray series) look differently because of its ability to incorporate weather information. In this case information about the measure Mean($°C^{\text{Min}}$) and in particular the last 3 weekly changes. The respective weekly changes are plotted in the lower half of each figure. The x-Axis denotes the week of first usage in forecasting. For example, a forecast for week 7 is the first period which can use the temperature change listed at x-value 7. Together with the temperature changes 6 and 5, because the last 3 changes can be used.

How this works can be showcased for Berlin, figure 5.7. A decrease in the observed counts (middle gray) can be observed from week 5 to 6. This decrease is used in the prediction for

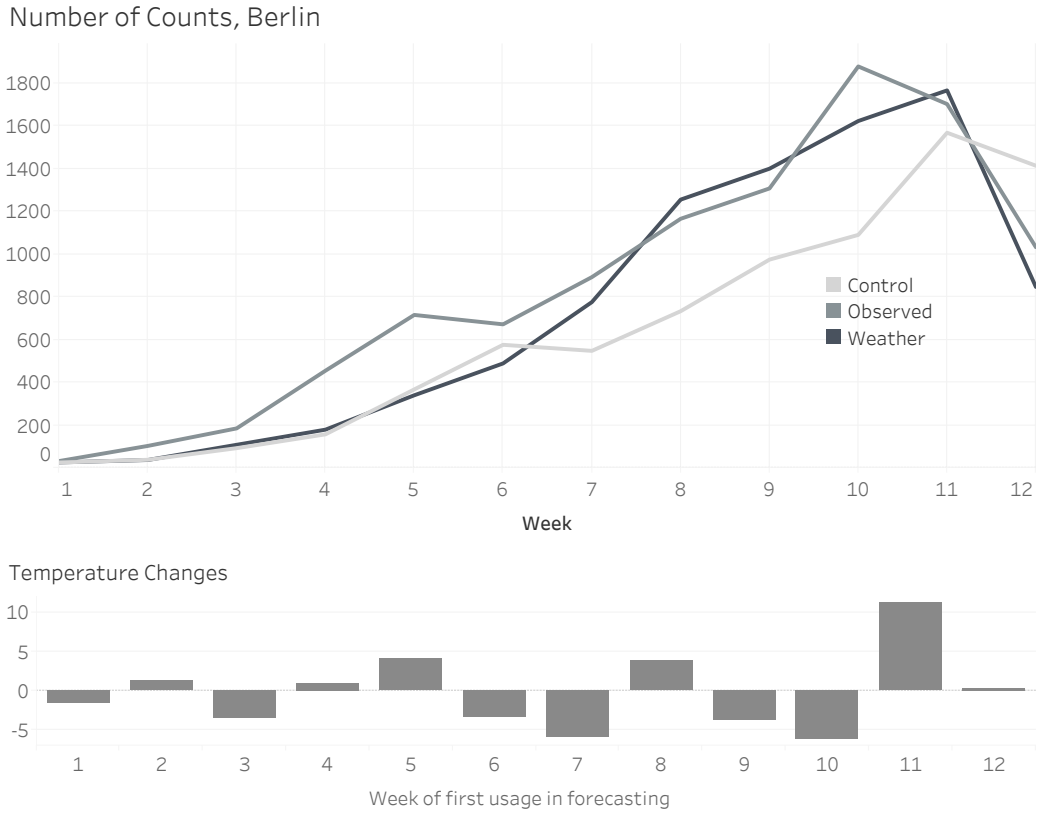


Figure 5.7: Observed weekly counts in Berlin 2018, together with weather and control model predictions [upper figure]. The change in Mean($°C^{\text{Min}}$) and its first usage for forecasting [lower figure].

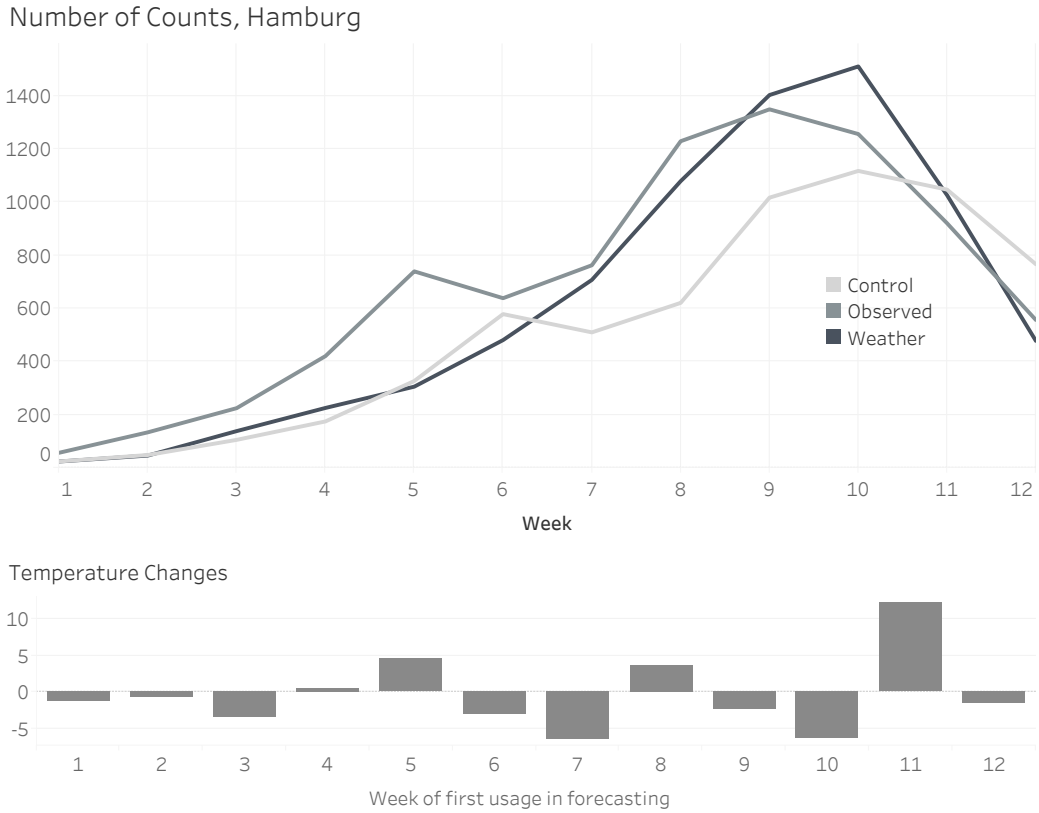


Figure 5.8: Observed weekly counts in Hamburg 2018, together with weather and control model predictions [upper figure]. The change in Mean($°C^{\text{Min}}$) and its first usage for forecasting [lower figure].

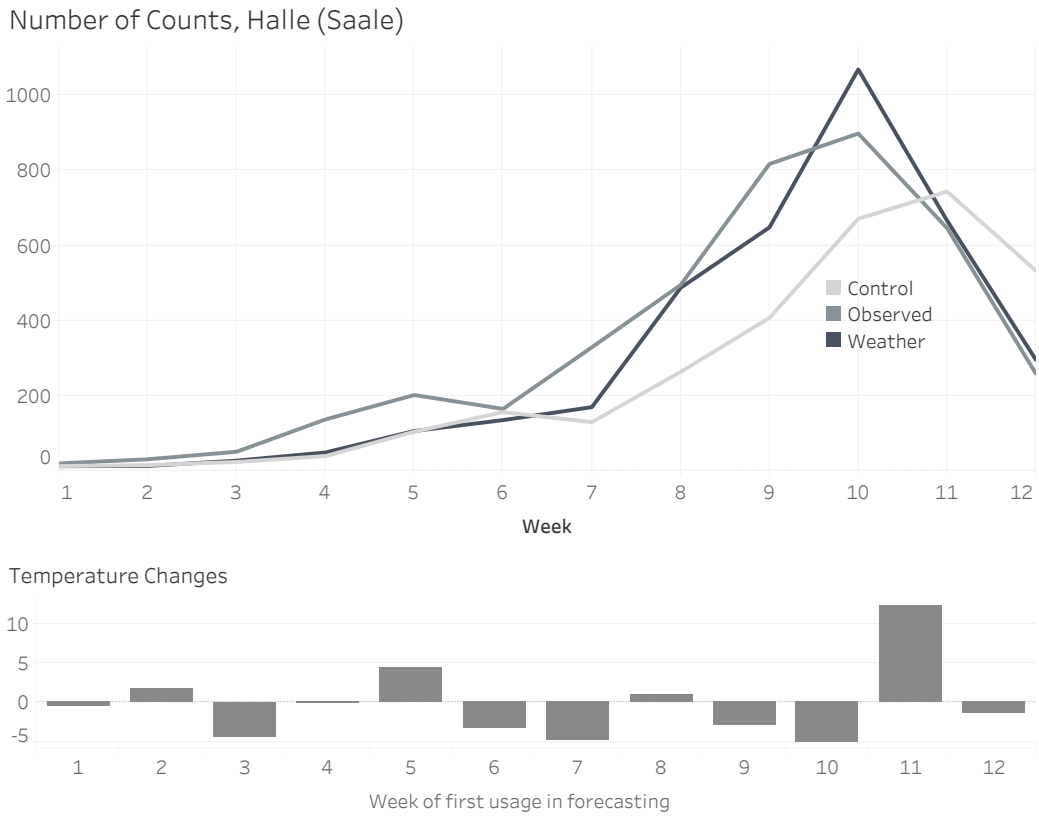


Figure 5.9: Observed weekly counts in Halle (Saale) 2018, together with weather and control model predictions [upper figure]. The change in Mean($°C^{\text{Min}}$) and its first usage for forecasting [lower figure].

week 7, resulting in a corresponding decrease in the control model. The weather model, however, observes a large drop in the temperature variable of over 5°C, which sharply increases its own forecast.

Another example is the forecast for week 11. The observed counts just showed a jump of almost 600 cases (to its peak). The prediction of the control model increases equally strong, but not the prediction of the weather model. That model just observed a quick rise in the temperature measure, which dampens the infection process. It is not enough to completely reverse the prediction, given the strong increase in observed cases and temperature decreases observed in the two weeks before, but it is enough to considerably change the prediction.

It is as if the weather model, at this point in time, is able to foresee the beginning of the subsiding phase of the influenza season. But how often do those situations appear? The weather model was considerably better only for 2018. Is the weather model suited for a particular sort of (extraordinary) situation that did not happen in any of the other evaluation years, or was it just pure luck?

6 Discussion

The goal of this study was to investigate the usefulness of meteorological data in surveillance systems for disease outbreaks. The ability of those systems hinges on the quality of forecasting the number of observed cases. Extraordinary events, like jumps and dips, can only be detected by comparing the observed case number to some kind of expected number - a forecast made earlier.

The data that typically arises from surveillance systems is the number of reported cases in a given region during a given time period. A statistical model able to handle such event count data with a space and time dimension, while simultaneously incorporating data from additional data sources, is the Endemic-Epidemic Model [15, 18, 32]. Its characteristic feature is the additive decomposition of the region- and time-specific mean into an endemic and epidemic component. The central distinction is the epidemic component's dependency on observed cases from the previous week - either from the same region (autoregressive) or from different regions (neighborhood). The endemic component comprises all other persistent effects that are independent of disease counts. Each of these components can be enriched with time- and region-varying covariates, e.g. weather and population data.

Estimations of this model were run on case data, provided by the Robert Koch Institute [35], for two diseases that exhibit a strong seasonality when plotted over time. Campylobacteriosis, a gastrointestinal infection caused by Campylobacter bacteria, and the viral infection influenza. Monitoring each disease is important: Campylobacter is the food-based disease with the highest incidence rate in Europe for every year since 2005. Influenza's ability to spread rapidly from person-to-person, together with its high mutation rate, makes the disease very dangerous - as observed e.g. during the 1918 influenza pandemic known as Spanish flu.

Additional employed data sources include a polygon map for Germany, used in the neighborhood component, from the Database of Global Administrative Areas (GADM), population data from the Federal Statistical Office as part of the census 2011 [40], and various types of weather data, measured and compiled by the Deutsche Wetterdienst

[9]. Measurements include relative humidity, hours of sunlight, amount of rain and temperature.

Several necessary preprocessing steps were needed to be carried out. The two most important were the deletion of data from calendar weeks 53, because the statistical model works on the assumption of 52 weeks per year, and the creation of a mapping between the weather measurements and counties. To achieve the latter, OpenStreetMap [28] was used to get addresses, including the zip code, from the longitude and latitude information provided for each weather station. The zip code was in turn transformed into a county name. Counties lacking a weather station got assigned the closest station as a replacement.

Various configurations of the endemic-epidemic model were run and evaluated using mean squared error (MSE) and mean absolute error (MAE) computed from rolling one-week-ahead forecasts for the 52 weeks of 2018. First tests revealed the helpfulness of a calendar week 2 indicator variable for Campylobacter and a preference for using only the last 2 years of data, instead of more prior years, for fitting the model.

The first set of experiments used the endemic component to incorporate the weather information. Considered was the inclusion of a single variable, a differenced variable, a polynomial, pairs of variables and three lags of a differenced variable. However, none of the specifications were able to reliably improve upon the control model, which excluded any weather information. In fact, it hardly mattered what weather measure or what specific formulation was used.

The epidemic component, in particular the autoregressive component, was used instead to incorporate weather information in the second set of experiments. Two formulations were tested and both proved useful. For Campylobacter, the basic single variable formulation decreased the MSE from 8.20 to 8.07. For influenza, 3 lags of the differenced variable hugely improved the MSE from 531.8 to 321.5, an improvement equivalent to 39.5%. The best weather measure for both specifications turned out to be Mean($°C^{\text{Min}}$), the weekly mean of the daily minimum temperature.

Letting the meteorological features enter the model simultaneously through the endemic and autoregressive component did not show much promise as the both-component formulation was incapable of improving upon the autoregressive-only formulation. In fact allowing both entry channels was sometimes even worse.

The best model for each disease was further evaluated on varying evaluation periods. It turned out that the model for Campylobacter was each year slightly better than a control

model. The influenza model was similarly evaluated. However, the huge 39.5% improvement for the 2018 evaluation period was a one-of-a-kind event. No major improvement was observable for any other evaluation period.

A thorough investigation revealed several facts about the influenza season in 2018. First, it was an exceptional season with 3 to 4 times more observed cases than in the years before. Second, the advantage of the model incorporating weather effects can be traced to a better forecasting performance for weeks 8, 10 and 12. For each of these three weeks, a handful of counties exist (Berlin, Hamburg and Halle) with a huge impact on the average mean squared error.

Plotting the case numbers observed in these three counties for the evaluation period 2018 and the two years used for fitting the model, 2016 and 2017, showed that there is a lot more variation over time than across counties. Not only does the number of infected people differ considerably each year, but the timing is quite different. This leaves the endemic component's seasonality very ill-equipped to forecast the correct number of cases. Seasonality is seen as something regularly occurring, with the same strength and timing. The observed shifts do not fit into that picture and the way seasonality is formulated in the model equation. The autoregressive component, on the other hand, has the necessary ignorance of calendar time to help forecasting too-early, too-strong or too-late increases or decreases.

A simultaneous examination of the observed case numbers and the weather changes in the three counties uncovered that large temperature changes coincided with large changes in the number of infected people. The autoregressive component was able to make good use of the weather information in that case. However, it remains unclear if the model would be as useful in a future situation as it was in 2018. It is very concerning that the model showed no considerable advantage over the control model in any other year.

For Campylobacter one can conclude that meteorological effects are certainly not the primary driver behind infections. A result in line with the literature, that produces very ambivalent results about the role of weather [19, 10], besides documenting a clear seasonality in the case data. However, the inconsistent results for influenza are a bit surprising given how well the role of temperature and humidity is understood [23].

Thus other factors might play a role in preventing the model to pick up the information present in the weather data. It could be that the variation of the weather data is too low. For example because too many weather stations were excluded during the filtering. Instead of entirely removing stations with gaps in its data, the available data could be used and measurements from close stations used to fill the gaps. Maybe the remaining

variation of the weather variables is too low after weekly averaging balanced out daily differences across regions.

Another aspect to consider is the employed model. Its functional form assumptions with regard to using the weather are very restricting. A zero-inflated model might be an alternative, given the large number of observations with a value of 0. There might be too many counties relative to weeks used in the model. Especially if the used specification does not really make use of the regional aspect besides the neighborhood component. Models that allow for more county-specific effects might be appropriate.

Due to time and computation restrictions, several of the available model specifications could not be tested. These include random and fixed effects, county-specific overdispersion, other formulations of neighborhood weights, seasonality in the epidemic component and the HHH4 add-on [4], that allows for conditioning on more past count values in the autoregressive component. However, it is unclear how this has changed the weather variable's ability to pick up its true effects: it might have been hindered or improved given that no other competing county-varying variables were present.

Lastly it must be said that even if more information were to be extracted from weather data by some other means, it would be useful only for a handful of diseases - most likely influenza. Many infectious diseases neither exhibit a seasonality nor is there anything known about a role for weather in the infection process.

Bibliography

[1] Roy M. Anderson, B. Anderson, and Robert M. May. *Infectious diseases of humans: dynamics and control*. Oxford university press, 1992.

[2] Cici Bauer and Jon Wakefield. "Stratified space-time infectious disease modelling, with an application to hand, foot and mouth disease in China". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67.5 (May 2018), pp. 1379–1398. DOI: 10.1111/rssc.12284.

[3] Roger S. Bivand, Edzer Pebesma, and Virgilio Gomez-Rubio. *Applied spatial data analysis with R, Second edition*. Springer, NY, 2013. URL: http://www.asdar-book.org/.

[4] Johannes Bracher and Leonhard Held. *Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction*. 2019. arXiv: http://arxiv.org/abs/1901.03090v2 [stat.AP].

[5] Bundesamt für Justiz. *Gesetz über den Deutschen Wetterdienst (DWD-Gesetz)*. URL: https://www.gesetze-im-internet.de/dwdg/index.html (visited on 11/07/2019).

[6] Bundesamt für Justiz. *Gesetz zur Verhütung und Bekämpfung von Infektionskrankheiten beim Menschen*. URL: https://www.gesetze-im-internet.de/ifsg/__7.html (visited on 11/07/2019).

[7] Bundesamt für Justiz. *Verordnung zur Festlegung der Nutzungsbestimmungen für die Bereitstellung von Geodaten des Bundes (GeoNutzV)*. URL: https://www.gesetze-im-internet.de/geonutzv/ (visited on 11/07/2019).

[8] C. P. A. Skarp, M.-L. Hänninen, and H. I. K. Rautelin. "Campylobacteriosis: the role of poultry meat". In: *Clinical Microbiology and Infection* 22.2 (Feb. 2016), pp. 103–109. DOI: 10.1016/j.cmi.2015.11.019.

[9] Deutscher Wetterdienst. *Climate Data Center*. URL: https://cdc.dwd.de/portal/ (visited on 10/29/2019).

[10] Abdelmajid Djennad et al. “Seasonality and the effects of weather on Campylobacter infections”. In: *BMC Infectious Diseases* 19.1 (Mar. 2019). DOI: 10.1186/s12879-019-3840-7.

[11] European Centre for Disease Prevention and Control. *Facts about campylobacteriosis*. URL: https://www.ecdc.europa.eu/en/campylobacteriosis/facts (visited on 11/19/2019).

[12] European Food Safety Authority and European Centre for Disease Prevention and Control. “The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2017”. In: *EFSA Journal* 16.12 (2018).

[13] Jeremy Ginsberg et al. “Detecting influenza epidemics using search engine query data”. In: *Nature* 457.7232 (Feb. 2009), pp. 1012–1014. DOI: 10.1038/nature07634.

[14] J. D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994. Chap. 2. ISBN: 9780691042893.

[15] Leonhard Held, Michael Höhle, and Mathias Hofmann. “A statistical framework for the analysis of multivariate infectious disease surveillance counts”. In: *Statistical Modelling: An International Journal* 5.3 (Oct. 2005), pp. 187–199. DOI: 10.1191/1471082x05st098oa.

[16] Leonhard Held, Sebastian Meyer, and Johannes Bracher. “Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture”. In: *Statistics in Medicine* 36.22 (June 2017), pp. 3443–3460. DOI: 10.1002/sim.7363.

[17] Leonhard Held and Michaela Paul. “Modeling seasonality in space-time infectious disease surveillance data”. In: *Biometrical Journal* 54.6 (Oct. 2012), pp. 824–843. DOI: 10.1002/bimj.201200037.

[18] Leonhard Held et al. “A two-component model for counts of infectious diseases”. In: *Biostatistics* 7.3 (Dec. 2005), pp. 422–437. DOI: 10.1093/biostatistics/kxj016.

[19] R. Sari Kovats et al. “Climate variability and campylobacter infection: an international study”. In: *International Journal of Biometeorology* 49.4 (Nov. 2004), pp. 207–214. DOI: 10.1007/s00484-004-0241-3.

[20] Landkreis Göttingen and Landkreis Osterode am Harz. *Gebietsänderungsvertrag*. 2013. URL: https://www.landkreisgoettingen.de/pics/medien/1_1384788061/Gebietsaenderungsvertrag_letzte_Fassung.pdf (visited on 11/07/2019).

[21] E. Lofgren et al. “Influenza Seasonality: Underlying Causes and Modeling Theories”. In: *Journal of Virology* 81.11 (Dec. 2006), pp. 5429–5436. DOI: 10.1128/jvi.01680-06.

[22] V. R. Louis et al. “Temperature-Driven Campylobacter Seasonality in England and Wales”. In: *Applied and Environmental Microbiology* 71.1 (Jan. 2005), pp. 85–92. DOI: 10.1128/aem.71.1.85-92.2005.

[23] Anice C. Lowen et al. “Influenza Virus Transmission Is Dependent on Relative Humidity and Temperature”. In: *PLoS Pathogens* 3.10 (2007), e151. DOI: 10.1371/journal.ppat.0030151.

[24] Sebastian Meyer and Leonhard Held. “Power-law models for infectious disease spread”. In: *The Annals of Applied Statistics* 8.3 (Sept. 2014), pp. 1612–1639. DOI: 10.1214/14-aoas743.

[25] Sebastian Meyer, Leonhard Held, and Michael Höhle. “Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance”. In: *Journal of Statistical Software* 77.11 (2017). DOI: 10.18637/jss.v077.i11.

[26] D. G. Newell and C. Fearnley. “Sources of Campylobacter Colonization in Broiler Chickens”. In: *Applied and Environmental Microbiology* 69.8 (Aug. 2003), pp. 4343–4351. DOI: 10.1128/aem.69.8.4343-4351.2003.

[27] G. Nylen et al. “The seasonal distribution of campylobacter infection in nine European countries and New Zealand.” In: *Epidemiology and infection* 128 (3 June 2002), pp. 383–390. ISSN: 0950-2688. DOI: 10.1017/s0950268802006830.

[28] OpenStreetMap contributors. *Nominatim*. https://nominatim.openstreetmap.org/. 2019.

[29] OpenStreetMap contributors. *Zip code to county mapping*. https://www.suche-postleitzahl.org/download_files/public/zuordnung_plz_ort_landkreis.csv. 2019. (Visited on 07/15/2019).

[30] M. E. Patrick et al. “Effects of Climate on Incidence of Campylobacter spp. in Humans and Prevalence in Broiler Flocks in Denmark”. In: *Applied and Environmental Microbiology* 70.12 (Dec. 2004), pp. 7474–7480. DOI: 10.1128/aem.70.12.7474-7480.2004.

[31] M. Paul and L. Held. “Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts”. In: *Statistics in Medicine* (2011), pp. 1118–1136. DOI: 10.1002/sim.4177.

[32] M. Paul, L. Held, and A. M. Toschke. “Multivariate modelling of infectious disease surveillance data”. In: *Statistics in Medicine* 27.29 (Dec. 2008), pp. 6250–6267. DOI: 10.1002/sim.3440.

[33] S. Paynter. “Humidity and respiratory virus transmission in tropical and temperate settings”. In: *Epidemiology and Infection* 143.6 (Oct. 2014), pp. 1110–1118. DOI: 10.1017/s0950268814002702.

[34] Edzer J. Pebesma and Roger S. Bivand. “Classes and methods for spatial data in R”. In: *R News* 5.2 (Nov. 2005), pp. 9–13. URL: https://CRAN.R-project.org/doc/Rnews/.

[35] Robert Koch-Institut. *SurvStat@RKI 2.0*. URL: https://survstat.rki.de (visited on 10/23/2019).

[36] S. J. Evans and A. R. Sayers. “A longitudinal study of campylobacter infection of broiler flocks in Great Britain”. In: *Preventive Veterinary Medicine* 46.3 (Aug. 2000), pp. 209–223. DOI: 10.1016/s0167-5877(00)00143-4.

[37] Jeffrey Shaman et al. “Absolute Humidity and the Seasonal Onset of Influenza in the Continental United States”. In: *PLoS Biology* 8.2 (Feb. 2010). Ed. by Neil M. Ferguson, e1000316. DOI: 10.1371/journal.pbio.1000316.

[38] Radina P. Soebiyanto, Farida Adimi, and Richard K. Kiang. “Modeling and Predicting Seasonal Influenza Transmission in Warm Regions Using Climatological Parameters”. In: *PLoS ONE* 5.3 (Mar. 2010). Ed. by Matthew Baylis, e9450. DOI: 10.1371/journal.pone.0009450.

[39] J. P. Southern, R. M. M. Smith, and S. R. Palmer. “Bird attack on milk bottles: possible mode of transmission of Campylobacter jejuni to man”. In: *The Lancet* 336.8728 (Dec. 1990), pp. 1425–1427. DOI: 10.1016/0140-6736(90)93114-5.

[40] Statistische Ämter des Bundes und der Länder. *Zensus 2011: Bevölkerung nach Geschlecht für Kreise und kreisfreie Städte*. Apr. 10, 2014. URL: https://www.zensus2011.de/SharedDocs/Downloads/DE/Pressemitteilung/DemografischeGrunddaten/1A_EinwohnerzahlGeschlecht.xls?__blob=publicationFile&v=5 (visited on 10/18/2019).

[41] Tyler H. Koep et al. “Predictors of indoor absolute humidity and estimated effects on influenza virus survival in grade schools”. In: *BMC Infectious Diseases* 13.1 (Feb. 2013). DOI: 10.1186/1471-2334-13-71.

[42] J. S. Wallace et al. “Seasonality of thermophilic Campylobacter populations in chickens”. In: *Journal of Applied Microbiology* 82.2 (Feb. 1997), pp. 219–224. DOI: 10.1111/j.1365-2672.1997.tb02854.x.

[43] World Health Organization. *Campylobacter Fact Sheet*. URL: https://www.who.int/en/news-room/fact-sheets/detail/campylobacter (visited on 11/19/2019).

[44] World Health Organization. *Fact Sheet on Influenza (Seasonal)*. URL: https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal) (visited on 11/19/2019).

[45] World Health Organization. *Fact sheets: Infectious diseases*. URL: https://www.who.int/topics/infectious_diseases/factsheets/en/.

[46] World Health Organization. *Influenza Vaccines*. URL: https://www.who.int/biologicals/vaccines/influenza/en/.