Held, Höhle, Hofmann:

A statistical framework for the analysis of multivariate infectious disease surveillance data

Sonderforschungsbereich 386, Paper 402 (2004)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

A statistical framework for the analysis of multivariate infectious disease surveillance data

Leonhard Held, Michael Höhle and Mathias Hofmann
Department of Statistics
University of Munich
Ludwigstr. 33, 80539 Munich
Germany

Email: {leonhard.held, hoehle, mhofmann}@stat.uni-muenchen.de

30th November 2004

Abstract

A framework for the statistical analysis of counts from infectious disease surveillance databases is proposed. In its simplest form, the model can be seen as a Poisson branching process model with immigration. Extensions to include seasonal effects, time trends and overdispersion are outlined. The model is shown to provide an adequate fit and reliable one-step-ahead prediction intervals for a typical infectious disease surveillance time series. Furthermore, a multivariate formulation is proposed, which is well suited to capture space-time interactions caused by the spatial spread of a disease over time. Analyses of uni- and multivariate times series on several infectious diseases are described. All analyses have been done using general optimization routines where ML estimates and corresponding standard errors are readily available.

Key words: Branching Process with Immigration; Infectious Disease Surveillance; Maximum Likelihood; Multivariate Time Series of Counts; Observation-driven; Parameter-driven; Space-Time-Models

1 Introduction

There has been much recent interest in the statistical analysis of multivariate time series of counts, where each component, for example, corresponds to the number of disease cases in a specific geographical region or in a certain age group. Such data arise naturally in surveillance systems on infectious diseases and are typically collected on a weekly or daily basis. Statistical analyses are typically done with computer-intensive Markov chain Monte Carlo (MCMC) methods, see for example Mugglin *et al.* (2002), Svensson and Lindbäck (2002) and Knorr-Held and Richardson (2003).

For simplicity, consider first the simple univariate time-series case. Approaches to analyze such data typically employ a log-linear Poisson regression model, perhaps allowing for overdispersion, and model the disease incidence with unknown latent parameters, which exhibit temporal (Farrington *et al.*, 1996) and possibly also seasonal dependence. For example, the number of counts y_t at time $t = 1, \ldots, n$ may be assumed to be Poisson with mean $\exp(\eta_t)$ where

$$\eta_t = \alpha_0 + \alpha_1 t + \sum_{s=1}^{S} \left(\beta_s \sin(\omega_s t) + \gamma_s \cos(\omega_s t)\right), \tag{1}$$

where the Fourier frequencies ω_s are $\omega_s = 2s\pi/52$ for weekly data (Diggle, 1990). Following the terminology of Cox (1981), this class of models can be called *parameter-driven*. Note that similar parameter-driven formulations with suitable prior distributions on latent parameters are used in the analysis of non-infectious diseases, for example counts of cancer incidence or mortality (Knorr-Held and Besag, 1998, Knorr-Held, 2000).

However, it has soon been recognized that a purely *parameter-driven* approach is often not able to describe localized epidemics, and further model extensions were needed. In particular, a fruitful approach is to add the number of cases in the past as additional explanatory variables in the model. In the terminology of Cox (1981), this part of the model is called *observation-driven* and, combined with (1), the complete model could thus be called *parameter- and observation-driven*.

However, certain complications arise. Adding the observed counts y_{t-1} in the linear predictor (1), i.e. y_t is Poisson distributed with mean

$$\mu_t = \exp(\eta_t + \lambda y_{t-1}),$$

say, is implausible because this model can only describe negative association but no positive association without growing exponentially in time (Diggle *et al.*, 2002, Section 10.4). Zeger and Quaqish (1988) have therefore introduced a modification, where essentially the logarithm of the observed counts (with $\log 0$ replaced by $\log d$, $0 < d < 1$), minus the linear predictor η_{t-1} of y_{t-1}, enters as an explanatory variable, i.e.

$$\mu_t = \exp(\eta_t + \lambda(\max(\log y_{t-1}, \log d) - \eta_{t-1})).$$

This model can be seen as a size-dependent branching process (Diggle *et al.*, 2002) and has nicer properties, in particular it allows for positive association between successive counts. The constant d prevents $y_{t-1} = 0$ from being an absorbing state, forcing all future responses to be zero. When $\lambda > 0$, we have an increased expectation when the previous outcome exceeds $\exp(\eta_{t-1})$.

However, the introduction of the parameter d and the regularization of past counts y_{t-1} through their non-epidemic expected values η_{t-1} is complicated and seems slightly unnatural. Interpretation of the autoregressive parameter λ is not straightforward in this formulation. Alternatively, Knorr-Held and Richardson (2003) let the logarithm of $1+y_{t-1}$ enter as an explanatory variable, but avoid the need to regularize. This is achieved by modulating the dependence on the previous counts by latent 0-1-indicators, which are assumed to follow a two-stage hidden Markov model.

In this paper we take a different model perspective, motivated from a branching process model with immigration (e.g. Guttorp, 1995, Section 2.11). Essentially, our proposal is to let previous counts act directly on the conditional mean μ_t of $y_t|y_{t-1}$ (and *not* on the log mean), so - in its simplest version without temporal or seasonal trends - we use an identity link rather than a log link:

$$\mu_t = \nu + \lambda y_{t-1}. \tag{2}$$

It is easy to show (Guttorp, 1995) that, for $\nu > 0$ and $0 < \lambda < 1$, this process is stationary with mean and variance

$$\begin{aligned} \mu &= \nu/(1-\lambda), \\ \sigma^2 &= \nu/\{(1-\lambda)(1-\lambda^2)\}. \end{aligned} \tag{3}$$

The advantage of model (2) is that, without the immigration, it has a nice interpretation as an approximation to a chain binomial model (see Becker, 1989, for further details) in the absence of information on the number of disease susceptibles. Information on the number of susceptibles is only ever available in very special, much analysed dataset, see for example Finkenstädt *et al.* (2002). It is seldom, if ever, available in a surveillance setting (Farrington *et al.,* 2003), so the approximation appears to be justified. Furthermore, under certain assumptions, the autoregressive parameter λ can be interpreted as the basic reproduction number R_0 (Farrington *et al.,* 2003), the key quantity in infectious disease epidemiology (e.g. Dietz, 1996).

The additional influx of immigrated case with mean ν ensures that the process will not die out with probability one, in contrast to the ordinary branching process. This is a useful addition, as infectious disease surveillance data often displays a mixture of an endemic and an epidemic behaviour. Indeed, for λ close to one, simulations from this model display occasional epidemic outbreaks, so the formulation seems more realistic than a purely *parameter-driven* formulation. However, in applications there may be need to replace the Poisson with a more flexible observation model to allow for overdispersion. We will use a negative binomial model, where the mean structure of the models remains the same but the variance σ_t^2 increases to

$$\sigma_t^2 = \mu_t + \mu_t^2/\psi$$

with the additional parameter $\psi > 0$, to be estimated from the data at hand. Note that for $\psi \to \infty$, the negative binomial model equals the simple Poisson model.

Clearly, model (2) will still be not sufficient for most data on infectious diseases. In particular, it does not allow for seasonality and temporal trends. A simple adjustment is to replace ν with a time-changing ν_t, where $\log \nu_t = \eta_t$ from (1). Note that the autoregressive parameter λ remains independent of time. However, we comment on extensions with time- or area-dependent λ in Section 4. As a side comment, unequally spaced time series can easily be casted into this framework by including the log length of the underlying reporting period as an additional offset variable.

A further advantage of this formulation is that the extended model is easily estimated by Maximum Likelihood using generic optimization routines, e.g. the function `optim()` in `R`. Note that the simple model (2) is just a generalized linear model (GLM) with Poisson observation model

and identity link, whereas the extended model with time-changing ν_t as in (1) does no longer fit into the GLM framework.

We will show that the model can readily be extended to typical research questions that appear in infectious disease surveillance. We will discuss the following: (a) Model-based prediction of epidemics, (b) Modelling multivariate times series, and (c) Space-time modelling of longitudinal count data on infectious diseases. Further extensions are outlined in Section 4.

2 The time series case

For illustration consider Figure 1, a univariate time series of weekly counts for *Salmonella Agona*, 1990-95, reported also in Farrington *et al.* (1996). Note that the time series shown in Farrington *et al.* (1996) is slightly different (and also slightly shorter), due to later modifications in the data-file.

To these data we fitted model (2) with $\log \nu$ replaced by $\log \nu_t$ from (1) and $S = 1$. Higher terms for seasonality did not lead to a significant improvement in the likelihood. The term for the linear time trend has always been included. We have thus fitted 8 different models, depending on the observation model (Poisson or negative binomial), whether the seasonality terms have been included or not, and whether the autoregressive component λy_{t-1} has been included in the linear predictor (2) or not. The results are summarized in Table 1.

Model	Distribution	Seasonality	Autoregression	$\hat{\lambda}(SE)$	$\hat{\psi}$ (SE)	$\log L$	DF
1	Poisson	No	No	–	–	−744.0	309
2	Poisson	No	Yes	0.49 (0.03)	–	−664.8	308
3	Poisson	Yes	No	–	–	−660.0	307
4	Poisson	Yes	Yes	0.29 (0.03)	–	−637.8	306
5	Neg. Binomial	No	No	–	2.1 (0.3)	−673.4	308
6	Neg. Binomial	No	Yes	0.48 (0.04)	3.8 (0.8)	−636.5	307
7	Neg. Binomial	Yes	No	–	4.0 (0.8)	−632.3	306
8	Neg. Binomial	Yes	Yes	0.27 (0.04)	5.2 (1.3)	−620.2	305

Table 1: Summary of the ML estimates (standard errors in brackets) of the different models for the *Salmonella Agona* data.

There are several interesting features to see:

- There is clear evidence for overdispersion, since the negative binomial models result in a significant increase in terms of maximized log-likelihood, denoted by $\log L$, compared to the corresponding Poisson models.

- Inclusion of seasonality terms in models with the autoregressive component leads to a considerably smaller estimated autoregressive parameter λ. This illustrates that the autoregressive component captures the *residual* temporal dependence in the time series, after adjusting for seasonal effects.

- Nevertheless, from the log-likelihood values it can be seen that both seasonality and the autoregressive component have to be included in the model and hence, model 8 appears to be the best.

The fit of the final model is compared with the observed data in Figure 1. Furthermore, Figure 2 also gives a plot of the Pearson residuals

$$r_t = \frac{y_t - \hat{\mu}_t}{\sqrt{\hat{\sigma}_t^2}}$$

where $\hat{\mu}_t$ and $\hat{\sigma}_t^2$ is the estimated mean and variance of y_t respectively. Note that, for the negative binomial model, we used

$$\hat{\sigma}_t^2 = \hat{\mu}_t + \hat{\mu}_t^2/\hat{\psi}.$$

Figure 2 also display a nonparametric smooth fit through the residuals and the estimated autocorrelation function of the residuals; The nonparametric fit does not indicate any strong evidence for non-stationarity of the residuals; the 95% pointwise confidence intervals for the autocorrelation give also only weak evidence of positive autocorrelation at lag 3, 7, 9, which may be spurious. In summary it appears that this model gives a reasonable (but perhaps not perfect) fit to the data.

As a further check, we also looked at the *out-of-sample* predictive quality of the models. The appropriateness of the predictions is of paramount importance, if warnings for the possibility of future outbreaks are based on the predictive distribution for y_{t+1}. Based on the data up to a certain week t, we have estimated all parameters in the model and then computed the predicted number

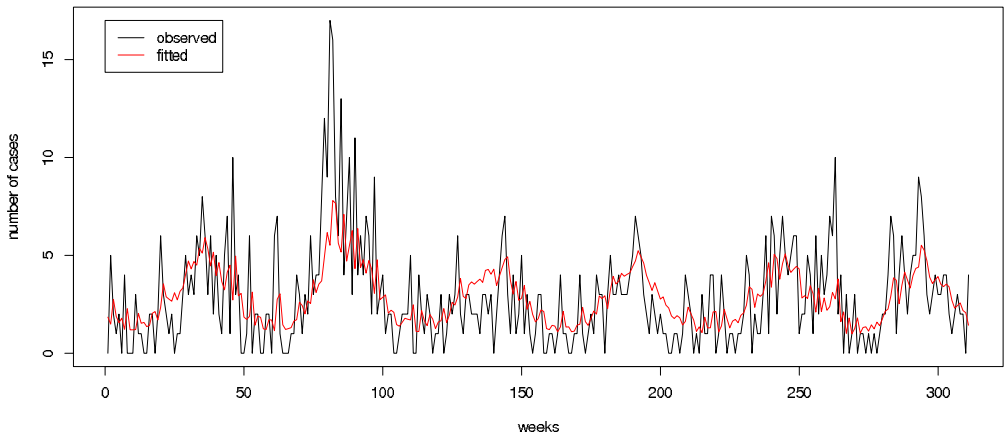

Figure 1: Observed and fitted counts of *Salmonella agona, 1990-95*

of cases from the chosen model for the next week. Furthermore, an upper limit for the number of cases has been computed based on the quantiles of the Poisson and negative binomial distribution, respectively. This procedure has been iterated over the last 100 observations of the time series and over all models.

The predictive quality of the different models, based on those one-step-ahead-predictions of the last 100 observations is summarized in Table 2. We have computed (a) the mean squared prediction error (MSPE) based on the square root counts (we have used the square root transformation in order to stabilize the variance of the counts) and (b) the empirical coverage of the upper prediction limits to the confidence levels 90, 95 and 99%, i.e. the proportion of observed counts that are smaller or equal to the corresponding upper prediction limit.

The following results should be highlighted:

- In terms of MSPE, the two models with seasonality and with the autoregressive component perform best. In particular, the MSPE is slightly smaller than the purely parameter-driven formulation with $\lambda = 0$.

- The MSPE is nearly independent of the chosen observation model. This is not surprising, as the negative binomial model does not change the mean structure, it only changes the variance structure.

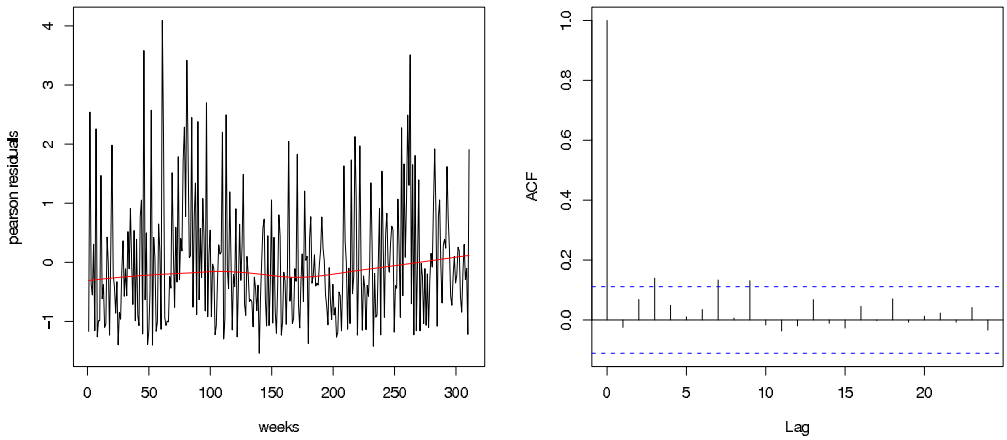

Figure 2: Pearson residuals and corresponding autocorrelation function

Model	Distribution	Seasonality	Autoregression	MSPE	Coverage		
					90%	95%	99%
1	Poisson	No	No	0.637	0.80	0.86	0.95
2	Poisson	No	Yes	0.558	0.84	0.92	0.98
3	Poisson	Yes	No	0.505	0.81	0.90	0.97
4	Poisson	Yes	Yes	0.484	0.83	0.89	0.97
5	Neg. Binomial	No	No	0.635	0.90	0.96	1.00
6	Neg. Binomial	No	Yes	0.557	0.93	0.98	0.99
7	Neg. Binomial	Yes	No	0.507	0.88	0.93	0.98
8	Neg. Binomial	Yes	Yes	0.484	0.87	0.94	0.99

Table 2: Predictive performance of the different models.

- The coverage is too low for the Poisson models (again not surprising, because the variance of the Poisson model is too low), but seems very reasonable for the negative binomial models.

Figure 3 illustrates this procedure for the last model. Shown are the one-step-ahead predictions, the corresponding 99% upper limits of the predictive distribution, and the actually observed data. Such an upper prediction limit is typically used as a threshold to flag outbreaks of infectious diseases (Farrington and Andrews, 2003).

In conclusion, we have shown that for this time series, a fairly simple model with only 6 unknown parameters produces an adequate fit and reasonable one-step-ahead predictions that do not indicate any serious inappropriateness of the model. All models can be estimated easily using the general

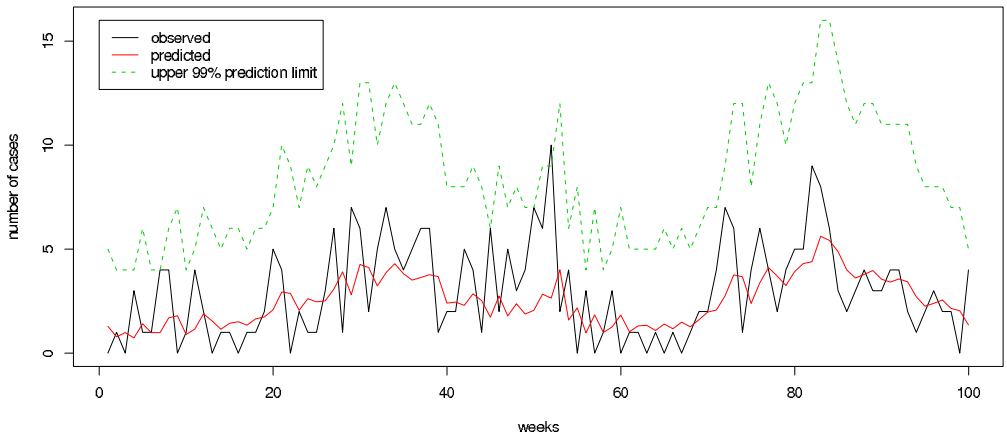

Figure 3: Observed, predicted, and upper 99% prediction limits for one-step-ahead predictions of the last 100 observations of the *Salmonella Agona* time series

purpose `R` routine `optim()` for numerical optimization. Results are immediately available and are numerically stable. Note that this function also returns standard errors based on the Hessian matrix, without any need to supply analytic derivatives of the log likelihood function (Venables and Ripley, 2002, Chapter 16).

As a further illustration, we have analysed the weekly number of *Enterohaemorrhagic Escherichia coli* (EHEC) infections in Bavaria, 2001-2003, shown in Figure 4. Fitting the same models as for *Salmonella Agona* to this time series, gives different results, summarized in Table 3. In particular, once the models are adjusted for seasonality, there is no need to include the autoregressive component and a purely *parameter-driven* model is sufficient. This can be seen as for the models with an additional autoregressive component, the estimate $\hat{\lambda}$ is not significantly different from zero and does also not lead to a significant reduction in maximized likelihood. Incidentally, there is still evidence for overdispersion, for example comparing the seventh with the third model using a likelihood ratio test gives a test statistic of $2(377.1 - 373.2) = 7.8$ with a p-value of 0.005. Note that Figure 4 also includes the fitted values from model 7 in Table 3. In contrast to Figure 1, now the sinusoidal parametric form for the seasonal pattern is clearly visible, due to the lack of the autoregressive component.

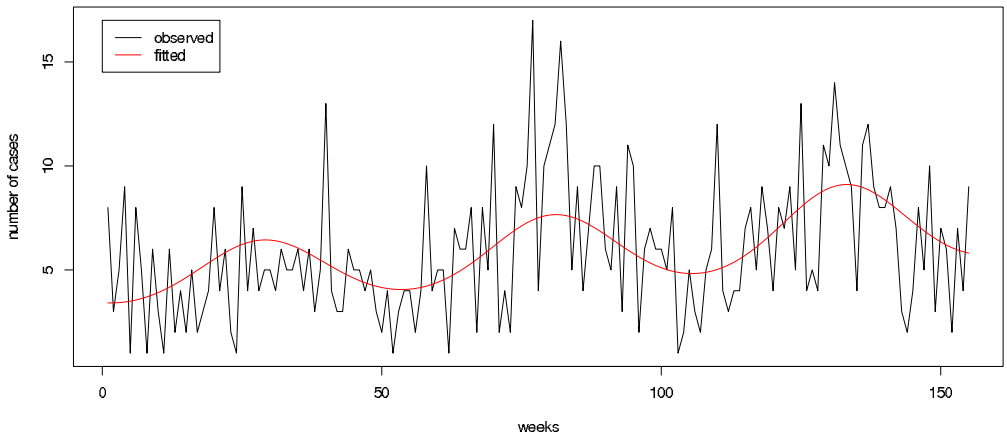

Figure 4: Observed and fitted counts of *Enterohaemorrhagic Escherichia coli*, 2001-2003

3 A multivariate model extension

Consider now the case where multivariate time series data is available. For example, we might consider the number of cases in different age groups or different geographical regions. We assume that we have $i = 1, \ldots, m$ "units" and denote with y_{it} the number of cases in unit i at time t.

Suppose now that (in the simplest model without time or seasonal trends) the mean structure is

$$\mu_{it} = \lambda y_{i,t-1} + n_{it}\nu$$

where n_{it} are - possibly standardized - population counts in area i.

This model has a nice aggregation property, since the aggregated counts $y_t = \sum_{i=1}^{n} y_{it}$ have mean

$$\mu_t = \lambda y_{t-1} + n_t \nu \tag{4}$$

where $n_t = \sum_{i=1}^{n} n_{it}$. So the parameter λ has the same interpretation for the aggregated counts as for the individual counts y_{it} and ν is adjusted with the corresponding population counts n_t. Note that - under a Poisson model for y_{it} - y_t will still be Poisson distributed. However, this does no longer hold for the negative binomial distribution.

Also note, that, in the Poisson case, the model can be written as a *multivariate* or *multitype*

Model	Distribution	Seasonality	Autoregression	$\hat{\lambda}(SE)$	$\hat{\psi}$ (SE)	$\log L$	DF
1	Poisson	No	No	–	–	-394.5	153
2	Poisson	No	Yes	0.13 (0.03)	–	-392.4	152
3	Poisson	Yes	No	–	–	-377.1	151
4	Poisson	Yes	Yes	-0.04 (0.03)	–	-376.9	150
5	Neg. Binomial	No	No	–	9.9 (3.0)	-384.6	152
6	Neg. Binomial	No	Yes	0.12 (0.04)	10.7 (3.4)	-383.4	151
7	Neg. Binomial	Yes	No	–	17.1 (7.7)	-373.2	150
8	Neg. Binomial	Yes	Yes	-0.05 (0.03)	17.1 (7.7)	-373.1	149

Table 3: Summary of the ML estimates (standard errors in brackets) of the different models for EHEC.

branching process with immigration (Mode, 1971) where

$$\boldsymbol{\mu}_t = \boldsymbol{\Lambda}\boldsymbol{y}_{t-1} + \boldsymbol{\nu}$$

with suitable defined vectors $\boldsymbol{\mu}_t$, $\boldsymbol{y}_{t-1}$, $\boldsymbol{\nu}$, and matrix $\boldsymbol{\Lambda}$. In model (4), $\boldsymbol{\Lambda}$ is simply diagonal with entries equal to λ; more elaborate specifications will be presented later.

As before, the assumption of a constant ν in (4) is too strict, and we may, for example, replace ν by ν_{it}, where

$$\log \nu_{it} = \alpha_i + \alpha_1 t + \sum_{s=1}^{S} (\beta_s \sin(\omega_s t) + \gamma_s \cos(\omega_s t)) \,. \tag{5}$$

Compared to (1), the additional unit-dependent parameter α_i allow for different incidence levels in the different units. For example, if a series of geographical units are analysed, there may be different reporting rates and they will be captured by α_i. Note that model (5) decomposes the incidence into unit-specific and time-dependent parameters, in the spirit of Knorr-Held and Besag (1998) and Knorr-Held and Richardson (2003). Inclusion of dependence *across* the different series in the *parameter-driven* part of the model is more difficult. Below we propose an extension where such dependence is captured in the *observation-driven* part of the model through an additional regression on the number of cases in other units. Depending on the context, this may be geographically *neighbouring* units, see Section 3.2, or, in the case of units corresponding to age groups, simply all other age groups.

3.1 Application to meningococcal infections in France

For illustration we now consider monthly counts of meningococcal incidence in France, 1985-1995. These data have previously been analysed in Knorr-Held and Richardson (2003) with focus on geographical variations. Here we split the data into $m = 4$ age groups (< 1, $1-5$, $5-20$, > 20) and obtain a multivariate time series of dimension four.

Model (5) has been fitted to these data with $S = 1$ (note that $\omega_s = 2\pi/12$ now), and the results are as follows: The ML-estimate of λ is 0.12 (0.02), indicating some evidence for a weak dependence on the number of counts in the last month - after adjustments for seasonal effects. The value of the maximized log likelihood is -1543.1, which should be compared to -1547.4, the value of the log likelihood in the purely *parameter-driven* model without the component $\lambda y_{i,t-1}$. Hence, there is evidence that the autoregressive component is needed (p-value $= 0.003$) in the model. Note that these results are based on a negative-binomial model, as there was evidence for residual overdispersion. The observed and fitted times series are displayed in Figure 5 while the Pearson residuals are shown in Figure 6.

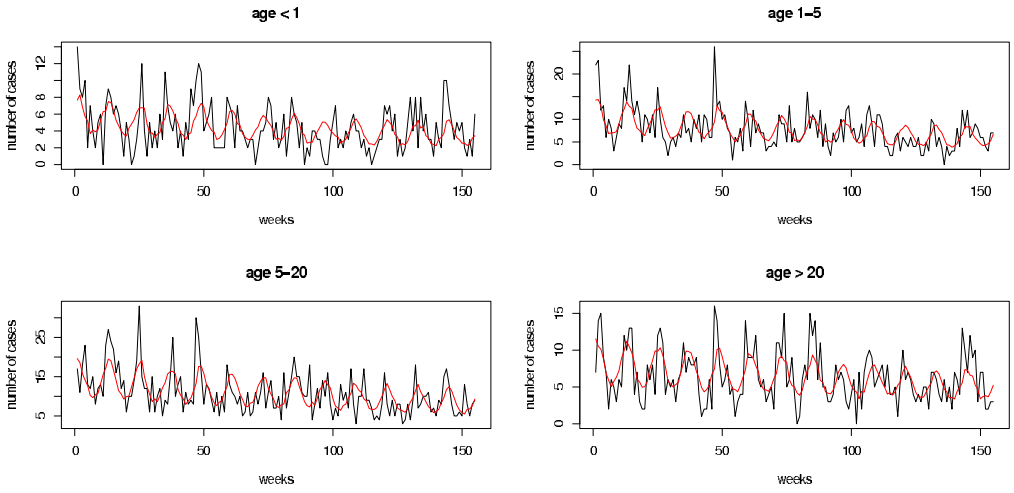

Figure 5: Observed and fitted counts of meningococcal infections in the different age groups

As said earlier, a more general model may also consider the past number of cases in other age

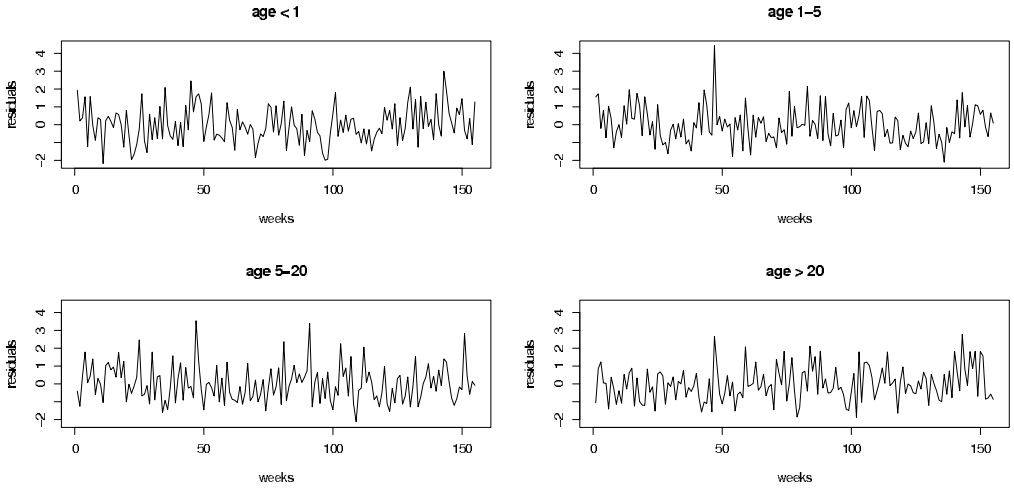

Figure 6: Pearson residuals

groups as potential explanatory variables for y_{it}. In the simplest case,

$$\mu_{it} = \lambda y_{i,t-1} + \phi \sum_{j \neq i} y_{j,t-1} + n_{it}\nu_{it} \tag{6}$$

with ν_{it} as in (5), which introduces one additional parameter ϕ for the autoregressive effect of the other age groups. Written as a multivariate branching process, the matrix $\mathbf{\Lambda}$ now has diagonal entries λ and off-diagonal entries equal to ϕ. Note that the effective basic reproduction number R_0 now equals the largest eigenvalue of $\mathbf{\Lambda}$ (Anderson and Britton, 2000, Chapter 6). If $R_0 < 1$, the process is ergodic with mean $\boldsymbol{\nu}(\boldsymbol{I} - \mathbf{\Lambda})^{-1}$ (Mode, 1971, Section 2.7). This formula is just the multivariate analogue of equation (3).

For the meningitis data, the ML-estimates of λ are nearly identical to the model without ϕ while the ML-estimate of ϕ is -0.0004 (0.005), very close and not significantly different from zero. Furthermore, the maximized log likelihood is still -1543.1, which clearly indicates, that the component $\phi \sum_{j \neq i} y_{j,t-1}$ is not needed for these data. Incidentally, we also considered the model including this term, but excluding $\lambda y_{i,t-1}$. The ML-estimate of ϕ is now 0.017 (0.005) and the maximized log likelihood is -1546.8, which is not significantly different from the purely *parameter-driven* model (p-value $= 0.27$). This indicates that, after adjusting for seasonality, an epidemic

component can be isolated *within* the age groups, but not *between* the age groups. This may relate to different contact rates in different age groups, see Farrington *et al.* (2001) and Whitaker and Farrington (2004) for a discussion of suitable choices of contact rates in different age groups in the context of serological surveys.

Incidentally, we have also analysed the four age groups separately, using the basic formulation (2) and (1). The ML-estimates of λ are 0.25 (0.04), 0.10 (0.03), 0.07 (0.03) and 0.05 (0.03) for the age groups < 1, $1-5$, $5-20$ and > 20. This suggests that there is some heterogeneity in the autoregressive effect, decreasing for older age groups. However, again using a likelihood ratio test, the separate analysis does not indicate a significant improvement over the joint model.

3.2 A space-time application: Measles epidemics in Lower Saxony

In the administrative district "Weser-Ems", located in the eastern part of the German state Lower Saxony, two measles epidemics occurred in the years 2001 and 2002. Here we analyse the weekly counts of those measles cases from the corresponding $m = 15$ spatial areas of this district, see Figure 8 for a map of the area considered. Note that we have omitted two areas with zero counts to avoid problems with non-existing ML-estimates. The data are shown in Figure 7, but see also

`http://www.nlga.niedersachsen.de/infekt/infekt_dat.htm`

for an animated movie of the 2002 epidemic (Click on "Interaktiver Infektionsbericht 2003" and then select "Masern" and "Diagramme, Zeitverlauf").

To these data we fitted a model adopted from (6),

$$\mu_{it} = \lambda y_{i,t-1} + \phi \sum_{j \sim i} y_{j,t-1} + n_{it}\nu_{it},$$

with population fractions n_{it} and ν_{it} as in (5), but where now the sum of the cases in other areas is only over *spatially adjacent* areas $j \sim i$. Two areas have been defined to be adjacent if they share a common border. With $S = 1$ Fourier frequencies, the model thus has 21 parameters. Again, estimation of this model with the function `optim()` was fairly straightforward and results have been computed in just a few seconds.

The estimates in the full model are $\hat{\lambda} = 0.62$ (0.08), $\hat{\phi} = 0.016$ (0.003) and $\hat{\psi} = 0.49$ (0.06) with $\log L = -942.6$. Thus, the space-time interaction effect of the counts in neighboring areas is

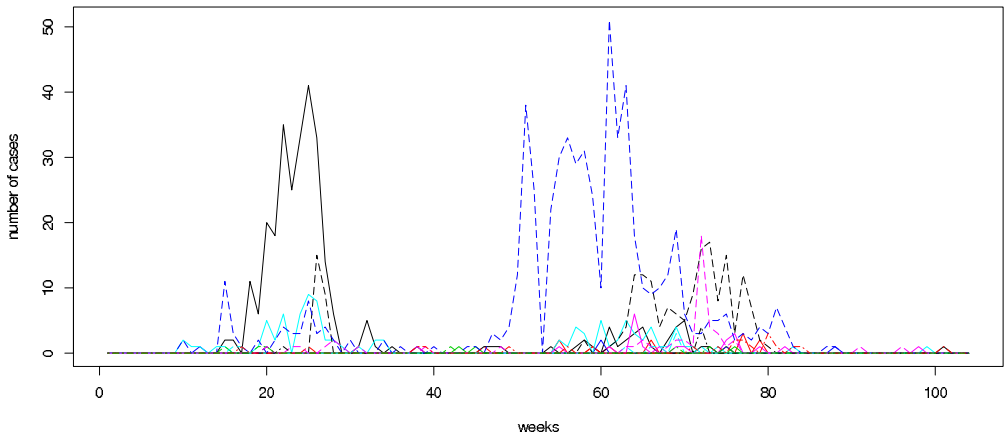

Figure 7: Weekly counts of new measles cases in $n = 15$ areas in the district "Weser-Ems"

significant. This can also be seen from a pronounced decrease in $\log L$ of the model with $\phi = 0$, where $\hat{\lambda} = 0.66$ (0.08), $\hat{\psi} = 0.45$ (0.06) and $\log L = -954.3$. This is in good agreement with the spatial spread of the disease over time, which is already visible from the animated movie mentioned above. We note that in the full model, the corresponding value of R_0 can be calculated from the eigenvalues of the matrix $\mathbf{\Lambda}$ and turns out to be $\hat{R}_0 = 0.69$.

4 Discussion

The attractiveness of the proposed framework is immediate: All models and analyses shown in this paper can be easily repeated within seconds using standard optimization software. Thus, in contrast to methods based on MCMC, the model is particularly well suited for routine analysis in infectious disease surveillance. On the other hand, we believe that the model constitutes a useful extension of the purely parameter-driven GLM formulation by Farrington *et al.* (1996).

However, some caveats are appropriate. First, the interpretation of the branching process as an approximation to a chain-binomial model is only appropriate if the generation time equals the observation time, typically days, weeks on months. However, we have conducted simulation studies which showed that a Poisson branching process, aggregated to coarser time intervals, can be

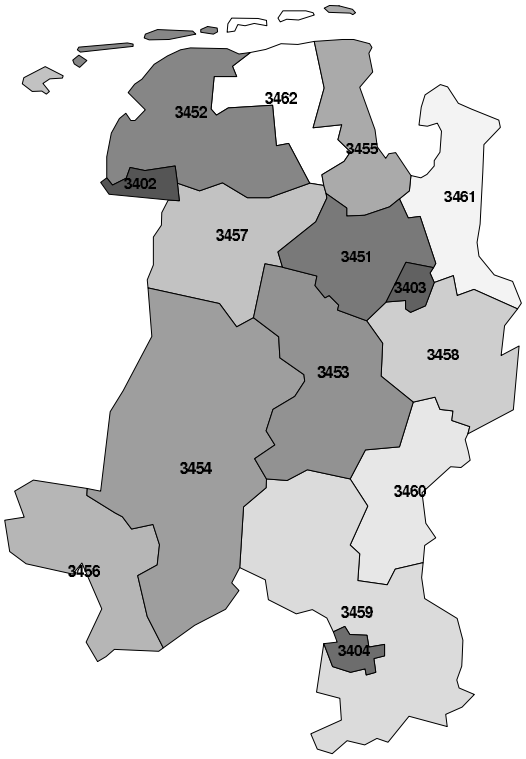

Figure 8: A map of the $m = 15$ areas in the administrative district "Weser-Ems"

approximated by a branching process with additional overdispersion. Indeed, in all our analysis, the switch from Poisson to negative binomial was needed. Another practical limitation of the model is that it does not allow for under-reporting, a typical feature of surveillance data. However, detailed information on under-reporting is rarely available. Furthermore, as long as the under-reporting rate is roughly constant across units or areas, it can well be absorbed by the area-specific parameter α_i. For example, in the analysis described in Section 3.2, the area effects (adjusted for population counts) showed a considerable variation, which may be both due to differences in incidence as well as different reporting rates.

Of course, further generalizations may require a Bayesian approach and more advanced MCMC techniques for statistical inference. For example, we are currently working on an extension where the parameter λ is allowed to vary over time, according to a Bayesian change-point model with unknown number of change points (Denison *et al.,* 2002). A time-changing λ_t will be appropriate in situations where the invectiveness of an agent varies over time, for example due to vaccination

programs, other interventions, or due to a sudden outbreak, where $\lambda_t > 1$ for some limited time period, to be estimated from the data. Alternatively, one may assume a smooth latent process for λ_t, perhaps suitably transformed. Similarly, random effects, possibly correlated, may be introduced at area-level. In both cases, Gaussian Markov random fields (Rue and Held, 2005) will be useful as prior distributions.

Another extension we currently consider in the space-time context is to include covariate information on area level. Such covariates could be introduced in ν, but also in λ, perhaps suitably transformed. The aim is here to bring together *spatial ecological regression* (Clayton *et al.*, 1993) and infectious disease epidemiology.

Acknowledgments

This work is supported by the German Science Foundation (DFG), SFB 386, Projekt B9: "Statistical methodology for infectious disease surveillance". We thank Paddy Farrington, Open University, UK, Klaus Stark and Christina Frank at the Robert-Koch Institute (RKI), Berlin, Germany, Daniel Lévy-Bruhl, Institut National de la Veille Sanitaire, Saint Maurice, France, and Johannes Dreesmann, Public Health Agency of Lower Saxony, Hannover, Germany for helpful cooperations and for providing the data on *Salmonella Agona*, EHEC, meningococcal infections, and measles, respectively.

References

Anderson, H. and Britton, T. (2000). *Stochastic Epidemic Models and their Statistical Analysis*, New York: Springer.

Becker, N. (1989) *Analysis of Infectious Disease Data*. London: Chapman and Hall.

Clayton, D.G., Bernardinelli, L. and Montomoli, C. (1993). Spatial correlation in ecological analysis. *International Journal of Epidemiology*, **22**, 1193-1202.

Cox, D. (1981) Statistical analysis of time series. Some recent developments. *Scandinavian Journal of Statistics*, **8**, 93-115.

Denison, D.G.T., Holmes, C.C., Mallick, B.K. and Smith, A.F.M. (2002). *Bayesian Methods for Nonlinear Classification and Regression.* Chichester: Wiley.

Dietz, K. (1996). Biometric advances in infectious disease epidemiology. In: *Advances in Biometry* (eds. P. Armitage and H.A. David), New York: Wiley, 319-338.

Diggle, P.J. (1990) *Time Series. A Biostatistical Introduction.* Oxford: Oxford University Press.

Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data.* 2nd edition. Oxford: Oxford University Press.

Farrington, C.P. and Andrews, N. (2003). Outbreak detection: Application to infectious disease surveillance. In: *Monitoring the Health of Populations* (eds. R. Brookmeyer and D.F. Stroup), Oxford: Oxford University Press, 203-231.

Farrington, C.P., Andrews, N., Beale, A.D. and Catchpole, M.A. (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society Series A*, **159**, 547-563.

Farrington, C.P., Kanaan, M.N. and Gay, N.J. (2001). Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Applied Statistics*, **50**, 251-292.

Farrington, C.P., Kanaan, M.N. and Gay, N.J. (2003). Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics*, **4**, 279-295.

Finkenstädt, B.F., Bjornstad, O.N and Grenfell, B.T. (2002). A stochastic model for extinction and recurrence of epidemics: estimation and inference for measles outbreaks. *Biostatistics*, **3**, 493-510.

Guttorp, P. (1995). *Stochastic Modelling of Scientific Data.* London: Chapman and Hall.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, **19**, 2555-2567.

Knorr-Held, L. and Besag, J. (1998). Modelling risk from a disease in time and space. *Statistics in Medicine*, **17**, 2045-2060.

Knorr-Held, L. and Richardson, S. (2003). A hierarchical model for space-time surveillance data on meningococcal disease incidence. *Applied Statistics*, **52**, 169-183.

Mode, C.J. (1971) *Multitype Branching Processes – Theory and Applications*. American Elsevier Publishing Company, Inc.

Mugglin, A.S., Cressie, N. and Gemmell, I. (2002). Hierarchical modeling of influenza-epidemic dynamics in space and time. *Statistics in Medicine*, **21**, 2703-2721.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields. Theory and Applications*. CRC/Chapman and Hall.

Svensson, A. and Lindbäck, J. (2002). Statistical analysis of temporal and spatial distribution of reported Campylobacter infections. Proceedings of the International Biometric Conference 2002 in Freiburg, Germany, 7-20.

Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S*. Fourth Edition. New-York: Springer.

Whitaker, H.J. and Farrington, C.P. (2004). Infections with varying contact rates: application to Varicella. *Biometrics*, **60**, 615-623.

Zeger, S.L. and Qaqish, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics*, **44**, 1019-1031.