**University of
Zurich**^{UZH}

**Zurich Open Repository and
Archive**

University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2011

Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts

Paul, M ; Held, L

Abstract: Infectious disease counts from surveillance systems are typically observed in several administrative geographical areas. In this paper, a non-linear model for the analysis of such multiple time series of counts is discussed. To account for heterogeneous incidence levels or varying transmission of a pathogen across regions, region-specific and possibly spatially correlated random effects are introduced. Inference is based on penalized likelihood methodology for mixed models. Since the use of classical model choice criteria such as AIC or BIC can be problematic in the presence of random effects, models are compared by means of one-step-ahead predictions and proper scoring rules. In a case study, the model is applied to monthly counts of meningococcal disease cases in 94 departments of France (excluding Corsica) and weekly counts of influenza cases in 140 administrative districts of Southern Germany. The predictive performance improves if existing heterogeneity is accounted for by random effects.

DOI: <https://doi.org/10.1002/sim.4177>

Posted at the Zurich Open Repository and Archive, University of Zurich

ZORA URL: <https://doi.org/10.5167/uzh-53037>

Journal Article

Published Version

Originally published at:

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

DOI: <https://doi.org/10.1002/sim.4177>

Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts

M. Paul^{*†} and L. Held

Infectious disease counts from surveillance systems are typically observed in several administrative geographical areas. In this paper, a non-linear model for the analysis of such multiple time series of counts is discussed. To account for heterogeneous incidence levels or varying transmission of a pathogen across regions, region-specific and possibly spatially correlated random effects are introduced. Inference is based on penalized likelihood methodology for mixed models. Since the use of classical model choice criteria such as AIC or BIC can be problematic in the presence of random effects, models are compared by means of one-step-ahead predictions and proper scoring rules. In a case study, the model is applied to monthly counts of meningococcal disease cases in 94 departments of France (excluding Corsica) and weekly counts of influenza cases in 140 administrative districts of Southern Germany. The predictive performance improves if existing heterogeneity is accounted for by random effects. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: multivariate time series of counts; infectious diseases; random effects; proper scoring rules

1. Introduction

Surveillance data on various notifiable diseases collected by national surveillance systems usually consist of multiple time series of counts of new infections. Data for a single disease are typically reported in several strata defined through administrative geographical areas and/or age groups [1, Chapter 13]. The statistical analysis of the resulting multivariate time series of daily, weekly, or monthly counts is an important task in infectious disease epidemiology.

The data are not available at an individual level but are aggregated and often also subject to under-reporting and reporting delays, which may give rise to overdispersion and blurred dependencies. Thus, detailed (mechanistic) modelling of the epidemic process is not feasible. On the other hand, purely empirical models such as log-linear Poisson regression are not able to adequately capture occasional outbreaks. Such temporal dependence beyond regular patterns can be addressed by including past counts as additional explanatory variables in the predictor [2–4]. In particular, Held *et al.* [5] suggest a Poisson regression model with identity link for the mean, i.e. the disease incidence, which is divided into three additive components: the first two components represent an autoregression on past counts in the same and in other regions, respectively, whereas the third component deals with trends and seasonal variation in a usual log-linear formulation. Overdispersion can be allowed for by replacing the Poisson with a negative binomial distribution. Maximum likelihood (ML) estimates are obtained using general optimization routines because the resulting model no longer fits into a generalized linear model (GLM) framework.

When analyzing spatially stratified time series, the assumption of equal transmission rates or incidence levels across all regions is questionable. For example, disease transmission might be influenced by factors such as age, sex, vaccination status, genetic variation in individuals, or environmental factors [6].

Biostatistics Unit, Institute of Social and Preventive Medicine, University of Zurich, Switzerland

^{*}Correspondence to: M. Paul, Biostatistics Unit, Institute of Social and Preventive Medicine, University of Zurich, Hirschengraben 84, CH-8001 Zurich, Switzerland.

[†]E-mail: michaela.paul@ifspm.uzh.ch

Such factors could be incorporated into the model framework of Held *et al.* [5] as covariates if they are observable and available. Alternatively, to allow for varying transmission of a pathogen across a few regions, region-specific autoregressive parameters are introduced in Paul *et al.* [7]. However, this approach is limited to a small to moderate number of strata. For highly multivariate time series the estimation procedure can get unstable and identifiability problems may occur.

A common approach to account for unobserved heterogeneity is by means of random effects. For infectious disease data, Li *et al.* [8] allowed for household-dependent heterogeneity in infection rates by assuming that the probability of avoiding infection varies randomly in a chain binomial model. Similarly, Davis *et al.* [9] considered heterogeneity in transmission probabilities due to household-specific random effects when estimating vaccine efficacy based on outbreak size household data. Zeger and Karim [10] and Lin and Zhang [11] used a generalized linear (additive) mixed model (GLMM) for the analysis of longitudinal data on respiratory infection in Indonesian children.

In this paper, we introduce random effects in the model discussed in Held *et al.* [5] and extended in Paul *et al.* [7] to deal with heterogeneous disease transmission and incidence levels. In particular, we focus on how to obtain parameter estimates and on how to compare different models. The estimation of parameters involves integration of the likelihood with respect to the random effects which cannot be done analytically. There are several possible methods for approximating of the integral. For instance, the integral can be solved numerically using adaptive Gauss–Hermite (AGH) quadrature or Monte Carlo integration techniques. Alternatively, the integrand can be approximated such that the integral over the resulting approximation is a closed-form expression. This can be done using first or higher order Laplace approximations or penalized quasi-likelihood (PQL) approaches. See e.g. Breslow [12] or Tuerlinckx *et al.* [13] and the references therein for details about these methods in GLMMs.

However, since the model considered in this paper is not linear in parameters, methods for GLMMs cannot be used. Inference is thus based on the methodology suggested by Kneib and Fahrmeir [14] for the analysis of survival data. Estimates of the regression coefficients are derived using penalized likelihood and estimates of variance components are obtained by optimizing the (approximated) marginal likelihood. The estimation procedure corresponds to a variant of the PQL approach as discussed in Breslow and Clayton [15].

Alternatively, a Bayesian approach could have been used for inference. For example, Held *et al.* [16] have proposed a Markov chain Monte Carlo (MCMC) algorithm for a univariate time series model with an autoregressive parameter following a piecewise constant change-point model with unknown number of change-points. This model is extended to a multivariate setting in Hofmann [17]. An extension to a multivariate additive time series model with random effects seems feasible. Knorr-Held and Richardson [18] have considered Bayesian inference via MCMC in a somewhat related multivariate model with multiplicative structure. However, model choice in such highly parameterized Bayesian hierarchical models remains a challenge. Of course, the deviance information criterion (DIC) [19] can be computed, but recent work [20] has shown that it tends to prefer too complex models in situations with many random effects.

Similar challenges arise in a frequentist setting where model choice is often based on Akaike's information criterion [21]

$$\text{AIC} = -2\log(L(\hat{\theta}|\text{data})) + 2p, \quad (1)$$

where $\log(L(\hat{\theta}|\text{data}))$ is the maximized log-likelihood of a model and p is the number of estimable parameters in this model, or related criteria. Extending the AIC to mixed models raises some issues: First of all, one has to decide on parameters of primary interest in order to choose the appropriate likelihood and the correct number of parameters in Equation (1) [22]. From a marginal point of view the marginal likelihood integrated over the random effects is used and the number of estimable parameters is given as the sum of the number of fixed effects and the number of variance parameters. This formulation is appropriate if fixed population effects are of interest. However, if future observations share the same random effects as the observed data (as is the case in our situation), the AIC should be based on the conditional likelihood given the random effects [22, 23]. The effective number of parameters then lies between that of a model without random effects and that of a model where the random effects are replaced by corresponding fixed effects. Obtaining an unbiased estimator for the effective number of parameters is computationally quite complex already for linear mixed models [22, 24] and not clear for GLMMs. Besides, the use of either the conditional or the marginal AIC for model selection has

serious shortcomings e.g. if one wants to decide whether the inclusion of a random effect is necessary or not, i.e. whether the variance is non-zero or zero [23].

When using the Bayesian information criterion $BIC = -2 \log(L(\hat{\theta}|\text{data})) + p \log(n)$ the above considerations concerning the choice of an appropriate likelihood and the corresponding number of parameters remain. Furthermore, the effective sample size n is also ambiguous because of the correlation between observations, see e.g. [25, 26].

A more natural approach for model selection in time series models than the use of classical model choice criteria is the comparison of successive one-step-ahead forecasts with the actually observed data [27, 28]. The usage of proper scoring rules has been recently advocated in this context [29]. The most widely known scoring rule is the logarithmic score which corresponds to the log predictive density at the observed value. In the case of time series where data are inherently ordered, BIC is an approximation of the log marginal likelihood (integrated over unknown parameters), which can be written as the sum of the one-step-ahead logarithmic scores. See Dawid [27] or Gneiting and Raftery [29] for further details.

Calculation of successive one-step-ahead forecasts with MCMC is very difficult. Specific algorithms such as particle filters have been suggested [30] but application to highly multivariate time series data seems not feasible. In contrast, the likelihood-based approach proposed in this paper allows for successive re-fitting the model and calculating the one-step-ahead forecasts.

The remainder of the paper is organized as follows. In Section 2 the model described in Paul *et al.* [7] is extended for the analysis of highly multivariate time series with heterogeneity in some of the model coefficients by means of (correlated) random effects. Statistical inference based on mixed model methodology is described. Software for fitting the models is currently incorporated in the development version of the R package *surveillance* [31] at <http://surveillance.r-forge.r-project.org/>. The predictive properties are investigated by means of one-step-ahead predictions and proper scoring rules, as outlined in Section 3. In Section 4.1 we apply our model to monthly meningococcal disease counts in 94 departments of France excluding Corsica from 1985 to 1999, previously analyzed in Knorr-Held and Richardson [18]. The incidence for meningococcal disease is rather low, resulting in a sparse data set. In a second example in Section 4.2 we consider a more common disease, namely influenza cases observed in the 140 districts of the two German states Baden-Württemberg and Bavaria from 2001 to 2008. We end with a discussion in Section 5.

2. Model formulation and inference

2.1. Model

To begin with, let y_{rt} denote the number of cases of a specific disease in ‘unit’ $r = 1, \dots, R$ at time $t = 1, \dots, T$. Units might represent several geographical regions or different age groups. In the following, we consider spatially stratified counts where a unit corresponds to a certain region. The counts are assumed to be Poisson distributed, $y_{rt}|y_{t-1} \sim \text{Po}(\mu_{rt})$, with conditional mean

$$\mu_{rt} = \lambda_r y_{r,t-1} + \phi_r \sum_{q \neq r} w_{qr} y_{q,t-1} + v_{rt}, \quad \lambda_r, \phi_r, v_{rt} > 0 \quad (2)$$

see [7]. The first two components of μ_{rt} include as covariates the number of past cases at time $t-1$ observed in the same and in other regions, respectively. The weights w_{qr} are assumed to be known and define how cases in other regions relate to cases in region r . In this paper, we use $w_{qr} = 1(q \sim r)/n_q$, where 1 is the indicator function, $q \sim r$ denotes that region q is a neighbor of region r , and n_q denotes the number of neighbors of region q . See Paul *et al.* [7] for a discussion of other possible weights. This observation-driven part of the conditional mean rate should capture occasional outbreaks and, following Held *et al.* [5], is called the ‘epidemic component’. The third component v_{rt} is the ‘endemic component’ and parametrically models seasonal variation and trends.

The specification of the unknown quantities, λ_r , ϕ_r , and v_{rt} , in the predictor (2) is discussed in more detail later. At first, we motivate the chosen additive decomposition. There has been much literature on how to generalize the well-known Gaussian autoregressive moving average (ARMA) models to observation-driven models for non-Gaussian time series data, e.g. [2–4]. In the case of a univariate Poisson time series, many models use the GLM framework with log link function and regress $\log(\mu_t)$ on past values of the response to address autocorrelation (e.g. references in [4]). However, it is not possible to account for positive association in a stationary model if past counts are directly included

as explanatory variables. Instead, a suitable transformation of y_{t-1} is necessary, see e.g. [3, 7]. Models belonging to the class of conditional linear autoregressive models use an identity link and directly regress μ_t on past cases without further transformations. For instance, Fokianos *et al.* [4] suggest the following model for a time series of counts $\{y_t\}$:

$$y_t | (\text{History up to time } t) \sim \text{Po}(\mu_t), \quad \mu_t = a\mu_{t-1} + by_{t-1} + d, \quad a, b, d > 0.$$

Without the moving average term $a\mu_{t-1}$, this model represents a univariate version of model (2). It corresponds to a Bienaymé–Galton–Watson branching process with immigration [32]. Branching processes play a fundamental role in the mechanistic modelling of infectious diseases and provide an approximation of the epidemic process assuming an unlimited amount of susceptibles [33]. In a surveillance setting, the number of susceptibles is rarely available and branching processes provide a useful model approximation.

Formulation (2) of our empirical model is based on this branching process formulation and can also be written in a multivariate fashion as

$$\mu_t = \Lambda y_{t-1} + v_t \quad (3)$$

with suitably defined column vectors μ_t, y_{t-1} , and v_t . The matrix Λ has entries λ_r on the diagonal and off-diagonal entries $\phi_r w_{qr}$ [7]. Stationarity holds if the largest eigenvalue of the matrix Λ is smaller than 1 [5].

The unknown quantities in (2) are now decomposed additively on the log scale

$$\log(\lambda_r) = \alpha^{(\lambda)} + b_r^{(\lambda)}, \quad (4)$$

$$\log(\phi_r) = \alpha^{(\phi)} + b_r^{(\phi)}, \quad (5)$$

$$\log(v_{rt}) = \alpha^{(v)} + b_r^{(v)} + \gamma_1 t + \sum_{s=1}^S \{\gamma_{2s} \sin(\omega_s t) + \gamma_{2s+1} \cos(\omega_s t)\} + \log(e_{rt}), \quad (6)$$

where $\alpha^{(\lambda)}, \alpha^{(\phi)}, \alpha^{(v)}$ are component-specific intercepts, $b_r^{(\lambda)}, b_r^{(\phi)}, b_r^{(v)}$ are random effects, γ_1 is a trend parameter, $\gamma_2, \dots, \gamma_{2S+1}$ are seasonal parameters, ω_s denotes a specific Fourier frequency, and e_{rt} is a region-specific and time-dependent offset. For instance, $\omega_s = 2\pi s/52$ is used for weekly data [34]. The offset e_{rt} might comprise yearly varying population numbers. All in all, the vector of the fixed (unpenalized) effects is given by $\beta = (\alpha^{(\lambda)}, \alpha^{(\phi)}, \alpha^{(v)}, \gamma_1, \dots, \gamma_{2S+1})^\top$. However, the parametric modelling of trends and seasonal variation is not restricted to the formulation given in (6). For instance, one could choose a quadratic instead of a linear trend.

Note that to address heterogeneity, the component-specific intercepts $\alpha^{(\lambda)}, \alpha^{(\phi)}$, and $\alpha^{(v)}$ were allowed to vary across regions in Paul *et al.* [7]. Such a formulation is feasible as long as the number of regions is low to moderate. In this paper, each component (4)–(6) instead includes an (optional) random effect. The stacked vector $\mathbf{b} = (\mathbf{b}^{(\lambda)\top}, \mathbf{b}^{(\phi)\top}, \mathbf{b}^{(v)\top})^\top$ containing all random effects from the three components is assumed to be normally distributed with mean $\mathbf{0}$ and positive definite covariance matrix

$$\Sigma = \text{diag}(\sigma_\lambda^2 \mathbf{I}, \sigma_\phi^2 \mathbf{I}, \sigma_v^2 \mathbf{I}), \quad (7)$$

where $\sigma_\lambda^2, \sigma_\phi^2$, and σ_v^2 are unknown variance parameters, and $\mathbf{I}$ is the $R \times R$ identity matrix. Note that we use the identity matrix for all components, i.e. all elements of $\mathbf{b}$ are assumed to be uncorrelated.

However, this assumption might not always be realistic. To allow for correlation between different components, the covariance matrix for the stacked vector $\mathbf{b}$ can be specified as

$$\Sigma = \Omega \otimes \mathbf{I}, \quad (8)$$

where Ω is an unknown 3×3 covariance matrix, and $\otimes$ denotes the Kronecker product. Positive definiteness of Σ can be ensured by using a suitable parameterization for Ω . See Pinheiro and Bates [35] for a comparison of several parameterizations for unstructured covariance matrices. We use a computationally stable factorization in terms of standard deviations and correlations which is based on the Cholesky decomposition of Ω in spherical coordinates, see Appendix A for further details.

In hierarchical models for spatio-temporal data, the independence assumption for a vector of random effects might be questionable and random effects are often assumed to be spatially correlated. As the observation-driven formulation (2) is able to incorporate temporal and spatio-temporal correlation, the

inclusion of independent and identically distributed (IID) Gaussian effects in (2) might be sufficient to address possible further heterogeneity. However, the IID random effects within a component might also be replaced by spatially correlated ones.

For example, one might adopt a conditional autoregressive (CAR) model [36, 37] for $\mathbf{b}^{(v)}$, say, i.e. effects in neighboring regions are assumed to be more alike than effects in distant regions. The respective identity matrix in (7) is then replaced by a matrix $\mathbf{K}$ with non-diagonal entries $k_{qr} = -1$ if $q \sim r$, and diagonal entries $k_{qq} = n_q$ [37, p. 102]. As the rows and columns of $\mathbf{K}$ sum to zero, this matrix is not of full rank but of rank $g < R$, leading to an improper distribution. The rank-deficiency of $\mathbf{K}$ equals $R - g = 1$ if all regions are connected and there are no islands. The inverse of $\mathbf{K}$, which is needed to obtain a marginal likelihood estimate for the unknown variance parameter, thus does not exist. However, the vector $\mathbf{b}^{(v)}$ can be re-expressed via a one-to-one transformation in terms of a $(R - g)$ -dimensional fixed, i.e. unpenalized parameter vector, and a g -dimensional parameter vector which is IID Gaussian [14, 37, p. 91].

Model (2) assumes a Poisson distribution for the counts. This is usually suitable for counts of less prevalent diseases such as the meningococcal disease data analyzed in Section 4.1. For counts of more prevalent diseases, such as influenza, overdispersion is more likely [38]. The model formulation can be easily adjusted for overdispersion by replacing the Poisson distribution with the negative binomial, i.e. assuming $y_{rt} | \mathbf{y}_{t-1} \sim \text{NegBin}(\mu_{rt}, \psi)$, where the conditional mean μ_{rt} is specified exactly as in the Poisson case but the conditional variance increases from μ_{rt} to $\mu_{rt}(1 + \psi\mu_{rt})$ [5, 7]. Thus, the negative binomial model simplifies to the Poisson model for $\psi = 0$. Also note that in applications, each component (4)–(6) of the conditional mean may be omitted in parts or as a whole and the vector $\boldsymbol{\beta}$ then only contains the included fixed effects. Analogously, the vector $\mathbf{b}$ only contains the specified random effects with corresponding covariance matrix $\boldsymbol{\Sigma}$.

2.2. Inference

ML estimates in the model formulation without random effects can be obtained by directly maximizing the respective Poisson or negative binomial log-likelihood using numerical optimization routines. See Paul *et al.* [7] for further details. In the presence of random effects, inference is more complex. In this paper, we use penalized likelihood approaches [14, 15, 39–41] to obtain parameter estimates. That is, variance components are treated as fixed when estimating the fixed and random effects. The variance components itself are estimated through maximizing the marginal likelihood after integration with respect to the fixed and random effects. As there is no analytical solution to this integral, a Laplace approximation to the marginal likelihood is used (see e.g. [14, 15]).

In brief, the following alternating algorithm for the estimation of all parameters is used: First update regression coefficients given the current variance parameters, then update variance components given current regression coefficients via Newton steps. Iterate these two steps until convergence is reached, i.e. parameter estimates no longer change.

Inference for the regression parameters $\boldsymbol{\beta}, \mathbf{b}$, given known variance components, is based on the penalized log-likelihood

$$\ell_{\text{pen}}(\boldsymbol{\beta}, \mathbf{b}; \boldsymbol{\Sigma}) = \ell(\boldsymbol{\beta}, \mathbf{b}) + \log p(\mathbf{b} | \boldsymbol{\Sigma}). \quad (9)$$

For simplicity, we assume that the counts are Poisson distributed. The corresponding log-likelihood in (9) is then given by

$$\ell(\boldsymbol{\beta}, \mathbf{b}) = \sum_{r,t} y_{rt} \log(\mu_{rt}) - \mu_{rt} - \log \Gamma(y_{rt} + 1),$$

where $\Gamma(\cdot)$ denotes the Gamma function. In the case of overdispersed data, the above Poisson log-likelihood is replaced by the negative binomial log-likelihood, see Paul *et al.* [7]. The penalty term $\log p(\mathbf{b} | \boldsymbol{\Sigma})$ in (9) corresponds to the log distribution of the vector of random effects $\mathbf{b}$, i.e. a multivariate Gaussian. After dropping terms that are constant with respect to $\mathbf{b}$, we have

$$\log(p(\mathbf{b} | \boldsymbol{\Sigma})) = -\frac{1}{2} \mathbf{b}^T \boldsymbol{\Sigma}^{-1} \mathbf{b}.$$

The score vector $s_{\text{pen}}(\boldsymbol{\beta}, \mathbf{b}; \boldsymbol{\Sigma})$ can be partitioned into two parts defined by the first-order derivatives of (9) with respect to the fixed and random vectors $\boldsymbol{\beta}$ and $\mathbf{b}$. Analogously, the observed Fisher information matrix $\mathbf{F}_{\text{pen}}(\boldsymbol{\beta}, \mathbf{b}; \boldsymbol{\Sigma})$, i.e. the negative second-order derivatives of (9), can be partitioned with respect to

β and $\mathbf{b}$. See Appendix A.1 for details. These quantities $s_{\text{pen}}(\cdot)$ and $\mathbf{F}_{\text{pen}}(\cdot)$ allow the computation of updated estimates for the regression coefficients given the variances in a Newton algorithm.

Estimates of the variance components are obtained by maximizing the marginal likelihood

$$L_{\text{marg}}(\Sigma) = \int \exp\{\ell_{\text{pen}}(\beta, \mathbf{b}; \Sigma)\} d\beta d\mathbf{b}. \quad (10)$$

The above integral cannot be solved analytically and one has to resort to numerical methods or approximations. Solving the integral in (10) by means of a Laplace approximation [42] is a relatively simple approach and computationally less complex than other methods mentioned in Section 1. AGH quadrature can provide a more accurate approximation in some situations since the Laplace approximation corresponds to AGH quadrature with a single quadrature point [43]. However, the computational complexity of AGH quadrature rapidly increases with the dimension of the random effects vector.

In general, the performance of first-order Laplace approximations depends on the validity of the normal approximation. PQL approaches for GLMMs [15] have been found to yield seriously biased estimates of variances and regression parameters with sparse binary outcome data where data are far from normal [44, 45]. However, in many situations Laplace approximations work quite well and there is no need for more elaborate approximations. Breslow [12] investigated several situations and concluded that PQL performs adequately for Poisson GLMMs with mean greater than 5 or even lower for many problems. Kneib and Fahrmeir [14] showed that the use of a Laplace approximation in their mixed model-based inference performed quite similar compared with MCMC-based inference for survival data.

Applying a Laplace approximation to the marginal likelihood (10) results in

$$\ell_{\text{marg}}(\Sigma) \approx \ell(\hat{\beta}, \hat{\mathbf{b}}) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} \hat{\mathbf{b}}^{\top} \Sigma^{-1} \hat{\mathbf{b}} - \frac{1}{2} \log |\mathbf{F}_{\text{pen}}(\beta, \mathbf{b}; \Sigma)|,$$

where $|\cdot|$ denotes the determinant of a matrix. Note that both $\hat{\beta}$ and $\hat{\mathbf{b}}$ are estimated for given Σ and thus depend on the variance parameters. In GLMMs [15], generalized additive models [46] and Cox-type hazard rate models [14], small changes in the variance parameters hardly affect the regression coefficients. We thus assume also that both $\ell(\hat{\beta}, \hat{\mathbf{b}})$ and $\hat{\mathbf{b}}$ vary only slowly when changing the variance components. Then, the first term of the approximation can be ignored and the log marginal likelihood reduces to

$$\ell_{\text{marg}}(\Sigma) \approx -\frac{1}{2} \log |\Sigma| - \frac{1}{2} \mathbf{b}^{\top} \Sigma^{-1} \mathbf{b} - \frac{1}{2} \log |\mathbf{F}_{\text{pen}}(\beta, \mathbf{b}; \Sigma)| \quad (11)$$

where $\mathbf{b}$, β denote a fixed value not depending directly on the variances, i.e. a current estimate. This expression can be maximized numerically via Newton steps to obtain estimates of the variance parameters.

First and second derivatives of the approximate log marginal likelihood (11) can be derived based on differentiation rules for matrices (see e.g. [47]). See Appendix A for further details about the estimation procedure and the implementation in the R package *surveillance* [31].

3. Predictive model assessment using proper scoring rules

A classical approach to measure the performance of a model is to use e.g. the mean squared error of several point predictions. However, such an approach does not take into account the uncertainty associated with the point predictions. Instead, probabilistic predictions in the form of a predictive probability distribution P should be considered.

In the recent years, the use of strictly proper scoring rules has been advocated to evaluate probabilistic predictions [29, 48]. Scoring rules measure the predictive quality by assigning a numerical score, $S(P, y)$, based on a stated predictive distribution P and the later observed true value y . They can be seen as negatively oriented penalties that are to be minimized. A scoring rule is said to be proper, when its expected value under the stated probability distribution P becomes minimal if the observed value y is indeed a realization from P [29]. It is strictly proper if this minimum is unique. Strict propriety is thereby essential to ensure that a scoring rule simultaneously addresses sharpness—the concentration of the predictive distribution—and calibration—the statistical consistency between the predicted and the later observed probability distributions. See Gneiting and Raftery [29] for further details.

Each scoring rule has differing properties and strengths. Unless there is a clearly defined unique underlying decision problem which calls for a specific scoring rule, it may often be appropriate to investigate several scores in applications [48]. In the following, we briefly describe a selection of proper scoring rules for count data discussed in Czado *et al.* [48]. Denote $P(Y=k)$ the probability mass function, $P(Y \leq k)$ the cumulative distribution function and μ_P the first (finite) moment of the predictive probability distribution. Furthermore, let y denote the count that materializes.

A traditional summary measure of predictive performance is the squared error score

$$\text{SES}(P, y) = (y - \mu_P)^2$$

defined in analogy to the mean squared error. This scoring rule depends on the predictive distribution only through the first moment μ_P and is proper, but not strictly proper. Nevertheless, we consider this score in Section 4 for comparison, due to its widespread use. Perhaps the most popular strictly proper scoring rule is the logarithmic score [49]

$$\log S(P, y) = -\log(P(Y=y)).$$

The logarithmic score is a local score. It gives no credit for assigning high probabilities to values near but not identical to the count y that materializes. At the same time it is highly sensitive to extreme cases as it strongly penalizes low probability events. A strictly proper scoring rule which is less sensitive to extreme events than the logarithmic score is the ranked probability score [48, 50]

$$\text{RPS}(P, y) = \sum_{k=0}^{\infty} (P(Y \leq k) - 1(y \leq k))^2.$$

The RPS adds particular weight to situations with unusually high observed or predicted counts. In such situations, differences between competing models are blown up [48].

Typically, mean scores over a set of predictions are used to rank and compare different models. This is mostly done informally by ordering the obtained mean scores. However, the question whether score differences between models are significant can also be addressed more formally via tests [51, 52]. Here we will look at a series of one-step-ahead predictions for each considered model. Two models are compared with a Monte Carlo permutation test for paired individual scores [53]. As test statistic the difference between mean scores from model A and mean scores from model B is used. Both scores are averaged over a set of n individual scores. Under the null hypothesis of no difference, the actually observed difference between mean scores should not be notably different from the distribution of the test statistic under permutation. For each permutation, we first randomly assign the membership of the n individual scores to either model A or B with probability 0.5. We then compute the respective difference in mean for models A and B in this permuted set of scores. As the computation of all possible permutations is only feasible for small data sets, a random sample of permutations is used to obtain the null distribution. The Monte Carlo p -value is then given by

$$\frac{1 + \{\text{number of permuted differences larger than observed difference (in absolute value)}\}}{1 + \text{number of permutations}}.$$

For the model presented in Section 2, the Poisson distribution is used as predictive distribution. The required parameters of the distribution are obtained through plugging in the estimates of the fixed and random effects. Although such an approach ignores parameter uncertainty, it is the most common and feasible method to obtain a predictive distribution. Fixing the parameters at their estimated values may in exceptional cases lead to identical scores for models of different complexities. For instance, a model with an estimated zero random effects variance would have the same predictions and thus the same scores as a corresponding simpler model without these random effects. In that case a heuristic solution is to always prefer the more parsimonious model among models with equal scores. Such a situation would not occur e.g. in a fully Bayesian approach where the model with random effects would yield a larger predictive variance than the model without these effects and thus lead to different scores.

Note that for a series of one-step-ahead predictions, the model needs to be re-fitted at each time-point. However, the optimizing algorithm usually converges after a few steps when using the previous estimates as initial values because estimates hardly change. A simpler but cruder approach would be not to update the parameter estimates (see e.g. [54]).

4. Applications

In the following we analyze and assess the predictive performance of the proposed model using two spatio-temporal data sets. The first data set consists of monthly cases of meningococcal disease caused by the *Neisseria meningitidis* bacterium and observed in 94 departments of France. The incidence of meningococcal disease is low and there is evidence of geographical heterogeneity within France [18]. The second data set comprises the weekly number of laboratory confirmed influenza A and B cases in 140 administrative districts in Southern Germany, obtained from the German national surveillance system operated by the Robert Koch Institute (RKI) [55]. The incidence of influenza is very high and the disease spreads across regions in seasonal waves.

In both applications, we consider several models that differ depending on whether autoregressive parameters are included or not, and on whether parameters are treated as fixed or as random. At first, we only consider models with fixed and IID random effects. In the case where there are IID random effects in more than one component, these effects are assumed to be correlated, as this correlated formulation is more flexible than the formulation with uncorrelated effects. Among all models considered, the best model with the lowest mean scores is selected. This model is then compared with a corresponding model with CAR instead of IID effects.

4.1. Meningococcal disease in France

Knorr-Held and Richardson [18] proposed a complex hierarchical model for the analysis of meningococcal disease incidence. Latent parameters are used to capture temporal, seasonal, and spatial trends. Potential autoregressive effects of counts in the same or neighboring regions at previous time points are modulated by latent binary indicators. These indicators are assumed to follow a two-stage hidden Markov model. Thus past counts act multiplicatively on the disease incidence instead of additively as in (2). Fully Bayesian inference is done using MCMC techniques.

In a case study, the model was applied to data on monthly counts of meningococcal disease in 94 departments of France (excluding Corsica) from 1985 to 1997. Different specifications for the functional form of the autoregressive term, including and excluding spatial neighbors, were compared based on the criterion DIC. Additional data for the years 1998 and 1999 were available for the purpose of prediction but have not been included in the analysis. The main focus was placed on model fit and predictive performance was touched only briefly.

Figure 1 shows the monthly number of cases for the whole of France for the years 1985–1999. A clear seasonal pattern with peaks in winter can be seen. Also shown is a map of the mean yearly incidence per 100 000 habitants averaged over the 15 years. The incidence is regionally varying with the lowest incidence concentrated in the central regions of France. Note that this heterogeneity could be partly due to the differences in reporting practices since the data correspond to the compulsory cases notified by clinicians [18]. The location of four selected departments that also served as illustration in Knorr-Held and Richardson [18] are indicated in the map.

Note that in a few regions there seem to have been artefacts in the data collection because a period with no cases at all is followed by a period with regular pattern of cases. To determine such regions we assumed a Poisson-gamma model with a single change-point in the mean of unknown location and compared it with a Poisson-gamma model with constant mean [56, Section 3.2.1 and Appendix B.5] for each region. If the posterior probability of the change-point model, assuming equal prior probability for both models, was larger than 0.8, the zero counts for the first month up to the estimated change-point were set to missing for the analysis. Eventually, the first two years in department ‘Maine-et-Loire’ and the first three years in departments ‘Calvados’ and ‘Marne’ were treated as missing.

Since the data are quite sparse we assume a Poisson model. Table I summarizes the results for different model formulations. All models contain a quadratic trend, $S=1$ seasonal component and include expected cases e_{rt} calculated by indirect age–sex standardization as offset in (6). See Knorr-Held and Richardson [18] for further details on the calculation of expected cases. Table I also shows the log-likelihood values for each model. For models without random effects, there is no penalty term involved and accordingly, the marginal likelihood ℓ_{marg} is denoted by NA in the table. In this case, ℓ_{pen} corresponds to the usual Poisson likelihood and could be used for model comparison. In the presence of random effects, both likelihoods are given. However, note that they cannot be used for model comparison.

Figure 1. Monthly number of meningococcal disease cases in the whole of France (left) and mean yearly incidence per 100 000 habitants in 94 departments, averaged over the years 1985–1999 (right).

Table I. Analysis of meningococcal disease in 94 departments of France.							
Model	$\exp(\hat{\alpha}^{(\lambda)})$ (se)	$\hat{\alpha}^{(v)}$ (se)	$\hat{\sigma}_{\lambda}^2$	$\hat{\sigma}_v^2$	$\hat{\rho}_{\lambda v}$	maxEV	$\ell_{\text{pen}} (\ell_{\text{mar}})$
<i>No autoregression</i>							
A0	—	−0.168 (0.022)	—	—	—	—	−10781 (NA)
A1	—	★	—	—	—	—	−10431 (NA)
A2	—	−0.253 (0.046)	—	0.14	—	—	−10480 (−146)
<i>Fixed intercept $\alpha^{(\lambda)}$</i>							
B1	0.076 (0.009)	★	—	—	—	0.08	−10387 (NA)
B2	0.079 (0.009)	−0.327 (0.047)	—	0.14	—	0.08	−10436 (−139)
<i>IID random effects $\mathbf{b}^{(\lambda)}$</i>							
C2	0.087 (0.010)	−0.339 (0.047)	0.28	0.14	−0.27	0.22	−10422 (−162)
<i>Fixed intercept $\alpha^{(\lambda)}$, CAR random effects $\mathbf{b}^{(v)}$</i>							
D	0.079 (0.009)	−0.328 (0.025)	—	0.50	—	0.08	−10438 (−139)

Shown are parameter estimates with standard errors in brackets. All models contain a quadratic trend and $S=1$ seasonal term. The maximum eigenvalue of the estimated matrix $\mathbf{A}$ in (3) is denoted by maxEV. Fixed but region-specific intercepts are indicated by ★. The numerals in the model labels correspond to 0: fixed intercept $\alpha^{(v)}$, 1: region-specific intercept $\alpha_r^{(v)}$ 2: IID random effects $\mathbf{b}^{(v)}$.

The first three models in Table I contain no autoregression: The intercept $\alpha^{(v)}$ is either treated as fixed (denoted by 0 in the label of the model) or as fixed region-specific (1) or as IID (2). The estimates for the quadratic trend and the seasonal components are very similar for all three models and thus not included in the table. Including the expected cases as offset causes a varying incidence level in model A0. However, it is not clear whether the heterogeneity of the disease incidence as visible in the map in Figure 1 can thus be captured adequately. Therefore, each region gets its own (fixed) intercept in model A1 which substantially improves the log-likelihood from −10781 to −10431. Finally, model A2 assumes that the incidence levels are IID random effects with estimated variance $\hat{\sigma}_v^2=0.14$.

Table II. Mean scores based on 94×84 one-step-ahead predictions for the meningococcal disease data.

Model	$\overline{\log S}$	(<i>p</i> -Value)	$\overline{RPS}$	(<i>p</i> -Value)	$\overline{SES}$	(<i>p</i> -Value)
<i>No autoregression</i>						
A0	0.5920	(0.0001)	0.2009	(0.0001)	0.3245	(0.0001)
A1	0.5845	(0.0011)	0.1973	(0.0016)	0.3146	(0.0033)
A2	0.5823	(0.0420)	0.1970	(0.0185)	0.3140	(0.0240)
<i>Fixed intercept $\alpha^{(\lambda)}$</i>						
B1	0.5825	(0.0022)	0.1965	(0.0016)	0.3128	(0.0003)
B2	0.5802	(0.2260)	0.1962	(0.2730)	0.3121	(0.0451)
<i>IID random effects $b^{(\lambda)}$</i>						
C2	0.5813	(0.0090)	0.1966	(0.0255)	0.3133	(0.0301)
<i>Fixed intercept $\alpha^{(\lambda)}$, CAR random effects $b^{(v)}$</i>						
D	0.5799		0.1961		0.3119	

Model D is compared with the remaining models, the Monte Carlo *p*-values are based on permutation tests for paired observations (9999 permutations).

Bold indicates the best model, i.e. the model with lowest score.

The predictive quality of the models is assessed through one-step-ahead predictions of the last 7 years. Table II shows the mean scores based on these 94×84 predictions. The results clearly indicate that a single fixed intercept $\alpha^{(\lambda)}$ (A0) yields the worst predictive performance and regional heterogeneity should be taken into account. According to all scores, a formulation with random effects (A2, B2) is to be preferred compared with the corresponding formulation with region-specific fixed effects (A1, B1). Note that in a regression context it is well known that the use of plug-in ML or Least Squares estimates can be suboptimal in prediction problems due to overfitting. Copas [57, 58] suggests shrinkage estimates for normal and logistic regression to improve predictive performance.

Basically, all models including an autoregression perform better than models excluding autoregression with respect to all scores. The autoregressive parameter $\hat{\alpha}^{(\lambda)}$ in those models is estimated to be quite small, $\exp(\hat{\alpha}^{(\lambda)}) \approx 0.08$. Accordingly, the maximum eigenvalue of $\hat{\Lambda}$ is also small. We did not include the quantity ϕ_r , which measures the influence of adjacent regions, into the model, as it seems to have a very small effect and is hardly estimable from the meningococcal disease data set.

Note that a model formulation with region-specific fixed autoregressive parameters $\alpha_r^{(\lambda)}$ cannot be fitted, whereas a formulation with random effects $b^{(\lambda)}$ is possible. Figure 2 displays the estimated IID random effects from model C2. The scatterplot shows that there is a moderate negative correlation between $\hat{b}^{(v)}$ and $\hat{b}^{(\lambda)}$, in accordance with the estimate $\hat{\rho}_{\lambda v} = -0.27$, see Table I. The estimated autoregressive coefficients $\hat{\lambda}_r$ are also mapped. There is considerable heterogeneity across regions with highest values in some northern and western departments. However, no clear spatial pattern can be seen. Although the variance σ_{λ}^2 is estimated fairly large in model C2 with IID autoregressive parameter, mean scores for this model are larger than for the respective model B2 with a single fixed parameter.

The best model with the lowest mean score among models with only fixed or IID random effects is model B2, which includes a single autoregressive parameter $\alpha^{(\lambda)}$ and IID effects $b^{(v)}$ in the endemic component. For comparison, we also fitted a corresponding model (D) with CAR instead of IID effects. This formulation leads to a further (if only minor) reduction of all mean scores. Figure 3 shows the estimated mean for this model in the four selected departments. The mean in each department is separated into two additive components: the light gray area shows the estimated endemic component $\hat{v}_{rt}$ and the dark gray area corresponds to the autoregressive contribution $\exp(\hat{\alpha}^{(\lambda)})y_{r,t-1}$. The incidence is clearly dominated by the endemic component, which is in agreement with the maximum eigenvalue shown in Table I.

To judge the differences between the best model D and all competing models more formally, we carried out permutation tests based on 9999 permutations for paired observations. The corresponding Monte Carlo *p*-values for these tests can be found in Table II. There is a significant difference to models without autoregression. However, there seems to be no significant difference between the CAR and an IID formulation for the random effects in the endemic component according to both logS and RPS.

A comparison with the results obtained by Knorr-Held and Richardson [18] is not straightforward due to substantial differences in the model formulation. Also, Knorr-Held and Richardson [18] have included only data up to 1997 for fitting and have not considered one-step-ahead forecasts for model

Figure 2. Estimated random effects from model C2. Shown is a scatterplot of the estimated random effects in the endemic component $\hat{\mathbf{b}}^{(v)}$ and in the autoregressive component $\hat{\mathbf{b}}^{(\lambda)}$ (left), and a map of the estimated autoregressive parameters $\hat{\lambda}_r$ (right).

Figure 3. Fitted values for model D for meningococcal disease data in selected departments of France.

comparison. Instead, they have used DIC [19] which has selected a model with two autoregressive terms on the number of counts in the same region and the number of counts in neighboring regions, respectively, as the best model. Note that past counts only enter in a dichotomized form (1 if there was at least one case and 0 otherwise) modulated by additional region-specific latent hidden Markov models. Our approach has identified a significant autoregression on past counts in the same region, but not in neighboring regions.

4.2. Influenza in Southern Germany

As a second application we consider the weekly number of laboratory confirmed influenza cases in 140 administrative districts of the two Southern German states Baden-Württemberg and Bavaria for years from 2001 to 2008 [55]. Note that the number of regions is too large to fit models with

Figure 4. Weekly number of influenza cases in the whole of German states Baden-Württemberg and Bavaria (left) and mean yearly incidence per 100 000 habitants in the 140 administrative districts, averaged over the years 2001–2008 (right).

region-specific fixed intercepts. Instead, we apply the proposed methodology with random effects to account for regional heterogeneity.

Figure 4 shows the total number of influenza cases per week in all districts. A clear seasonal pattern with steep increases of the number of cases in winter can be seen. Also shown in Figure 4 is a map of the average yearly incidence per 100 000 habitants over the years 2001–2008. The incidence ranges from 0 to 70 cases and is much higher than for the meningococcal disease data. However, note that the influenza data suffer from substantial underreporting. Only a fraction of all influenza cases is actually recorded since the illness caused by influenza is often too slight to warrant medical attention.

We consider several models that differ depending on whether quantities λ_r , ϕ_r are included in the model or not, and if yes, whether they enter as fixed or random effects. The number of cases is assumed to be negative binomial distributed with overdispersion parameter ψ . All models include random incidence levels $b^{(v)}$, a linear trend, $S=3$ seasonal terms, and population fractions $p_r / \sum_r p_r$, where p_r denotes the population in region r at 31.12.2001 as offset. Note that $S=3$ components are used because the superposition of only one sine and cosine wave could not adequately model the distinct peaks, compare Paul *et al.* [7].

Table III shows the estimated parameters and standard errors for the different models. The overdispersion parameter ψ is always estimated to be larger than 1. There seems to be little variation in the autoregressive coefficient λ_r since the variance $\hat{\sigma}_\lambda^2$ is estimated to be quite small. Consequently, estimation of models including IID effects in the autoregressive component turned out to be more difficult than estimation of the remaining models in Table III. On the other hand, there is considerable variation concerning the ‘neighbor-driven’ coefficient ϕ_r with $\hat{\sigma}_\phi^2 \approx 1$. While the estimated correlation $\hat{\rho}_{\phi_v}$ in models A2, B2, C2, with IID effects in both the neighbor-driven and endemic component is considerable, IID effects in the autoregressive and endemic component are estimated to be nearly uncorrelated in models C0 and C2. Note that IID and CAR effects in the neighbor-driven and endemic component, respectively, are assumed to be uncorrelated in model D. Both quantities λ_r and ϕ_r characterizing the spatio-temporal spread seem to be important for the model fit. The maximum eigenvalue of $\hat{\mathbf{A}}$ is quite large in most models which indicates considerable ‘epidemic’ behavior.

The predictive performance of the models is assessed through one-step-ahead predictions of the last 104 weeks, i.e. the last two years. Mean scores based on these 140×104 predictions are shown in Table IV. Most models yield comparable mean scores, except for models without autoregressive and neighbor-driven component (A0, A1, A2) which have a clearly higher score. According to the mean logarithmic score, model B2 with a fixed autoregressive parameter performs best, closely followed by model C2 with IID autoregressive parameter. Both models contain IID effects in the neighbor-driven and in the endemic component. Replacing the IID effects in the endemic component of model B2 by CAR effects in model D leads to a slightly higher mean logS. It seems that such a spatial effect is not needed in addition to the neighbor-driven component, which already accounts for spatio-temporal

Table III. Analysis of influenza in the 140 administrative districts of Southern Germany.

Model	$\exp(\hat{\alpha}^{(\lambda)})$ (se)	$\exp(\hat{\alpha}^{(\phi)})$ (se)	$\hat{\alpha}^{(v)}$ (se)	$\hat{\psi}$ (se)	$\hat{\sigma}_{\lambda}^2$	$\hat{\sigma}_{\phi}^2$	$\hat{\sigma}_v^2$	$\hat{\rho}_{\lambda v}$	$\hat{\rho}_{\phi v}$	$\hat{\rho}_{\lambda \phi}$	maxEV	$\ell_{\text{pen}} (\ell_{\text{mar}})$
<i>No autoregression, IID random effects $\mathbf{b}^{(v)}$</i>												
A0	—	—	0.52 (0.11)	2.07 (0.05)	—	—	0.79	—	—	—	—	−20178 (−314)
A1	—	0.24 (0.01)	0.23 (0.12)	1.80 (0.05)	—	—	1.14	—	—	—	0.24	−19805 (−299)
A2	—	0.39 (0.04)	0.26 (0.11)	1.51 (0.04)	—	1.11	0.79	—	0.51	—	0.39	−19468 (−435)
<i>Fixed intercept $\alpha^{(\lambda)}$, IID random effects $\mathbf{b}^{(v)}$</i>												
B0	0.49 (0.02)	—	0.42 (0.10)	1.26 (0.04)	—	—	0.48	—	—	—	0.49	−19049 (−256)
B1	0.46 (0.02)	0.11 (0.01)	0.24 (0.11)	1.20 (0.04)	—	—	0.68	—	—	—	0.57	−18895 (−252)
B2	0.41 (0.02)	0.22 (0.02)	0.22 (0.10)	1.10 (0.03)	—	0.96	0.51	—	0.56	—	0.63	−18742 (−343)
<i>IID random effects $\mathbf{b}^{(\lambda)}$, IID random effects $\mathbf{b}^{(v)}$</i>												
C0	0.47 (0.02)	—	0.42 (0.10)	1.22 (0.04)	0.06	—	0.48	0.03	—	—	0.65	−19019 (−301)
C1	0.42 (0.02)	0.11 (0.01)	0.25 (0.11)	1.15 (0.04)	0.09	—	0.67	0.28	—	—	0.67	−18851 (−310)
C2	0.38 (0.02)	0.21 (0.02)	0.22 (0.10)	1.05 (0.03)	0.13	1.02	0.51	0.04	0.53	−0.18	0.75	−18691 (−410)
<i>Fixed intercept $\alpha^{(\lambda)}$, IID random effects $\mathbf{b}^{(\phi)}$, CAR random effects $\mathbf{b}^{(v)}$</i>												
D	0.41 (0.02)	0.22 (0.03)	0.22 (0.06)	1.10 (0.03)	—	1.12	1.66	—	—	—	0.62	−18744 (−368)

Shown are parameter estimates with standard errors in brackets. The maximum eigenvalue of the estimated matrix $\mathbf{A}$ in (3) is denoted by maxEV. All models contain a linear trend and $S=3$ seasonal terms. The numerals in the model labels correspond to 0: no neighbor-driven component, 1: fixed intercept $\alpha^{(\phi)}$, 2: IID random effects $\mathbf{b}^{(\phi)}$.

Table IV. Mean scores based on 140×104 one-step-ahead predictions for the influenza data.

Model	$\overline{\log S}$	(p -Value)	$\overline{\text{RPS}}$	(p -Value)	$\overline{\text{SES}}$	(p -Value)
<i>No autoregression, IID random effects $\mathbf{b}^{(v)}$</i>						
A0	0.5988	(0.0001)	0.5258	(0.0001)	8.1394	(0.0001)
A1	0.5913	(0.0001)	0.5030	(0.0001)	6.7258	(0.0001)
A2	0.5882	(0.0001)	0.4815	(0.0001)	6.2256	(0.0001)
<i>Fixed intercept $\alpha^{(\lambda)}$, IID random effects $\mathbf{b}^{(v)}$</i>						
B0	0.5686	(0.0006)	0.4472	(0.0001)	5.2607	(0.6144)
B1	0.5652	(0.0353)	0.4427	(0.0001)	5.1883	(0.9952)
B2	0.5633		0.4363		5.1878	
<i>IID random effects $\mathbf{b}^{(\lambda)}$, IID random effects $\mathbf{b}^{(v)}$</i>						
C0	0.5685	(0.0018)	0.4461	(0.0001)	5.2682	(0.5107)
C1	0.5648	(0.0830)	0.4414	(0.0007)	5.2182	(0.7227)
C2	0.5635	(0.5979)	0.4346	(0.0438)	5.2088	(0.7399)
<i>Fixed intercept $\alpha^{(\lambda)}$, IID random effects $\mathbf{b}^{(\phi)}$, CAR random effects $\mathbf{b}^{(v)}$</i>						
D	0.5638	(0.1840)	0.4361	(0.5485)	5.1969	(0.4263)

Model B2 is compared with the remaining models, the Monte Carlo p -values are based on permutation tests for paired observations (9999 permutations).

Bold indicates the best model, i.e. the model with lowest score.

dependence. The ranking of models with respect to $\overline{\text{RPS}}$ is similar except for the best two models with interchanged ordering.

Figure 5. Fitted values for model B2 for influenza data in selected administrative districts of Southern Germany.

We compared each model to model B2. Models including both autoregressive and neighbor-driven component have a lower mean score (p -values of the Monte Carlo permutation tests <0.0018 for $\log S$ and equal to 0.0001 for RPS). According to the Monte Carlo p -values from the respective permutation tests there is, however, no clear preference for the form of the parameters (i.e. fixed or random).

Figure 5 shows a plot of the fitted incidence from the best model according to $\log S$, B2, in four selected administrative districts indicated in the map in Figure 4. Light, medium, and dark gray areas correspond to the endemic, autoregressive, and neighbor-driven part of the mean. In some regions, such as LK Kitzingen or SK Stuttgart, the cases in surrounding districts considerable contribute to the mean incidence. The importance of ϕ_r and λ_r for the predictive performance is also clear when looking at successive maps of the weekly number of cases.

5. Discussion

In this paper, we have extended the model for the analysis of multiple time series of infectious disease counts suggested by Held *et al.* [5] and Paul *et al.* [7] to account for different incidence levels or varying disease transmission via possibly correlated random effects. Especially for highly multivariate time series, heterogeneity is very likely to exist and should be accounted for. This is done by including region-specific random effects which are assumed to be either IID Gaussians or follow a CAR model. Random effects in the different components of the mean may also be assumed to be correlated.

Inference is based on methodology for mixed models as discussed e.g. in Kneib and Fahrmeir [14]. Estimates for regression parameters are obtained by numerically optimizing the penalized likelihood and estimates of variance parameters are obtained by optimizing an approximate marginal likelihood in an alternating algorithm.

Extension of classical model choice criteria to mixed models is challenging. There are versions of AIC or BIC which can be used to address certain questions, e.g. which fixed effects to select in models with given random effects structure or whether or not to include a random effect in models with given fixed effects structure. However, these criteria are not suitable for the comparison of several models in any situation. Therefore, we compared models based on one-step-ahead predictions and proper scoring rules which can be used in any situation.

While a formulation where each region gets its own fixed intercept can be applied for the analysis of disease counts in a low to moderate number of regions, this is no longer feasible when the number

of regions is large. For instance, the analysis of the second data set would not be possible with the formulation in Paul *et al.* [7]. The analysis of the meningococcal disease and influenza data showed that the induced shrinkage of region-specific estimates towards the overall mean improves the predictive performance in the presence of heterogeneity.

Appendix A: Inference

In the following we provide further details about the estimation procedure described in Section 2.2 for the model with mean (2) specified as in Equations (4)–(6), and covariance matrix Σ for the IID random effects given by (8). To ensure positive definiteness of the covariance matrix, we use a parameterization based on the Cholesky decomposition of Ω in spherical coordinates [35]. Note that the diagonal elements of the lower triangular Cholesky matrix must be positive and the off-diagonal elements must be in $(0, \pi)$ to ensure uniqueness. These restrictions can be guaranteed by using suitable transformations (see e.g. [59]). The resulting parameterization is then given by

$$\Omega(\theta) = \begin{pmatrix} \exp(2s_1) & & \\ \frac{r_1 \exp(s_2 + s_1)}{\sqrt{r_1^2 + 1}} & \exp(2s_2) & \\ \frac{r_2 \exp(s_3 + s_1)}{\sqrt{r_2^2 + 1}} & \frac{(r_1 r_2 \sqrt{r_3^2 + 1} + r_3) \exp(s_3 + s_2)}{\sqrt{r_1^2 + 1} \sqrt{r_2^2 + 1} \sqrt{r_3^2 + 1}} & \exp(2s_3) \end{pmatrix},$$

where the vector $\theta = (s_1, s_2, s_3, r_1, r_2, r_3)^\top$ contains the unknown variance parameters which can take values on the whole real line. For the sake of clarity only the lower triangular part of the symmetric matrix $\Omega(\theta)$ is displayed. Note that the diagonal elements correspond to the variances of the random effects in the autoregressive (4), neighbor-driven (5), and endemic component (6). For instance, we have $\sigma_\lambda^2 = \exp(2s_1)$. Similarly, we obtain the correlation parameters, e.g. $\rho_{\lambda\phi} = r_1 / \sqrt{r_1^2 + 1}$.

A.1. Estimation of regression parameters

First, let i and j denote any element of the stacked vector of fixed and random effects. The score vector with first derivatives of (9) with respect to β and b is given by

$$s_{\text{pen}}(\beta, b; \Sigma) = \begin{pmatrix} \frac{\partial \ell_{\text{pen}}(\beta, b)}{\partial \beta} \\ \frac{\partial \ell_{\text{pen}}(\beta, b)}{\partial b} \end{pmatrix} = \begin{pmatrix} s(\beta) \\ s(b) - \Sigma^{-1}b \end{pmatrix},$$

where the elements of the unpenalized score vector are given by

$$s(i) = \sum_{r,t} \frac{y_{rt}}{\mu_{rt}} \frac{\partial \mu_{rt}}{\partial i} - \frac{\partial \mu_{rt}}{\partial i}.$$

Analogously, the observed Fisher information matrix is partitioned as

$$\mathbf{F}_{\text{pen}}(\beta, b; \Sigma) = \begin{pmatrix} -\frac{\partial^2 \ell_{\text{pen}}(\beta, b)}{\partial \beta \partial \beta^\top} & -\frac{\partial^2 \ell_{\text{pen}}(\beta, b)}{\partial \beta \partial b^\top} \\ -\frac{\partial^2 \ell_{\text{pen}}(\beta, b)}{\partial b \partial \beta^\top} & -\frac{\partial^2 \ell_{\text{pen}}(\beta, b)}{\partial b \partial b^\top} \end{pmatrix} = \begin{pmatrix} \mathbf{F}[\beta\beta] & \mathbf{F}[\beta b] \\ \mathbf{F}[b\beta] & \mathbf{F}[bb] + \Sigma^{-1} \end{pmatrix}, \quad (\text{A1})$$

where $\mathbf{F}[ij]$ denotes the block of the unpenalized Fisher information matrix corresponding to the parameter vectors i and j , and has elements

$$\mathbf{F}[ij] = -\sum_{r,t} \left\{ -\frac{y_{rt}}{\mu_{rt}^2} \frac{\partial \mu_{rt}}{\partial i} \frac{\partial \mu_{rt}}{\partial j} + \frac{y_{rt}}{\mu_{rt}} \frac{\partial^2 \mu_{rt}}{\partial i \partial j} - \frac{\partial^2 \mu_{rt}}{\partial i \partial j} \right\}.$$

In the following, let $\xi^{(\lambda)}$ contain all parameters in the autoregressive component (4), i.e. $\alpha^{(\lambda)}, \mathbf{b}^{(\lambda)}$, and let $\mathbf{x}_{rt}^{(\lambda)}$ denote the corresponding design vector in region r at time t , i.e. $\mathbf{x}_{rt}^{(\lambda)} = (1, \mathbf{u}_r)^\top$ where the r th element of the R -dimensional vector $\mathbf{u}_r$ is 1 and 0 otherwise. The vectors $\xi^{(\phi)}, \mathbf{x}_{rt}^{(\phi)}$ and $\xi^{(v)}, \mathbf{x}_{rt}^{(v)}$ are defined in analogy for components (5) and (6), respectively. First and second partial derivatives of μ_{rt} are then given by

$$\frac{\partial \mu_{rt}}{\partial \xi^{(\lambda)}} = \lambda_{rt} y_{r,t-1} \mathbf{x}_{rt}^{(\lambda)}, \quad \frac{\partial \mu_{rt}}{\partial \xi^{(\phi)}} = \phi_{rt} \sum_{q \neq r} w_{qr} y_{q,t-1} \mathbf{x}_{rt}^{(\phi)}, \quad \frac{\partial \mu_{rt}}{\partial \xi^{(v)}} = v_{rt} \mathbf{x}_{rt}^{(v)}$$

and

$$\begin{aligned} \frac{\partial^2 \mu_{rt}}{\partial \xi^{(\lambda)} \partial \xi^{(\lambda)\top}} &= \lambda_{rt} y_{r,t-1} \mathbf{x}_{rt}^{(\lambda)} \mathbf{x}_{rt}^{(\lambda)\top}, \quad \frac{\partial^2 \mu_{rt}}{\partial \xi^{(\phi)} \partial \xi^{(\phi)\top}} = \phi_{rt} \sum_{q \neq r} w_{qr} y_{q,t-1} \mathbf{x}_{rt}^{(\phi)} \mathbf{x}_{rt}^{(\phi)\top}, \\ \frac{\partial^2 \mu_{rt}}{\partial \xi^{(v)} \partial \xi^{(v)\top}} &= v_{rt} \mathbf{x}_{rt}^{(v)} \mathbf{x}_{rt}^{(v)\top}. \end{aligned}$$

Second partial derivatives of μ_{rt} with respect to parameters belonging to different components are zero.

A.2. Estimation of variance parameters

First note that

$$\log |\Sigma| = \log (|\Omega(\theta) \otimes \mathbf{I}|) = \log (|\Omega(\theta)|^R |\mathbf{I}|^3) = 2R \left(\sum_{i=1}^3 s_i - \frac{1}{2} \sum_{i=1}^3 \log(r_i^2 + 1) \right).$$

Let k and l denote any element of the vector with variance parameters θ . Furthermore, the dependence of the observed Fisher information matrix (A1) on both regression and variance parameters will be suppressed in the following for notational convenience. The score vector $\mathbf{s}_{\text{marg}}(\Sigma)$ with first derivatives of (11) with respect to θ then has elements

$$\frac{\partial \ell_{\text{marg}}(\Sigma)}{\partial k} = -R \sum_{i=1}^3 \left(1(k=s_i) - \frac{r_i}{r_i^2 + 1} 1(k=r_i) \right) + \frac{1}{2} \mathbf{b}^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial k} \Sigma^{-1} \mathbf{b} - \frac{1}{2} \text{tr} \left(\mathbf{F}_{\text{pen}}^{-1} \frac{\partial \mathbf{F}_{\text{pen}}}{\partial k} \right)$$

where $\text{tr}(\cdot)$ denotes the trace of a matrix. The elements of the observed Fisher information matrix $\mathbf{F}_{\text{marg}}(\Sigma)$ are given as

$$\begin{aligned} \frac{\partial^2 \ell_{\text{marg}}(\Sigma)}{\partial k \partial l} &= -R \sum_{i=1}^3 \left(\frac{r_i^2 - 1}{(r_i^2 + 1)^2} 1(k=l=r_i) \right) - \frac{1}{2} \mathbf{b}^\top \Sigma^{-1} \left(-\frac{\partial^2 \Sigma}{\partial k \partial l} + \frac{\partial \Sigma}{\partial k} \Sigma^{-1} \frac{\partial \Sigma}{\partial l} + \frac{\partial \Sigma}{\partial l} \Sigma^{-1} \frac{\partial \Sigma}{\partial k} \right) \Sigma^{-1} \mathbf{b} \\ &\quad - \frac{1}{2} \text{tr} \left(-\mathbf{F}_{\text{pen}}^{-1} \frac{\partial \mathbf{F}_{\text{pen}}}{\partial l} \mathbf{F}_{\text{pen}}^{-1} \frac{\partial \mathbf{F}_{\text{pen}}}{\partial k} + \mathbf{F}_{\text{pen}}^{-1} \frac{\partial^2 \mathbf{F}_{\text{pen}}}{\partial k \partial l} \right). \end{aligned}$$

The derivation of first and second derivatives of $\Omega(\theta)$ with respect to θ , and thus Σ and $\mathbf{F}_{\text{pen}}$, is straightforward and omitted here. Note that for uncorrelated random effects, the block diagonal structure of the covariance matrix Σ can be exploited to obtain simplified expressions for the score vector $\mathbf{s}_{\text{marg}}(\Sigma)$ and the Fisher information matrix $\mathbf{F}_{\text{marg}}(\Sigma)$.

A.3. Implementation

In applications, the likelihood functions described in Section 2 might be flat and contorted due to the additive decomposition of the mean. This complicates maximization. In principle, estimates both regression and variance components can be obtained using the iterative Newton algorithm. At each iteration, the score function is approximated by a local quadratic model and solved for the root, i.e. the maximum of the likelihood. Close to the maximum, the Fisher information matrix must be positive definite and the Newton method guarantees fast convergence [60]. However, far away from the maximum the Fisher information matrix might not be positive definite. If initial values are poor, the Newton algorithm may fail.

There are several modifications to obtain a globally convergent Newton algorithm without sacrificing the good local convergence properties, see e.g. [61, Chapters 5, 6]. For instance, the Fisher information matrix can be modified such that it is always positive definite. Convergence from poor initial values can be improved by using a trust-region approach [61, Chapter 6.4]. The trust-region approach provides at each iteration a region in which the quadratic approximation can be trusted to adequately model the score function. The root of the approximation is picked as new iterate and the goodness of the approximation is evaluated. If the approximation is not good enough, the region is shrunk until an acceptable new iterate is found.

The optimization problem gets harder the more parameters are involved and the choice of good initial values is especially important. In non-linear models, it is often advisable to use several initial values and select estimates resulting in the highest likelihood should the optimization algorithm lead to different results (e.g. [7]). In the current context; however, it is not clear how to deal with differing results for the alternating algorithm from Section 2.2. The penalized likelihood cannot be used for deciding between two sets of parameter estimates because it depends on the variance parameters. Thus, results with larger variance parameters tend to be preferred.

The estimation procedure is currently incorporated in the development version of the R package *surveillance* [31] available from <http://surveillance.r-forge.r-project.org/>. We use the R function `nlminb` as default optimizer which can either use analytical second derivatives or numerical approximations to the Hessian. Although a single computation of the analytical Fisher information matrix is more expensive than the use of numerical approximations, the use of the analytical Fisher information matrix makes convergence faster in the long run and is to be preferred. The function `nlminb` also implements a trust region approach which enhances convergence. In all applications considered, the algorithm always lead to the same solution after convergence when using different initial values.

Acknowledgements

We thank the reviewers for helpful comments and suggestions. Financial support by the Swiss National Science Foundation is gratefully acknowledged.

References

1. Giesecke J. *Modern Infectious Disease Epidemiology* (2nd edn). Hodder Arnold: London, 2002.
2. Zeger SL, Qaqish B. Markov regression models for time series: a quasi-likelihood approach. *Biometrics* 1988; **44**(4):1019–1031.
3. Benjamin MA, Rigby RA, Stasinopoulos DM. Generalized autoregressive moving average models. *Journal of the American Statistical Association* 2003; **98**(461):214–223.
4. Fokianos K, Rahbek A, Tjøstheim D. Poisson autoregression. *Journal of the American Statistical Association* 2009; **104**(488):1430–1439.
5. Held L, Höhle M, Hofmann M. A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling* 2005; **5**(3):187–199.
6. Becker NG, Britton T. Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society, Series B* 1999; **61**(2):287–307.
7. Paul M, Held L, Toschke AM. Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine* 2008; **27**(29):6250–6267.
8. Li N, Qian GQ, Huggins R. A random effects model for diseases with heterogeneous rates of infection. *Journal of Statistical Planning and Inference* 2003; **116**(1):317–332.
9. Davis X, Waller L, Haber M. Estimating vaccine efficacy from outbreak size household data in the presence of heterogeneous transmission probabilities. *Journal of Biopharmaceutical Statistics* 2006; **16**(4):499–516.
10. Zeger SL, Karim MR. Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* 1991; **86**(413):79–86.
11. Lin XH, Zhang DW. Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B* 1999; **61**(2):381–400.
12. Breslow NE. Whither PQL? In *Proceedings of the Second Seattle Symposium in Biostatistics—Analysis of Correlated Data*, Lin, DY, Heagerty PJ (eds), Lecture Notes in Statistics, vol. 179. Springer: New York, 2004; 1–22.
13. Tuerlinckx F, Rijmen F, Verbeke G, De Boeck P. Statistical inference in generalized linear mixed models: a review. *British Journal of Mathematical and Statistical Psychology* 2006; **59**(2):225–255.
14. Kneib T, Fahrmeir L. A mixed model approach for geospatial hazard regression. *Scandinavian Journal of Statistics* 2007; **34**(1):207–228.
15. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**(421):9–25.

16. Held L, Hofmann M, Höhle M, Schmid V. A two-component model for counts of infectious diseases. *Biostatistics* 2006; **7**(3):422–437.
17. Hofmann M. Statistical models for infectious disease surveillance counts. *Ph.D. Thesis*, Ludwig-Maximilians-Universität München, 2007. Available from: <http://edoc.ub.uni-muenchen.de/6601/>.
18. Knorr-Held L, Richardson S. A hierarchical model for space–time surveillance data on meningococcal disease incidence. *Journal of the Royal Statistical Society, Series C* 2003; **52**(2):169–183.
19. Spiegelhalter D, Best N, Carlin B, van der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B* 2002; **64**:583–616.
20. Plummer M. Penalized loss functions for Bayesian model comparison. *Biostatistics* 2008; **9**(3):523–539.
21. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference. A Practical Information—Theoretic Approach* (2nd edn). Springer: New York, 2002.
22. Vaida F, Blanchard S. Conditional Akaike information for mixed-effects models. *Biometrika* 2005; **92**(2):351–370.
23. Greven S, Kneib T. On the behavior of marginal and conditional Akaike information criteria in linear mixed models. *Biometrika* 2010; **97**(4):773–789.
24. Liang H, Wu HL, Zou GH. A note on conditional AIC for linear mixed-effects models. *Biometrika* 2008; **95**(3):773–778.
25. Pauler DK. The Schwarz criterion and related methods for normal linear models. *Biometrika* 1998; **85**(1):13–27.
26. Jiang JM, Rao JS, Gu ZH, Nguyen T. Fence methods for mixed model selection. *Annals of Statistics* 2008; **36**(4):1669–1692.
27. Dawid AP. Statistical theory. The prequential approach. *Journal of the Royal Statistical Society, Series A* 1984; **147**(2):278–292.
28. Smith JQ. Diagnostic checks of non-standard time series models. *Journal of Forecasting* 1985; **4**(3):283–291.
29. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 2007; **102**(477):359–378.
30. Doucet A, de Freitas N, Gordon N. *Sequential Monte Carlo Methods in Practice* (2nd edn), Statistics for Engineering and Information Science. Springer: Berlin, 2001.
31. Höhle M. Surveillance: an R package for the monitoring of infectious diseases. *Computational Statistics* 2007; **22**(4):571–582.
32. Haccou P, Jagers P, Vatutin VA. *Branching Processes: Variation, Growth, and Extinction of Populations*. Cambridge University Press: Cambridge, 2007.
33. Farrington CP, Kanaan MN, Gay NJ. Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics* 2003; **4**(2):279–295.
34. Diggle PJ. *Time Series. A Biostatistical Introduction*. Oxford University Press: Oxford, 1990.
35. Pinheiro JC, Bates DM. Unconstrained parameterizations for variance–covariance matrices. *Statistics and Computing* 1996; **6**(3):289–296.
36. Besag J, York J, Mollié A. Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 1991; **43**(1):1–20.
37. Rue H, Held L. *Gaussian Markov Random Fields*, vol. 104. Chapman and Hall/CRC Press: London, 2005.
38. Farrington CP, Andrews NJ, Beale AD, Catchpole MA. A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society, Series A* 1996; **159**:547–563.
39. Schall R. Estimation in generalized linear models with random effects. *Biometrika* 1991; **78**(4):719–727.
40. Ogata Y. Evaluation of spatial Bayesian models—two computational methods. *Journal of Statistical Planning and Inference* 1996; **51**(1):1–18.
41. Cai T, Hyndman RJ, Wand M. Mixed model-based hazard estimation. *Journal of Computational and Graphical Statistics* 2002; **11**(4):784–798.
42. Tierney L, Kadane JB. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* 1986; **81**(393):82–86.
43. Liu Q, Pierce DA. A note on Gauss–Hermite quadrature. *Biometrika* 1994; **81**(3):624–629.
44. Breslow NE, Lin XH. Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika* 1995; **82**(1):81–91.
45. Lin XH, Breslow NE. Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* 1996; **91**(435):1007–1016.
46. Kneib T. Mixed model based inference in structured additive regression. *Ph.D. Thesis*, LMU München: Fakultät für Mathematik, Informatik und Statistik, 2006. Available from: <http://edoc.ub.uni-muenchen.de/5011/>.
47. Harville DA. *Matrix Algebra from a Statistician's Perspective*. Springer: New York, 1997.
48. Czado C, Gneiting T, Held L. Predictive model assessment for count data. *Biometrics* 2009; **65**(4):1254–1261.
49. Good IJ. Rational decisions. *Journal of the Royal Statistical Society, Series B* 1952; **14**(1):107–114.
50. Epstein ES. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology* 1969; **8**(6):985–987.
51. Diebold FX, Mariano RS. Comparing predictive accuracy. *Journal of Business and Economic Statistics* 1995; **13**(3):253–263.
52. Jolliffe IT. Uncertainty and inference for verification measures. *Weather and Forecasting* 2007; **22**(3):637–650.
53. Ludbrook J, Dudley H. Why permutation tests are superior to *t* and *F* tests in biomedical research. *American Statistician* 1998; **52**(2):127–132.
54. Brix A, Diggle PJ. Spatiotemporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society, Series B* 2001; **63**(4):823–841.
55. Robert Koch Institute. SurvStat. Available from: <http://www3.rki.de/SurvStat> (March 2009).
56. Denison DGT, Holmes CC, Mallick BK, Smith AFM. *Bayesian Methods for Nonlinear Classification and Regression*. Wiley Series in Probability and Statistics. Wiley: Chichester, 2002.

57. Copas JB. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society, Series B* 1983; **45**(3):311–354.
58. Copas JB. Using regression models for prediction: shrinkage and regression to the mean. *Statistical Methods in Medical Research* 1997; **6**(2):167–183.
59. Rapisarda F, Brigo D, Mercurio F. Parameterizing correlations: a geometric interpretation. *IMA Journal of Management Mathematics* 2007; **18**(1):55–73.
60. Thisted RA. *Elements of Statistical Computing. Numerical Computation*. Chapman and Hall: New York, 1988.
61. Dennis JE, Schnabel RB. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial and Applied Mathematics: Philadelphia, 1996.