Statistical Modelling 2005; **5**: 187–199

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

Leonhard Held, Michael Höhle and Mathias Hofmann
Department of Statistics, University of Munich, Munich, Germany

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 time series. In addition, a multivariate formulation is proposed, which is well suited to capture space–time dependence caused by the spatial spread of a disease over time. An analysis of two multivariate time series is 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

Data and software link available from: http://stat.uibk.ac.at/SMIJ
Received March 2005; **revised** May 2005; **accepted** May 2005

1 Introduction

Recently there has been much 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 (e.g., Mugglin *et al.*, 2002; Svensson and Lindbäck, 2002; Knorr-Held and Richardson, 2003). Empirical Bayes techniques for time–space disease surveillance have been recently proposed by Böhning (2003).

For simplicity, consider first the simple univariate time-series case. Approaches to analyse 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 and possibly also serial dependence. Adjustments for seasonal correlation may be done through quasi-likelihood methods (Zeger, 1988;

Address for correspondence: Leonhard Held, Department of Statistics, University of Munich, Ludwigstr. 33, 80539 Munich, Germany. E-mail: leonhard.held@stat.uni-muenchen.de

 10.1191/1471082X05st098oa

Zeger and Qaqish; 1988). 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

$$\eta_t = \alpha + \beta t + \sum_{s=1}^{S} (\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t)) \tag{1.1}$$

where S is the number of harmonics to include and ω_s are the Fourier frequencies, for example $\omega_s = 2s\pi/52$ for weekly data (Diggle, 1990). Following the terminology of Cox (1981), this class of models can be called parameter-driven. Similar parameter-driven formulations with suitable prior distributions on latent parameters are used in state-space and related nonstationary models, where latent parameters are allowed to change over time (e.g., Jørgensen *et al.*, 1999 or Fahrmeir and Knorr-Held, 2000).

However, it was soon recognized that a purely parameter-driven approach is often not able to describe localized epidemics which are typical for infectious disease surveillance data, 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 Equation (1.1), 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 (1.1), that is 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 Qaqish (1988) 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, that is

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

This model can be seen as a size-dependent branching process (Diggle *et al.*, 2002) and has better properties; in particular, it allows for positive association between successive counts. However, the introduction of the parameter d and the regularization of past counts y_{t-1} through η_{t-1} are complicated and seem to be 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. They avoid regularization 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 article, 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:

$$\mu_t = \nu + \lambda y_{t-1}. \tag{1.2}$$

The disease incidence can thus be separated into two parts: an endemic part with rate ν and an epidemic part with conditional rate λy_{t-1}. Endemic incidence is persistent with a stable temporal, perhaps seasonal pattern. Nonendemic, that is, epidemic incidence will break out occasionally and eventually burn out (provided $\lambda < 1$). The distinction between endemic and epidemic incidences is quite common in dynamic models for infectious disease counts (e.g., Finkenstädt *et al.*, 2002).

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

$$\mu = \frac{\nu}{1-\lambda}, \quad \sigma^2 = \frac{\nu}{(1-\lambda)(1-\lambda^2)}. \tag{1.3}$$

Knowledge of the stationary mean $\mu = \nu/(1-\lambda)$ allows for a useful interpretation of λ. The epidemic incidence has stationary mean $\nu/(1-\lambda) - \nu = \lambda\nu/(1-\lambda)$ and hence, the ratio of epidemic to total stationary mean rate is simply λ.

The advantage of model (1.2) is that, without the endemic part, it can be interpreted as an approximation to a chain binomial model (Becker, 1989 for further details) without information on the number of disease susceptibles. Information on the number of susceptibles is only ever available in very special, much analysed data sets (e.g., Finkenstädt *et al.*, 2002). It is seldom, if ever, available in a surveillance setting, and so the approximation appears to be justified.

The additional influx of endemic cases with rate ν ensures that the process will not die out with probability 1, which is in contrast to the ordinary branching process. This is a useful addition, as infectious disease surveillance data often display a mixture of endemic and epidemic behaviours. Indeed, for λ close to 1, simulations from this model display occasional epidemic outbreaks, and so the formulation seems to be more realistic than a purely parameter-driven formulation. In contrast for $\lambda = 0$, the model reduces to a parameter-driven formulation with no epidemic incidence. Incidentally, model (1.2) is just a generalized linear model (GLM) with Poisson observation model and identity link, and so can be fitted with standard software.

In most applications, there may be a need to replace the Poisson distribution with a more flexible observation model to allow for overdispersion, for example, caused by the influence of unobserved covariates that affect the disease incidence. We will use a negative binomial model, where the conditional mean μ_t remains the same but the conditional variance σ_t^2 increases to

$$\sigma_t^2 = \mu_t + \frac{\mu_t^2}{\psi} = \mu_t\left(1 + \frac{\mu_t}{\psi}\right) \tag{1.4}$$

with the additional parameter $\psi > 0$, to be estimated from the available data. For $\psi \to \infty$, the negative binomial model equals the simple Poisson model.

Clearly, model (1.2) 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 Equation (1.1). The resulting model no longer fits into the GLM framework but numerical techniques can be used for likelihood inference. We will outline how the model can be extended further to the multivariate case, where the components of the time-series may relate to different age groups or to different geographical regions. In the latter case, spatio-temporal dependence can be even included in the model and we will illustrate this through an example of measles incidence in parts of Lower Saxony, Germany. Further extensions of the model with time- or area-dependent λ will be discussed in Section 4.

A particular and important advantage of our formulation is that the proposed model framework is easily estimated by ML using generic optimization routines, for example, the function optim() in R, so there is no need for simulation-based inference such as MCMC. Data and algorithms used will be made available in the R-package surveillance, available at www.statistik.lmu.de/ ~ hoehle/software/surveillance/.

This article is organized as follows. In Section 2, we show that this model can be fitted well to a typical time-series obtained from infectious disease surveillance. In Section 3, we extend the model to the multivariate case. The applicability of the multivariate models is illustrated with two-real world examples. We conclude in Section 4 with some additional discussions.

2 The univariate time-series case

Figure 1 illustrates a univariate time-series of length 312 of weekly counts for *Salmonella agona*, 1990–95, in the UK, provided by the Communicable Disease Surveillance Centre, UK. A similar time-series has been analysed in Farrington *et al.* (1996) in the

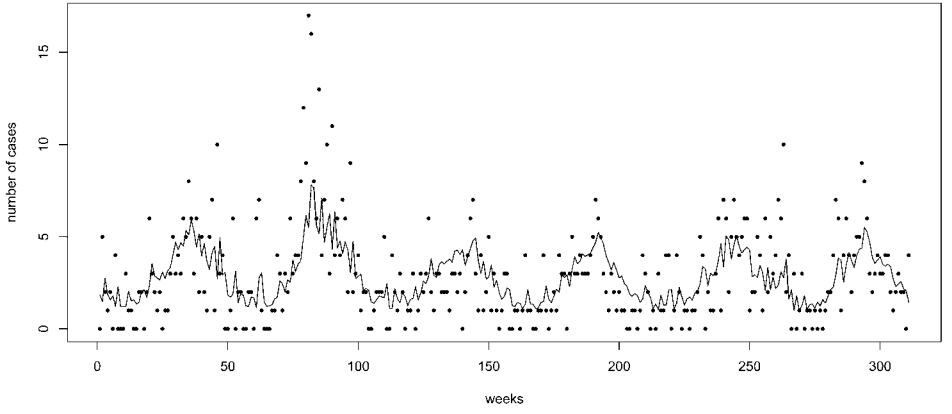

Figure 1 Observed (solid points) and fitted (solid line) counts of *Salmonella agona*, 1990–95

Table 1 Summary of the ML estimates (standard errors given in parantheses) of the different models for the *Salmonella agona* data

Model	Distribution	Seaso-nality	Auto-regression	$\hat{\lambda}$ (SE)	$\hat{\psi}$ (SE)	$\log L + 744$	Number of parameters
1	Poisson	No	No	–	–	0.0	2
2	Poisson	No	Yes	0.49 (0.03)	–	79.2	3
3	Poisson	Yes	No	–	–	84.0	4
4	Poisson	Yes	Yes	0.29 (0.03)	–	106.2	5
5	Negative binomial	No	No	–	2.1 (0.3)	70.6	3
6	Negative binomial	No	Yes	0.48 (0.04)	3.8 (0.8)	107.5	4
7	Negative binomial	Yes	No	–	4.0 (0.8)	111.7	5
8	Negative binomial	Yes	Yes	0.27 (0.04)	5.3 (1.3)	123.8	6

context of outbreak detection; however, the time-series used there is slightly different (and also slightly shorter) due to later modifications in the data file.

To these data, we fitted model (1.2) with ν replaced by $\nu_t = \exp(\eta_t)$ from Equation (1.1) with $S = 1$. 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 $2^3 = 8$ different models, depending on the observation model (Poisson or negative binomial), whether the seasonality terms have been included and whether the autoregressive component λy_{t-1} has been included in the linear predictor (1.2). The results are summarized in Table 1.

There are several interesting features to observe. There is clear evidence for over-dispersion, because the negative binomial models result in a significant increase in terms of maximized log-likelihood, denoted by $\log L$, compared with the corresponding Poisson models. With a mean incidence of $\mu = 2.875$ cases per week, the estimated variance/mean ratio is $1 + \mu/\hat{\psi} \approx 2.4$, 1.8, 1.7 and 1.5 in model 5–8, respectively, compare Equation (1.4). 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 the seasonality and the autoregressive components 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. It appears that this model gives a quite 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 different models. Predictive diagnostics are generally regarded as a useful tool to validate statistical models and are particularly easy to compute in the time-series context. On the basis of the data up to a certain week t, we have estimated all parameters in the model and then computed the predicted number of cases from the chosen model for the next week. In addition, an upper limit for the number of cases has been computed based on the quantiles of the Poisson and negative

Table 2 Predictive performance of the different models

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	Negative binomial	No	No	0.635	0.90	0.96	1.00
6	Negative binomial	No	Yes	0.557	0.93	0.98	0.99
7	Negative binomial	Yes	No	0.507	0.88	0.93	0.98
8	Negative binomial	Yes	Yes	0.484	0.87	0.94	0.99

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 (a square root transformation was used 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% that is, the proportion of observed counts that are smaller or equal to the corresponding upper prediction limit. We have chosen not to use two-sided prediction intervals in this validation step because the lower prediction limit will often be zero due to a relatively large predictive probability of observing zero cases. This will cause bias in the empirical coverage proportions because the observed counts can never be below zero.

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

In conclusion, we have shown that for this time series, a fairly simple model with only six 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-purpose R routine optim( ) for numerical optimization. Results are immediately available and are numerically stable. 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).

For comparison, we have fitted the same models as for *Salmonella agona* to weekly counts of Enterohaemorrhagic *Escherichia coli* (EHEC) infections in Bavaria, 2001–03 (not shown). The results have been quite different, particularly once the models are adjusted for seasonality; there was no need to include the autoregressive component and a purely parameter-driven model was sufficient. However, there was still evidence

for overdispersion, so the corresponding model 7 from Table 1 has been preferred on the basis of standard likelihood ratio tests. This example highlights a situation where the autoregressive term is not required.

3 A multivariate model extension

Consider now the case where multivariate time-series data are 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

$$\mu_{it} = \lambda y_{i,t-1} + n_{it}\nu \tag{3.1}$$

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

This model is aggregation consistent, because the aggregated counts $y_t = \sum_{i=1}^{m} y_{it}$ have mean

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

where $n_t = \sum_{i=1}^{m} n_{it}$. 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. Furthermore, under a Poisson model for y_{it}, y_t will still be Poisson-distributed, but this no longer holds for the negative binomial distribution.

Also note that in the Poisson case, the model can be written as a multivariate or multitype branching process with immigration (Mode, 1971) where

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

with suitable defined vectors μ_t, $\boldsymbol{y}_{t-1}$, ν and matrix Λ. For example, in model (3.1), Λ is simply diagonal with entries equal to λ; more elaborate specifications will be presented later. It is interesting that there exists a similar threshold theorem for multivariate branching process as in the univariate case. If the largest eigenvalue of Λ is smaller than unity, the process is ergodic with mean $\nu(\boldsymbol{I} - \Lambda)^{-1}$ (Mode, 1971, section 2.7). This formula is just the multivariate analogue of Equation (1.3).

As before, the assumption of a constant ν in Equation (3.1) is too strict, and we may, for example, replace ν by ν_{it}, where

$$\log \nu_{it} = \alpha_i + \beta t + \sum_{s=1}^{S} (\gamma_s \sin(\omega_s t) + \delta_s \cos(\omega_s t)). \tag{3.2}$$

Compared to Equation (1.1), the additional unit-dependent parameters α_i are allowed 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. Hence model (3.2) 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). Of course, one could also let the slope parameter β or the seasonal parameters γ_s and δ_s vary across areas, if appropriate, but we have not considered this further here.

Inclusion of dependence across the different series in the parameter-driven part of the model is more difficult. 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 subsequently. Depending on the context, this may be geographically neighbouring units as in the example of Section 3.2, or simply all other age groups, as in the example of Section 3.1.

3.1 Application to meningococcal infections in France

For illustration, we now consider monthly counts of meningococcal incidence in France, 1985–95. These data have previously been analysed in Knorr-Held and Richardson (2003) with focus on geographical variation. Here, we split the data into $m = 4$ age groups (<1, 1–5, 5–20, >20) and obtain a multivariate time-series of dimension 4.

Model (3.1) with ν replaced by ν_{it} from Equation (3.2) has been fitted to these data with $S = 1$ (now $\omega_s = 2\pi/12$). 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 with -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. Again, 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 2.

As mentioned earlier, a more general model may also consider the earlier number of cases in other age groups as potential explanatory variables for y_{it}. In the simplest case,

$$\mu_{it} = \lambda y_{i,t-1} + \phi \sum_{j \neq i} y_{j,t-1} + n_{it}\nu_{it} \tag{3.3}$$

with ν_{it} as in Equation (3.2), which introduces one additional parameter ϕ for the autoregressive effect of the other age groups. Written as a multivariate branching process, the matrix Λ now has diagonal entries λ and off-diagonal entries equal to ϕ.

For the meningitis data, the ML estimates of λ are nearly identical to the model without ϕ whereas the ML estimate of ϕ is -0.0004 (0.005), very close and not significantly different from zero. In addition, the maximized log-likelihood is still -1543.1, which clearly indicates that the component $\phi \Sigma_{j \neq i} y_{j,t-1}$ is not required 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

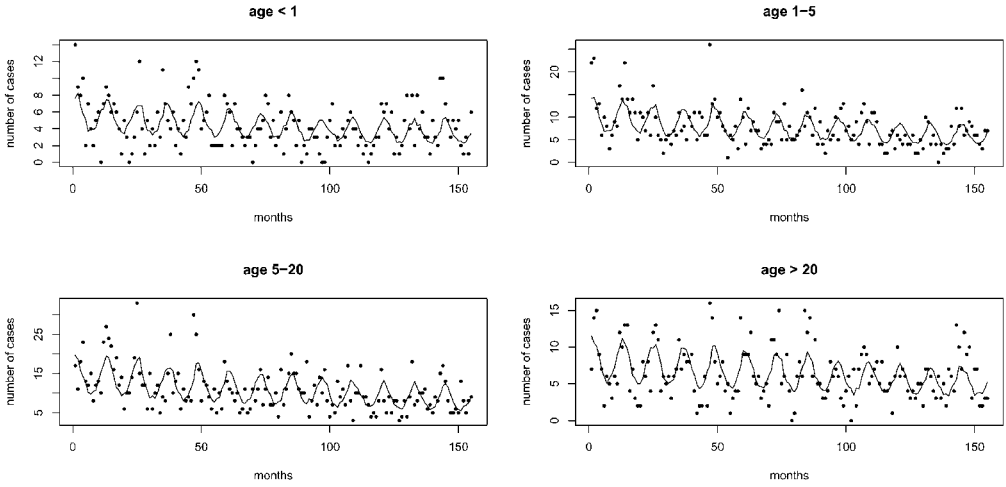

Figure 2 Observed (solid points) and fitted (solid lines) counts of meningococcal infections in the different age groups

(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, that is there is stronger contact within age groups than between age groups. We finally note that all these results are identical whether we used population counts n_{it} in Equation (3.1) or simply set $n_{it} = 1$. In fact, only the estimates of the intercept parameters α_i will change, but all other parameter estimates remain the same including the value of the maximized likelihood. This is because the available population data n_{it} did, in fact, not depend on t, in which case the intercepts α_i will completely adjust for the missing population counts.

We have also analysed the four age groups separately, using the basic formulations (1.2) and (1.1). The ML estimates of λ are 0.15 (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, respectively. 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 the measles cases from the corresponding $m = 15$ areas of this district (Figure 3 for a map of the area considered). Two areas have been omitted to avoid problems with nonexisting ML estimates. The data are shown in Figure 4 and also in http://www.nlga.niedersachsen.de for an animated movie of the 2002 epidemic (Click on 'Interaktiver Infektionsbericht 2003', then on 'Infektion & Hygiene', then on 'Infektionsdaten/(Epidemiologie', and then select 'Masern' and 'Diagramme, Zeitverlauf').

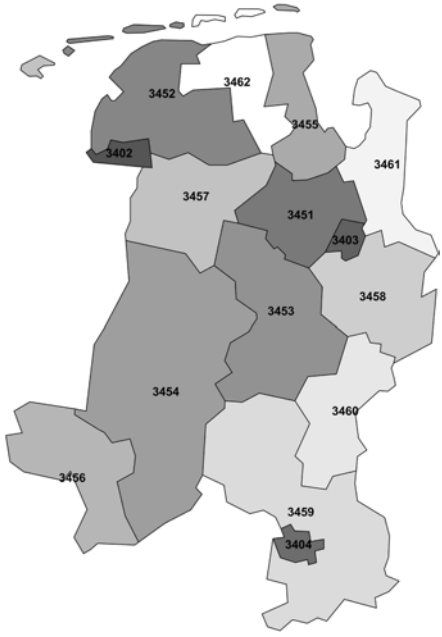

Figure 3 A map of the $m = 15$ areas in the administrative district 'Weser-Ems'

To these data, we fitted a model adopted from Equation (3.3),

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

with population fractions n_{it} and ν_{it} as in Equation (3.2), but the sum of the cases in other areas is now only over spatially adjacent areas $j \sim i$. Two areas have been defined

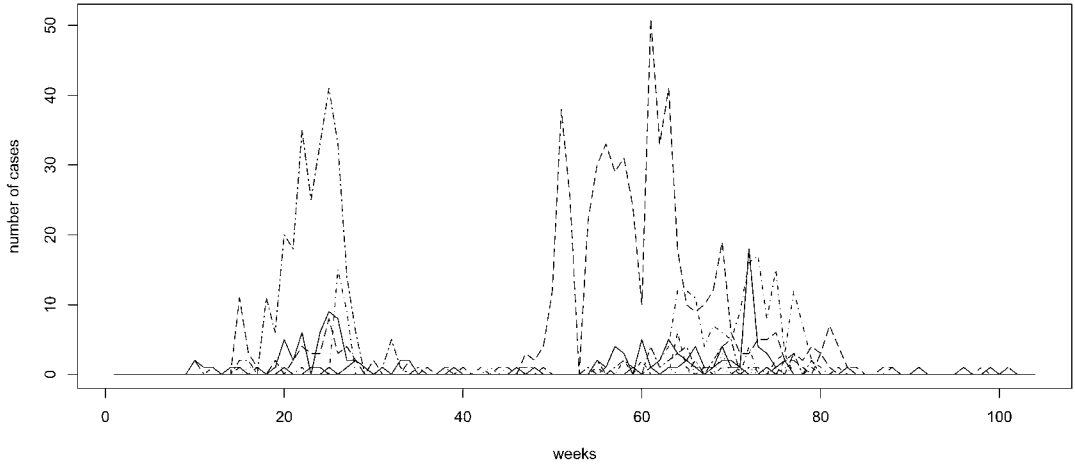

Figure 4 Weekly counts of new measles cases in $m = 15$ areas in the district 'Weser-Ems'

to be adjacent if they share a common border. With $S=1$ Fourier frequencies, the model thus has 21 parameters: the autoregressive parameters λ and ϕ, the linear time trend β, the seasonal parameters γ and δ, the overdispersion parameter ψ and $m=15$ area-specific intercepts α_i. 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$=-942.6$, which means that the space–time effect of previous counts in neighbouring areas is significant. This can also be seen from a pronounced decrease in log L of the model with $\phi=0$, leading to $\hat{\lambda}=0.66$ (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 previously. Incidentally, in the full model, the largest eigenvalue of the estimated matrix Λ is 0.69.

4 Discussion

In this paper we have proposed a statistical framework for the analysis of uni- and multivariate infectious disease counts. All models and analyses shown in this article can be easily repeated within seconds using standard optimization software. Thus, in contrast to methods based on MCMC, the model is particularly suited for statistical analysis in infectious disease surveillance. As a referee has noted, the model may also be applied to surveillance data on noninfectious diseases, although the branching process interpretation would then be slightly artificial.

However, some general caveats are appropriate. For infectious disease data, 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 or months. However, we have conducted simulation studies which showed that a Poisson branching process, aggregated to coarser time intervals, can be 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. As long as the under-reporting rate is roughly constant within an area over time, it can be absorbed well 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 due to both differences in incidence and different reporting rates. Finally, diagnostic test errors may also pose a problem, in particular, if notification of the disease requires laboratory reports of the isolation of the relevant pathogen. For more discussion of quality problems of surveillance data, refer Diggle *et al.* (2003).

Although spatial and temporal dependences can be accounted for in our model, the particular model chosen may not provide a good fit in some applications. In practice, one should assess whether there is any evidence for residual spatial or temporal dependence. If so, further model generalizations may be useful, but 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 infectiveness of an agent varies over time, for example, due to vaccination programmes, other interventions or 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. 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. Here the aim is to bring together spatial ecological regression (Clayton *et al.*, 1993) and infectious disease epidemiology.

Acknowledgements

This work is supported by the German Science Foundation (DFG), SFB 386, Projekt B9: 'Statistical methodology for infectious disease surveillance'. We thank the Communicable Disease Surveillance Center, UK, the Robert-Koch Institute (RKI), Berlin, Germany, the Institut National de la Veille Sanitaire, Saint Maurice, France and the Public Health Agency of Lower Saxony, Hannover, Germany, for providing the data on *Salmonella agona*, EHEC, meningococcal infections and measles, respectively. The article has improved through helpful comments made by three referees and Michael Toschke.

References

Becker N (1989) *Analysis of infectious disease data*. London: Chapman & Hall.

Böhning D (2003) Empirical Bayes estimators and non-parametric mixture models for space and time–space disease mapping and surveillance. *Environmetrics* **14**, 431–51.

Clayton DG, Bernardinelli L and Montomoli C (1993) Spatial correlation in ecological analysis. *International Journal of Epidemiology* **22**, 1193–202.

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

Denison DGT, Holmes CC, Mallick BK and Smith AFM (2002) *Bayesian methods for nonlinear classification and regression*. Chichester: Wiley.

Diggle PJ (1990) Time series. *A biostatistical introduction*. Oxford: Oxford University Press.

Diggle PJ, Heagerty P, Liang K-Y and Zeger SL (2002) *Analysis of longitudinal data*. 2nd edition. Oxford: Oxford University Press.

Diggle PJ, Knorr-Held L, Rowlingson B, Su T-L, Hawtin P and Bryant T (2003) On-line monitoring of public health surveillance data. In Brookmeyer R and Stroup DF, editors, *Monitoring the health of populations: statistical principles and methods for public health surveillance*, Oxford: Oxford University Press.

Fahrmeir L and Knorr-Held L (2000) Dynamic and semiparametric models. In Schimek M editors, *Smoothing and regression: approaches, computation and application*, New York: John Wiley & Sons.

Farrington CP, Andrews N, Beale AD and Catchpole MA (1996) A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society Series A* **159**, 547–63.

Finkenstädt BF, Bjornstad ON and Grenfell BT (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 & Hall.

Jørgensen B, Lundbye-Christensen S, Song PX-K and Sun L (1999) A state space model for multivariate longitudinal count data. *Biometrika* **86**, 169–81.

Knorr-Held L and Besag J (1998) Modelling risk from a disease in time and space. *Statistics in Medicine* **17**, 2045–60.

Knorr-Held L and Richardson S (2003) A hierarchical model for space–time surveillance data on meningococcal disease incidence. *Applied Statistics* **52**, 169–83.

Mode CJ (1971) *Multitype branching processes – theory and applications*. New York: Elsevier.

Mugglin AS, Cressie N and Gemmell I (2002) Hierarchical modeling of influenza-epidemic dynamics in space and time. *Statistics in Medicine* **21**, 2703–721.

Rue H and Held L (2005) *Gaussian Markov random fields. Theory and applications*. Boca Raton: CRC/Chapman & Hall.

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

Venables WN and Ripley BD (2002) *Modern applied statistics with S*. 4th Edition. New York: Springer.

Zeger SL (1988) A regression model for time series of counts. *Biometrika* **75**, 621–29.

Zeger SL and Qaqish B (1988) Markov regression models for time series: a quasi-likelihood approach. *Biometrics* **44**, 1019–31.