Applications of Mathematics

Marek Jiruše; Josef Machek; Viktor Beneš; Petr Zeman

A Bayesian estimate of the risk of tick-borne diseases

Applications of Mathematics, Vol. 49 (2004), No. 5, 389–404

Persistent URL: http://dml.cz/dmlcz/134575

49 (2004) APPLICATIONS OF MATHEMATICS No. 5, 389–404

A BAYESIAN ESTIMATE OF THE RISK OF TICK-BORNE DISEASES*

MAREK JIRUŠE, JOSEF MACHEK, VIKTOR BENEŠ, and PETR ZEMAN, Praha

(Received March 22, 2002, in revised version June 14, 2002)

Abstract. The paper considers the problem of estimating the risk of a tick-borne disease in a given region. A large set of epidemiological data is evaluated, including the point pattern of collected cases, the population map and covariates, i.e. explanatory variables of geographical nature, obtained from GIS.

The methodology covers the choice of those covariates which influence the risk of infection most. Generalized linear models are used and AIC criterion yields the decision. Further, an empirical Bayesian approach is used to estimate the parameters of the risk model. Statistical properties of the estimators are investigated. Finally, a comparison with earlier results is discussed from the point of view of statistical disease mapping.

Keywords: Bayesian estimation, generalized linear model, epidemiological data, statistical properties

MSC 2000: 62G05

1. INTRODUCTION

Statistical disease mapping aims to characterize the spatial variation of cases of a disease and to study connections with given covariates. Most analyses published in the literature are based on the area-level approach which means that cases are aggregated in subregions. If a point pattern of cases is available, a planar point process modelling is another possible approach. In Diggle [3] a survey of both approaches is given.

We are interested in estimating the risk that a person gets infected by a tick-borne disease at a specific location, which is the task usually dealt with by epidemiologists

* This work was supported by the grant No. 201/98/0090 of the Grant Agency of the Czech Republic, and by the project MSM 113200008.

and medical practitioners making decision on prophylactic measures in endemic areas (Zeman [9]). Suppose that the incidence of the disease has been registered for a representative period, and the total human population exposed is known. Further, assume that the area of interest is divided into n regions which may be either administrative or deliberately designed for the purposes of this study. In this area-level approach, for each region an average risk is enquired disregarding individual times of exposure to the risk factor.

Let the number of people living in the ith region be N_i. If the probabilities of being infected were the same for all of them, say equal to π_i, then the observed proportion of the infected $p_i = O_i/N_i$ would be an unbiased estimate of the probability π_i, the risk of infection. The reality, however, is more complex. Every inhabitant of the region has a different exposure intensity and hence a different probability of being infected and it is the average value of these probabilities which should be called the risk of infection for the region, say r_i. The approach used in the present paper is to consider this parameter as a random variable having some prior distribution that depends on a set of variables which describe the environmental conditions. This approach belongs to a class of empirical Bayesian approaches widely used in epidemiology, see e.g. Mollie and Richardson [6].

In this paper first the dataset is described and the process of selection of explanatory variables is discussed in detail. Using the statistical model the desired risk estimators are obtained. The results are compared with the earlier work of Mašata [4] based on a different Bayesian approach.

2. Description of the data

The whole area of the Central Bohemia (CB) was divided into 141 small squares, see Fig. 1. The hole in the middle of Fig. 1 corresponds to the capital Prague which is not a part of CB administratively and its inhabitants are excluded from the analysis. For each square we know, cf. Zeman [9]

- O ... the number of persons infected by tick-borne encephalitis reported in the years 1971–1993 (446 cases altogether)
- N ... the number of inhabitants living in each square (total about 1.1 million)
- *Overlap* ... the proportion of the square area which lies within CB. There are 54 squares partly outside of the CB, we call them border squares thereafter.

The following data that provide additional information about the individual regions will be called explanatory variables (related to each square area):

- $X2, X3, X4$ ... the proportion of the area of coniferous, mixed, and deciduous forest, respectively.

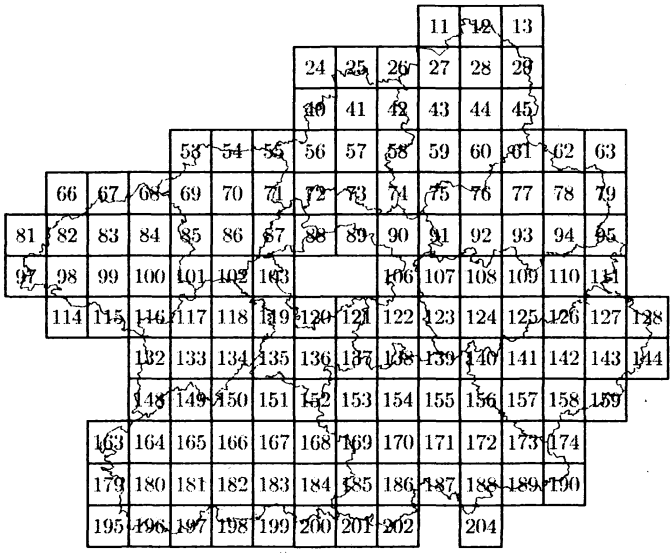

Figure 1. The area of the Central Bohemia divided into 141 squares sized 10×10 km. The administrative subdivision into districts (see the boundaries) is too coarse for the analysis.

- $V1, V2, V3, V4$ ... the proportion of the area having elevation < 300 m, 300–500 m, 500–700 m, and 700–900 m, respectively.
- $L1, L2, L3, L4, L5, L6$ ... the proportion of the area covered by forests of individual area 1–10 ha, 10–50 ha, 50–150 ha, 150–300 ha, 300–600 ha, and > 600 ha, respectively.

The original elevation model was obtained from the Institute of Military Topography, Dobruška, all the other explanatory variables were derived from the satellite images of LANDSAT-5 MSS with resolution power of 80×80 m.

3. Statistical methods

3.1. Empirical Bayesian approach

Suppose that the risk r_i in the ith square is a random variable with some prior distribution. We assume the exponential prior distribution with the mean $E(r_i) = \pi_i$. This is partly justified by the argument that smaller risk values are more probable than larger ones.

Assume that the number of infected persons O_i can be approximated by the Poisson distribution. The posterior distribution of the risk $r_i|O_i$ has gamma distribution

with parameters

$$\eta = O + 1 \quad \text{and} \quad \sigma = \frac{1}{N + 1/\pi}, \tag{1}$$

i.e. the posterior mean is

$$E(r_i|O_i) = \frac{O_i + 1}{N_i + 1/\pi_i}. \tag{2}$$

We will use this posterior mean for the risk quantification. In (2), the trivial estimate of the risk is corrected by the value π which is a function of explanatory variables.

3.2. Optimizing the set of explanatory variables

Here we assume again that the response variable O is approximated by the Poisson distribution with the mean λ. Suppose that the number of infected persons O after appropriate transformation depends on the explanatory variables and the relationship is linear

$$\lambda = f(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p). \tag{3}$$

At a first glance it may seem surprising that the number N_i of persons exposed to the risk does not appear explicitly in the function (3). It should be born in mind, however, that N_i enters the result in another way, see formula (2).

The suitable function and the estimation of the parameters can be found by the theory of GLM models, see McCullagh and Nelder [5]. GLM models extend the classical regression analysis to the case of the response variable coming from the distribution of exponential family. This is in many cases more natural than the assumption of normality. The distribution of exponential family can be written in the form

$$g(y;\theta,\varphi) = \exp\left\{\frac{(y\theta - b(\theta))w}{\varphi} + c(y,\varphi,w)\right\} \tag{4}$$

where $b(\cdot)$, $c(\cdot)$ are specific functions depending on this distribution, θ is the so-called canonical parameter, φ is the dispersion parameter, and w is a known weight.

The canonical parameter θ contains the 'natural' link function that yields the transformation of the mean of the response variable into the linear relationship with the explanatory variables. Another type of the link function can be chosen than the natural one. The use of the natural link function gives to the estimates better

theoretical properties (existence and uniqueness). The probability function for the Poisson distribution can be written in the form (4) as

$$P(O = o) = \exp\Big\{o \underbrace{\log \lambda}_{\theta} - \underbrace{\lambda}_{b(\theta)} \underbrace{- \log o!}_{c(o,\varphi)}\Big\}, \tag{5}$$

where $\theta = \log \lambda$ is the natural link function. Hence the function f in (3) is

$$\lambda = \mathrm{e}^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}. \tag{6}$$

The dispersion parameter for the Poisson distribution is $\varphi = 1$. For the mean and variance it holds

$$E(O) = b'(\theta) = \lambda, \quad \mathrm{var}(O) = b''(\theta)\varphi = \lambda. \tag{7}$$

The estimation of parameters is based on the maximum likelihood method when substituting (6) for the parameter λ into the likelihood function. The estimates are obtained iteratively using the IWLS (iterative weighted least squares regression) described in McCullagh and Nelder [5].

In the classical linear regression, the residual sum of squares is a measure of the goodness of fit, in GLM models residual deviance D^R is used instead (in the following bold letters are used for vectors, e.g. $\boldsymbol{o} = (o_i)_{i=1}^n$),

$$D^R = 2l(\boldsymbol{o}) - 2l(\hat{\boldsymbol{\lambda}}, \boldsymbol{o}). \tag{8}$$

Here $l(\boldsymbol{o})$ is the maximum likelihood function obtained by the absolute fit of observed data with the model, that is $\lambda = o$. Similary $l(\hat{\boldsymbol{\lambda}}, \boldsymbol{o})$ is the likelihood function with substituted ML estimates. D^R has approximately χ^2_{n-q} distribution, where q is the number of parameters. The residual deviance D^R is also a relative measure to compare two models. Denote $D^0 = 2l(\boldsymbol{o}) - 2l^0$ the null deviance where l^0 is the null model of likelihood function, i.e. no relationship is assumed and the predicted values of $\hat{\boldsymbol{\lambda}}$ are substituted by the overall mean $\hat{\lambda} = \sum_{i=1}^{n} o_i/n$.

The fit of our model can be quantified relatively when we substract the residual deviance D^R from null deviance D^0,

$$D = D^0 - D^R = 2l(\hat{\boldsymbol{\lambda}}, \boldsymbol{o}) - 2l^0; \tag{9}$$

D has approximately χ^2_1 distribution.

Accordingly, the contribution of explanatory variables added into the model can be tested. For comparison between the model G and F the deviance is obtained as

$D(\hat{\boldsymbol{\lambda}}^{\boldsymbol{G}}, \hat{\boldsymbol{\lambda}}^{\boldsymbol{F}}, \boldsymbol{o}) = 2l(\hat{\boldsymbol{\lambda}}^{\boldsymbol{G}}, \boldsymbol{o}) - 2l(\hat{\boldsymbol{\lambda}}^{\boldsymbol{F}}, \boldsymbol{o})$ where $D(\hat{\boldsymbol{\lambda}}^{\boldsymbol{G}}, \hat{\boldsymbol{\lambda}}^{\boldsymbol{F}}, \boldsymbol{o}) \sim \chi^2_{t-s}$, t is the number of parameters of the model G and s is the number of parameters of the model F. In classical regression analysis, the residual sum of squares is reduced by adding any explanatory variable. The same applies for residual deviance. The AIC criterion (Akaike Information Criterion) takes into account the number of estimated parameters

$$\text{AIC} = D^R + 2p\hat{\varphi} \tag{10}$$

where p is the number of parameters, and $\hat{\varphi}$ is the estimate of the dispersion parameter from the sample

$$\hat{\varphi} = \frac{1}{n-p} \sum \frac{(o_i - \hat{\lambda})^2}{\hat{\lambda}}. \tag{11}$$

The criterion for the choice among all possible models is the minimum of AIC.

3.3. Overdispersion

For the Poisson distributed response variable, the variance (7) and the dispersion parameter $\varphi = 1$ are given. High values of residual deviance, which is here the measure of the fit, can cause the rejection of the model. This phenomenon is called 'overdispersion', which means that the model has higher variability than assumed. The indicator that the overdispersion is present in the model can be also the estimated dispersion parameter φ. A value greater than 1 makes the variance of the model for Poisson data greater than in (7). To avoid the problem of overdispersed data we can replace the assumption of Poisson distributed response variable by negative binomial distribution which is more flexible due to an additional parameter τ. More details are described in Venables and Ripley [8]. The probability function of the negative binomial distribution in the form of (4) can be written as

$$P(O = o) = \exp\left\{ o \underbrace{\log\left(\frac{\lambda}{\tau + \lambda}\right)}_{\theta} - \tau \underbrace{\log\left(\frac{\tau + \lambda}{\tau}\right)}_{b(\theta)} + c(o, \theta) \right\}. \tag{12}$$

In (12), the natural link function is θ, and the dispersion parameter is $\varphi = 1$. It holds $E(O) = \lambda$, $\text{var}(O) = \lambda + \lambda^2/\tau$, so the variance is greater than the mean, which may be a better reflection of reality.

4. Numerical results

Here we deal with the squares where we have complete (or almost complete) information, i.e. with the squares in Fig. 1 having the overlap with the CB more than 90 %. Since it is assumed that the explanatory variables affect the response variable by a linear relationship after transformation of the response variable, the simple scatterplots between response and explanatory variables can be the first indicators of what the dependency might be.

We distinguish between the random variable O_i and its observed value o_i. In Fig. 2 the relationship between the number of infected persons o_i and the variables $X3$ and $X4$ is obvious. The variable $X2$ does not show any dependency.

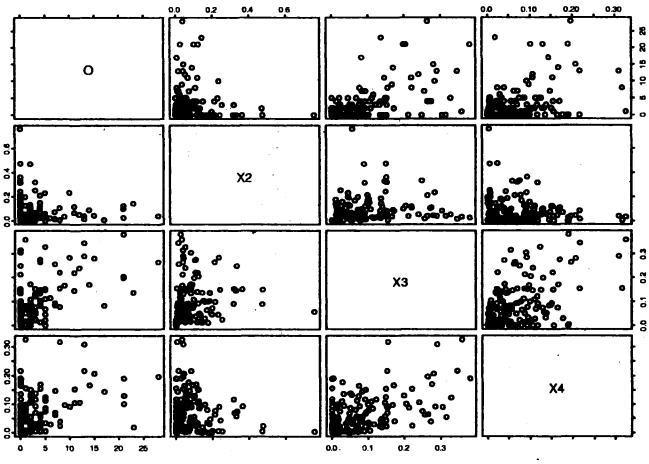

Figure 2. Scatterplots of variables o, $X2$–$X4$.

In Fig. 3, the variable $V2$ shows the best correlation with the number of infected, and so do the variables $L2$ and $L3$ in Fig. 4. The simple correlation coefficients are shown in the Tabs. 1a, 1b.

	$X2$	$X3$	$X4$	$V1$	$V2$	$V3$	$V4$
o	−0.02	0.56	0.40	−0.40	0.43	0.03	−0.08

Table 1a. Correlation coefficients between variables o and X, V, respectively.

	$L1$	$L2$	$L3$	$L4$	$L5$	$L6$
o	0.37	0.62	0.61	0.24	−0.02	−0.07

Table 1b. Correlation coefficients between variables o and L.

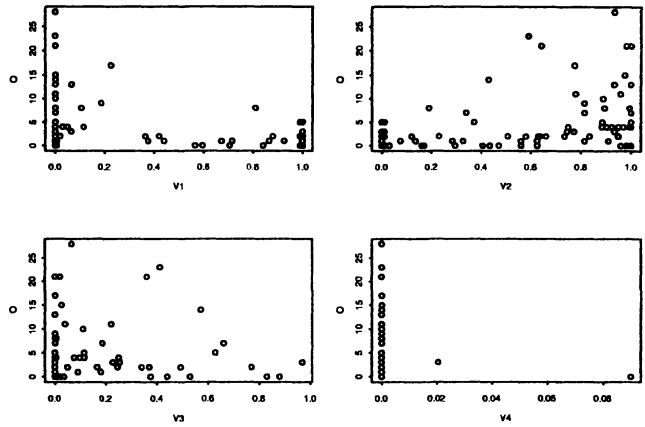

Figure 3. Scatterplots of variables o, $V1$–$V4$.

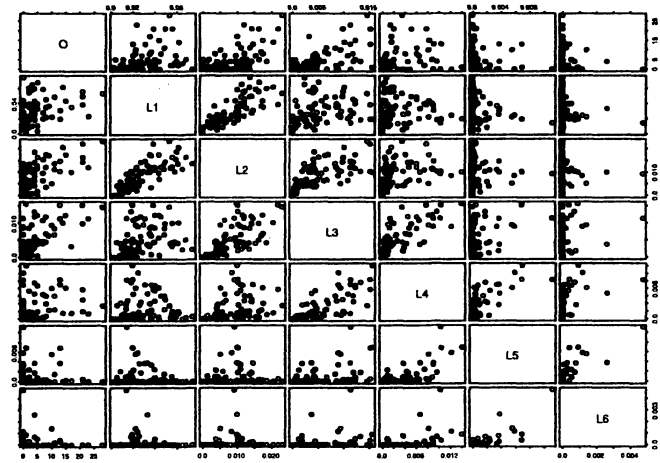

Figure 4. Scatterplots of variables o, $L1$–$L6$.

4.1. The choice of the model

In the first step, we have to find an appropriate model for the distribution of the response variable and the form of the link function. We deal with the variable $X3$ at first, which exhibited positive correlation with the number of infected, and assume the number of infected in the square area to be Poisson distributed. The estimation of the parameters using the natural link function coming from (5) has been carried out in the programming language S supported by the statistical software S-PLUS

$$\log \hat{\lambda}_i = 0.62 + 6.14(X3)_i.$$

The residual deviance exceeded the quantile of its approximative distribution $D^R = 388.3 > \chi^2_{0.95;85} = 107.5$. The estimated dispersion parameter from (11) is $\hat{\varphi} = 5$. The data show higher variability than expected by Poisson model. We try to estimate the parameters from the likelihood function computed for negative binomial distribution of the response variable. The program does not use the natural link function as in (12) and works with the same link function as in the Poisson model

$$\log \hat{\lambda}_i = 0.40 + 7.73(X3)_i, \quad \hat{\tau} = 1.01, \quad \hat{\varphi} = 1.02.$$

This time the residual deviance did not exceed the quantile of its approximative distribution $D^R = 96.6 < \chi^2_{0.95;85} = 107.5$ and the estimation of the dispersion parameter corresponds to its expected value 1. The goodness of fit with the model using the variable $X3$ is confirmed since in (9) it holds

$$D = D^0 - D^R = 136.0 - 96.6 = 39.4 > 3.84 = \chi^2_{0.95;1}.$$

The plot of the observed values o_i against the values $\hat{\lambda}_i$ predicted by the model with the explanatory variable $X3$ shows the goodness of fit in Fig. 5.

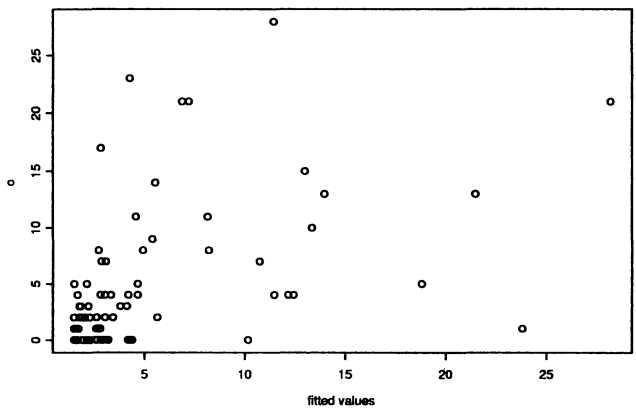

Figure 5. Observed cases o_i plotted against the fitted $\hat{\lambda}_i$.

The fitted values should in the ideal case lie on the diagonal. The correlation coefficient is $\text{cor}(o_i, \hat{\lambda}_i) = 0.49$. There is still a lack of fit using only the variable $X3$ and we keep looking for a set of available explanatory variables to reach a better fit of the data. The model for the negative binomial distribution of the response variable O is used for that. The estimate $\hat{\tau} \doteq 1$ is used fixed for better comparison among the models with other explanatory variables.

4.2. The choice of the set of the explanatory variables

In the first column of Tab. 2, there are six explanatory variables that are included into the model separately and ordered by the size of the residual deviance R^D. In the third column of Tab. 2 there is the value of the correlation coefficient between the observed and the fitted values $\mathrm{cor}(o_i, \hat{\lambda}_i)$. The value R^D that exceeds the quantile $\chi^2_{0.95;85} = 107.5$ is marked with the star. The rest of the variables showed poor fit to the data.

	R^D	$\mathrm{cor}(o_i, \hat{\lambda}_i)$
$L3$	90.1	0.58
$L2$	91.9	0.69
$X3$	95.8	0.49
$V2$	103.1	0.42
$V1$	103.1	0.40
$X4$	115.8*	0.32

Table 2. Residual deviance R^D and correlation coefficient $\mathrm{cor}(o_i, \hat{\lambda}_i)$ for explanatory variables included into the model separately.

If all variables are included into the model the algorithm of stepwise regression using the minimization of the AIC results in the final model with the variables $L2$, $L3$ (AIC= 90.0418). We should prefer a simpler model with different types of explanatory variables because the information they contain is less dependent. The variables $L2$ and $L3$ are highly correlated therefore including the variable $X3$ instead of $L3$ could improve the model (AIC $= D^R + 2p = 91.9 + 4 = 95.9$). The correlation coefficient between the observed and the fitted values has the value $\mathrm{cor}(o_i, \hat{\lambda}_i) = 0.70$. The variable $V2$ did not show bad fit to the data but its inclusion into the model with the variables $L2$ and $X3$ would not improve the fit significantly.

The size of the forest 10–50 ha ($L2$) and the ratio of the mixed forest ($X3$) seem to be satisfactory explanatory variables explaining the number of infected and represent the required prior information in the study. The final model used in the next section to get the predicted values is

$$\log \hat{\lambda}_i = -0.04 + 3.93(X3)_i + 91.89(L2)_i, \tag{13}$$

where $D^R = 91.9 < \chi^2_{0.95;84} = 107.5$, $\hat{\varphi} = 0.95$, $D = D^0 - D^R = 134.8 - 91.9 = 42.9 > \chi^2_{0.95;2} = 6.0$.

4.3. Estimation of the risk

Here we deal with all squares including the border squares. The explanatory variables are in these squares corrected by multiplication by the variable *Overlap*.

The empirical Bayesian estimation is used. We substitute the result (13) of the regression $\pi_i = \hat{\lambda}_i/N_i$ into the formula (2) to get the posterior mean estimator as the estimate of the risk. In Fig. 6 the trivial estimates of the risk are compared with the posterior ones. The values are ordered by the size of the trivial estimates of the risk (in ascending order). The highest values are shown in Tab. 3.

Id	o	o/N	$E(r_i\|o_i)$
155	4	117.10	119.80
119	17	155.27	127.14
168	11	165.66	154.56
157	5	183.62	170.58
156	4	197.92	205.15
139	8	204.55	206.52
136	15	232.67	232.19
41	9	310.45	295.23
140	23	371.21	313.53
101	13	393.82	380.65
183	14	429.18	424.34
100	21	639.07	637.56
167	21	748.13	728.62
182	21	848.48	795.11
152	28	1132.23	1113.71

Table 3. The number of the square (Id) in Fig. 1 with the number of the infected persons (o_i), trivial estimate (o_i/N_i) and Bayesian estimate $E(r_i|o_i)$ of the risk [$\times 10^5$] for the squares with the highest estimated risk (in ascending order).

The posterior mean is very close to the trivial estimate. If the trivial estimate equals zero the posterior mean tends to be higher, on the contrary by very high trivial estimate the posterior mean is lower. There are, though, some remarkable outliers that are marked in Fig. 6 by the numbers. The numbers 1, 29, 54, 59 in Fig. 6 represent the squares with the number 11, 81, 195, 204 in Fig. 1, respectively. These squares have the overlap with the area of the CB less than 9 %. The correction of the explanatory variables in the border squares is unsatisfactory.

The posterior mean shows higher departure from the trivial estimate in further squares with numbers 109, 116, 126, 140. As for the squares 109, 126 in Fig. 1 the low posterior mean is caused by extremely low value of one or both explanatory variables. $X3$ is on the contrary very high in the square 116 and $L2$ is very low in the square 140.

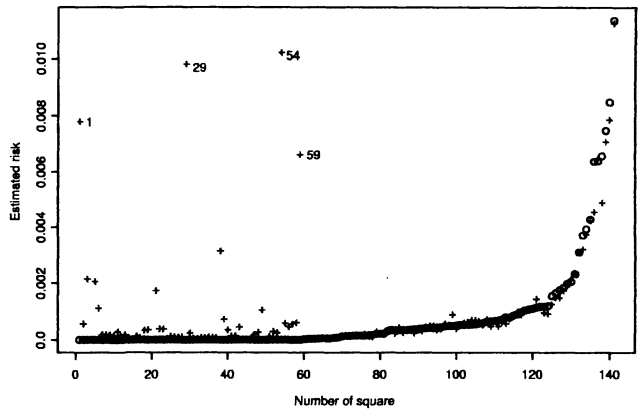

Figure 6. Trivial (o) and Bayesian estimates (+), mean value (in ascending order—the number of square is renumbered).

4.4. Properties of the risk estimators

Mean square error MSE is an appropriate measure of the quality of the estimation of the parameters comprising both bias and variability

$$\text{MSE} = \text{var}(\hat{r}) + (E(\hat{r}) - r)^2 \tag{14}$$

where $\hat{r}$ is the estimation of the unobservable risk r. Denote the mean value of the trivial (Bayesian) estimate by r_{BI} (r_{AP}), respectively.

In the Introduction, the possible bias of the trivial estimate r_{BI} was discussed. Intuitively, the zero risk obtained by the trivial estimate is less likely, and hence the Bayesian estimate never predicting zero value of the risk might be closer to the true risk and therefore less biased. If we do not consider the error due to regression in the estimation of the posterior mean we can compare the variability of the trivial and the Bayesian estimates in the range of the real data set.

The trivial estimate p, assuming that O_i are binomially distributed, has the mean r_{BI} and variance $r_{BI}(1 - r_{BI})/N$. The Bayesian estimate $\hat{r}_{AP}$ has the gamma distribution with the parameters (η, σ), see (1), with the mean (2) $r_{AP} = \eta\sigma = O + 1(N + 1/\pi)^{-1}$ and variance $\text{var}(\hat{r}_{AP}) = \eta\sigma^2 = O + 1(N + 1/\pi)^{-2}$.

In Fig. 7, the interquartile range $\widetilde{X}_{0.75} - \widetilde{X}_{0.25}$ of both distributions as a function of the mean value is shown. N is in both cases taken as $N = 2000$. The higher N the less different are both estimates. The vertical line can be understood as the risk, the maximum value 0.015 in Fig. 7 corresponds approximately to the maximum observed value in the square 152.

From the formula (2), it is obvious that the prior information makes always the denominator higher. To compare the variability of the trivial and the Bayesian

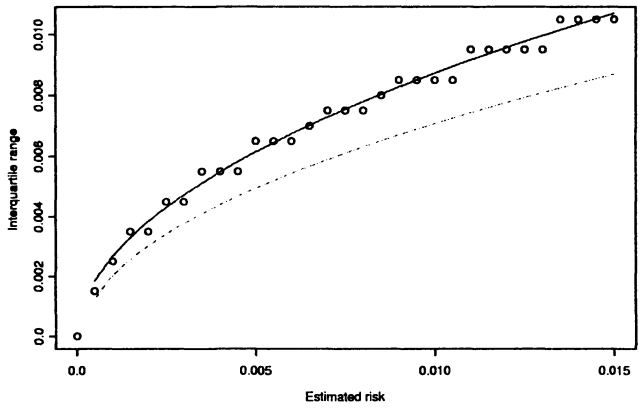

Figure 7. The comparison of the variability of the risk estimates as a function of $E(\hat{r})$. The points correspond to the binomial distribution with the mean $E(\hat{r}) = O/N$, the line represents the interquartile range of the gamma distribution with the parameters $\eta = O$, $\sigma = 1/N$ with the mean $E(\hat{r}) = \eta\sigma$, the dot-and-dash line corresponds the interquartile range of the gamma distribution with the $\sigma^* = (N + 1000)^{-1}$.

estimate for the same mean value the second parameter σ in gamma distribution has to be decreased. The interquartile range of the gamma distribution with the $\sigma^* = (N + 1000)^{-1}$ is drawn in Fig. 7 by dot-and-dash line. If both estimates were unbiased or equally biased the Bayesian one is that of lower variability for low level of risk $r < 0.015$ and hence of lower MSE.

If the trivial estimate underestimates the true risk there is no guaranty of lower MSE until the squared bias in the formula (14) exceeds the variance. This happens already for a bias of 3 infected from 2000 people in the square. Such an analysis applies for higher N values, too. It is worth mentioning that the variability of the trivial estimate is lower compared to the Bayesian estimate for the higher level of risk approximately $r > 0.1$, but these values do not appear in practice.

The pure comparison of credibility intervals without regard to the bias should be taken with caution but can still bring some information on reliability. The true bias can hardly be estimated. The 95 % credibility (confidence) intervals for the estimated risk in each square area based on the Bayesian (trivial) method are shown in Fig. 8. These intervals correspond to the squares with overlap greater than 90 % with the area of the CB, which are in ascending order according to mean estimated risk.

In 25 squares, there is no observed case of the infection and therefore the trivial estimates and the confidence intervals for the risk are equal to zero and are closer than the credibility intervals obtained from the Bayesian method. In the rest of the squares, 70 % of the credibility intervales obtained from the Bayesian method

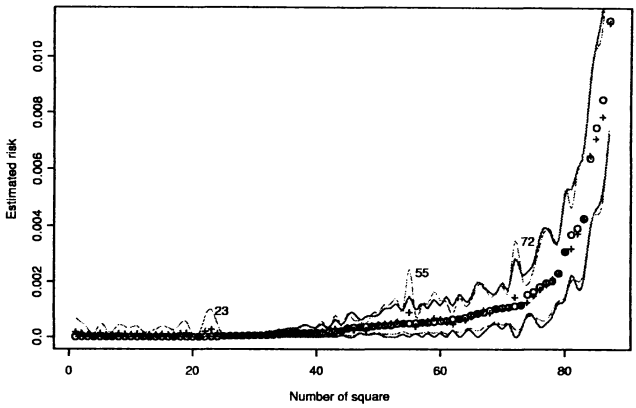

Figure 8. 95% intervals for trivial (full line) and Bayes risk estimates (broken line).

are closer than the trivial ones. The squares with high estimates of the risk by the Bayesian method are marked in Fig. 8 by the numbers 23, 55, 72. These numbers correspond to the squares 169, 116, 98 in Fig. 1. In all cases, the number of observed cases is zero or low and the Bayesian estimate is increased by prior information.

5. Conclusions

The whole paper is written in an area level data setting, an alternative approach discussed in Diggle [3] and Best et. al. [1] makes use of (marked) point process modelling of cases where intensity function is a crucial concept in the risk model. Both approaches make it possible to construct models with varying complexity, e.g. the ratio kernel estimator (Bithell [2]) of intensity function applied in Zeman [9] to the pattern of cases studied here is simpler than a hierarchical Bayesian model in Stern, Cressie [7] in an area level approach. Generally the point process approach may give a finer resolution of disease cases and it is both theoretically and computationally more demanding. The area level approach is finite-dimensional and useful in situations where a natural (geographical) division into subregions is desired or a fine resolution is not needed. E.g. for our pattern of 446 cases the area level approach with subregions in Fig. 1 must lead to satisfactory interpretations.

A substantial part of the paper is devoted to the choice of explanatory variables which influence most the risk of infection. The AIC criterion seems to be natural and enables to reduce large geographical information available to the most important factors.

For the risk estimation, a simple approach has been applied which misses modelling of spatial dependence and association, see e.g. Stern and Cressie [7]. We can compare

our approach with that of Stern and Cressie [7] as applied to our dataset in Mašata [4]. Mašata [4] used only explanatory variables X (type of forest) and a coarser division of CB into 41 subregions. Under the same conditions we obtain comparable results using our approach, see Fig. 9.

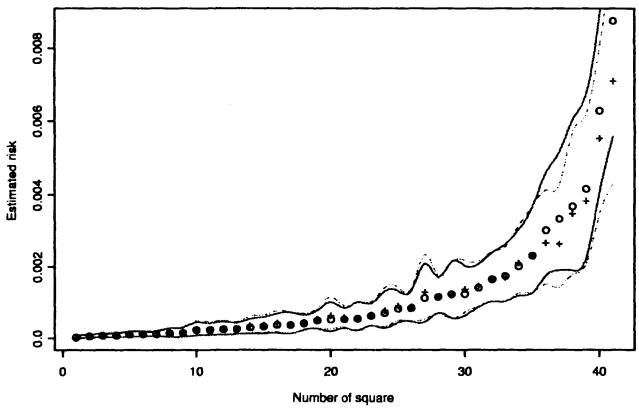

Figure 9. Comparison of the present method and that of Stern and Cressie [7]. Estimated risk is on the vertical axis, the number of the subarea is on the horizontal axis. Empirical Bayesian estimates are denoted by (o) and estimates obtained by Mašata [4] by (+). 95 % credibility intervals for empirical Bayesian estimates (full line) and Mašata [4] estimates (broken line) are drawn.

In Fig. 9, we observe that in six subregions with highest mean posterior risk the values obtained by us are about 10 % higher. This may be caused by the fact that in the model with spatial dependence the extreme risk estimates are dropped by the values in neighbouring subregions. But this does not lead to a conclusion that the more complex model is better. Unfortunately, we are not able to distinguish clustering and spatial variation from a single point pattern of cases. Since clustering is a departure from the assumption of independent cases and spatial variation is a departure from the assumption of equal risk (see Diggle [3]), spatial variation is expected in our case and not dependence. Concerning the credibility intervals in Fig. 9 the more complex model applied by Mašata [4] does not lead to a substantial precision improvement. Summarizing these observations we conclude that the presented approach yields an appropriate method for evaluation the tick-borne encephalitis infection risk.

References

[1] N. G. Best, K. Ickstadt, and R. L. Wolpert Spatial Poisson regression for health and exposure data measured at disparate resolutions. Journal of the American Statistical Association *95* (2000), 1076–1088.

[2] *J. F. Bithell*: An application of density estimation to geographical epidemiology. Statistics in Medicine *9* (1980), 691–701.

[3] *P. Diggle*: Overview of statistical methods for disease mapping and its relationship to cluster detection. In: Spatial Epidemiology: Methods and Applications (P. Elliott et al., eds.). Oxford University Press, Oxford, 2000, pp. 87–103.

[4] *M. Mašata*: Assessment of risk of infection by means of a Bayesian method. In: Proceedings S^4G International Conference on Stereology, Spatial Statistics and Stochastic Geometry (V. Beneš, J. Janáček, and I. Saxl, eds.). JČMF, Praha, 1999, pp. 197–202.

[5] *P. McCullagh, J. A. Nelder*: Generalized Linear Models. Chapman & Hall, London, 1992, pp. 26–43, 193–200.

[6] *A. Mollie, S. Richardson*: Empirical Bayes estimates of cancer mortality rates using spatial models. Statistics in Medicine *10* (1991), 95–112.

[7] *S. H. Stern, N. Cressie*: Inference for extremes in disease mapping. Methods of Disease Mapping and Risk Assessment for Public Health Decision Making (A. Lawson et al., eds.). Wiley, New York, 1999, pp. 63–84.

[8] *W. N. Venables, B. D. Ripley*: Modern Applied Statistics with S-PLUS. Springer, New York, 1997, pp. 242–243.

[9] *P. Zeman*: Objective assessment of risk maps of tick-borne encephalitis and lyme borreliosis based on spatial patterns of located cases. International Journal of Epidemiology *26* (1997), 1121–1130.

Authors' addresses: *M. Jiruše*, University of Economics, Department of Probability and Statistics, Nám. Winstona Churchilla 4, 130 00 Praha 3, Czech Republic, e-mail: `yjiruse@vse.cz`; *J. Machek*, Charles University, Department of Probability and Statistics, Sokolovská 83, 186 75 Praha 8, Czech Republic, e-mail: josef.machek@mff.cuni.cz; *V. Beneš*, Charles University, Department of Probability and Statistics, Sokolovská 83, 186 75 Praha 8, Czech Republic, e-mail: `Viktor.Benes@mff.cuni.cz`; *P. Zeman*, Regional Centre of Hygiene, Dittrichova 14, 120 07 Praha 2, Czech Republic.

49 (2004) APPLICATIONS OF MATHEMATICS No. 5, 389–404

A BAYESIAN ESTIMATE OF THE RISK OF TICK-BORNE DISEASES*

Marek Jiruše, Josef Machek, Viktor Beneš, and Petr Zeman, Praha

(Received March 22, 2002, in revised version June 14, 2002)

Abstract. The paper considers the problem of estimating the risk of a tick-borne disease in a given region. A large set of epidemiological data is evaluated, including the point pattern of collected cases, the population map and covariates, i.e. explanatory variables of geographical nature, obtained from GIS.

The methodology covers the choice of those covariates which influence the risk of infection most. Generalized linear models are used and AIC criterion yields the decision. Further, an empirical Bayesian approach is used to estimate the parameters of the risk model. Statistical properties of the estimators are investigated. Finally, a comparison with earlier results is discussed from the point of view of statistical disease mapping.

Keywords: Bayesian estimation, generalized linear model, epidemiological data, statistical properties

MSC 2000: 62G05

1. Introduction

Statistical disease mapping aims to characterize the spatial variation of cases of a disease and to study connections with given covariates. Most analyses published in the literature are based on the area-level approach which means that cases are aggregated in subregions. If a point pattern of cases is available, a planar point process modelling is another possible approach. In Diggle [3] a survey of both approaches is given.

We are interested in estimating the risk that a person gets infected by a tick-borne disease at a specific location, which is the task usually dealt with by epidemiologists

* This work was supported by the grant No. 201/98/0090 of the Grant Agency of the Czech Republic, and by the project MSM 113200008.

and medical practitioners making decision on prophylactic measures in endemic areas (Zeman [9]). Suppose that the incidence of the disease has been registered for a representative period, and the total human population exposed is known. Further, assume that the area of interest is divided into n regions which may be either administrative or deliberately designed for the purposes of this study. In this area-level approach, for each region an average risk is enquired disregarding individual times of exposure to the risk factor.

Let the number of people living in the ith region be N_i. If the probabilities of being infected were the same for all of them, say equal to π_i, then the observed proportion of the infected $p_i = O_i/N_i$ would be an unbiased estimate of the probability π_i, the risk of infection. The reality, however, is more complex. Every inhabitant of the region has a different exposure intensity and hence a different probability of being infected and it is the average value of these probabilities which should be called the risk of infection for the region, say r_i. The approach used in the present paper is to consider this parameter as a random variable having some prior distribution that depends on a set of variables which describe the environmental conditions. This approach belongs to a class of empirical Bayesian approaches widely used in epidemiology, see e.g. Mollie and Richardson [6].

In this paper first the dataset is described and the process of selection of explanatory variables is discussed in detail. Using the statistical model the desired risk estimators are obtained. The results are compared with the earlier work of Mašata [4] based on a different Bayesian approach.

2. Description of the data

The whole area of the Central Bohemia (CB) was divided into 141 small squares, see Fig. 1. The hole in the middle of Fig. 1 corresponds to the capital Prague which is not a part of CB administratively and its inhabitants are excluded from the analysis. For each square we know, cf. Zeman [9]

- O ... the number of persons infected by tick-borne encephalitis reported in the years 1971–1993 (446 cases altogether)
- N ... the number of inhabitants living in each square (total about 1.1 million)
- *Overlap* ... the proportion of the square area which lies within CB. There are 54 squares partly outside of the CB, we call them border squares thereafter.

The following data that provide additional information about the individual regions will be called explanatory variables (related to each square area):

- $X2, X3, X4$ ... the proportion of the area of coniferous, mixed, and deciduous forest, respectively.

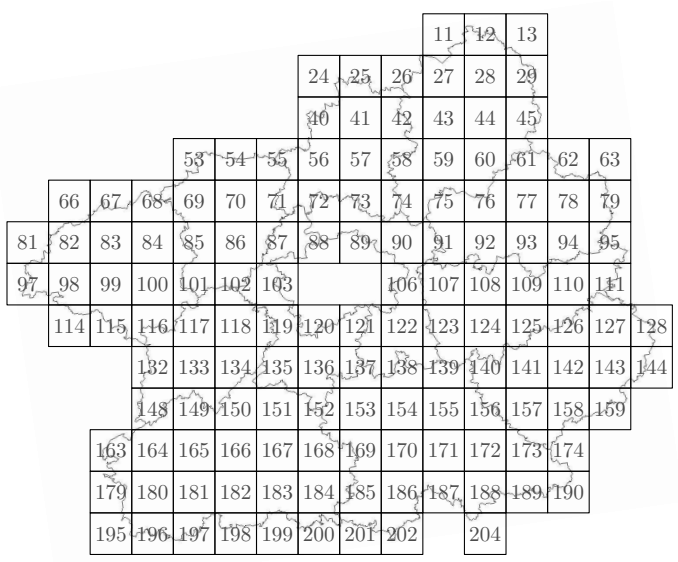

Figure 1. The area of the Central Bohemia divided into 141 squares sized 10×10 km. The administrative subdivision into districts (see the boundaries) is too coarse for the analysis.

- $V1, V2, V3, V4$ ... the proportion of the area having elevation < 300 m, 300–500 m, 500–700 m, and 700–900 m, respectively.
- $L1, L2, L3, L4, L5, L6$ ... the proportion of the area covered by forests of individual area 1–10 ha, 10–50 ha, 50–150 ha, 150–300 ha, 300–600 ha, and > 600 ha, respectively.

The original elevation model was obtained from the Institute of Military Topography, Dobruška, all the other explanatory variables were derived from the satellite images of LANDSAT-5 MSS with resolution power of 80×80 m.

3. Statistical methods

3.1. Empirical Bayesian approach

Suppose that the risk r_i in the ith square is a random variable with some prior distribution. We assume the exponential prior distribution with the mean $E(r_i) = \pi_i$. This is partly justified by the argument that smaller risk values are more probable than larger ones.

Assume that the number of infected persons O_i can be approximated by the Poisson distribution. The posterior distribution of the risk $r_i|O_i$ has gamma distribution

with parameters

$$\eta = O + 1 \quad \text{and} \quad \sigma = \frac{1}{N + 1/\pi}, \tag{1}$$

i.e. the posterior mean is

$$E(r_i|O_i) = \frac{O_i + 1}{N_i + 1/\pi_i}. \tag{2}$$

We will use this posterior mean for the risk quantification. In (2), the trivial estimate of the risk is corrected by the value π which is a function of explanatory variables.

3.2. Optimizing the set of explanatory variables

Here we assume again that the response variable O is approximated by the Poisson distribution with the mean λ. Suppose that the number of infected persons O after appropriate transformation depends on the explanatory variables and the relationship is linear

$$\lambda = f(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p). \tag{3}$$

At a first glance it may seem surprising that the number N_i of persons exposed to the risk does not appear explicitly in the function (3). It should be born in mind, however, that N_i enters the result in another way, see formula (2).

The suitable function and the estimation of the parameters can be found by the theory of GLM models, see McCullagh and Nelder [5]. GLM models extend the classical regression analysis to the case of the response variable coming from the distribution of exponential family. This is in many cases more natural than the assumption of normality. The distribution of exponential family can be written in the form

$$g(y; \theta, \varphi) = \exp\Big\{ \frac{(y\theta - b(\theta))w}{\varphi} + c(y, \varphi, w) \Big\} \tag{4}$$

where $b(\cdot)$, $c(\cdot)$ are specific functions depending on this distribution, θ is the so-called canonical parameter, φ is the dispersion parameter, and w is a known weight.

The canonical parameter θ contains the 'natural' link function that yields the transformation of the mean of the response variable into the linear relationship with the explanatory variables. Another type of the link function can be chosen than the natural one. The use of the natural link function gives to the estimates better

theoretical properties (existence and uniqueness). The probability function for the Poisson distribution can be written in the form (4) as

$$P(O=o)=\exp\Big\{o\underbrace{\log\lambda}_{\theta}-\underbrace{\lambda}_{b(\theta)}\underbrace{-\log o!}_{c(o,\varphi)}\Big\}, \tag{5}$$

where $\theta=\log\lambda$ is the natural link function. Hence the function f in (3) is

$$\lambda=\mathrm{e}^{\beta_0+\beta_1x_1+\ldots+\beta_px_p}. \tag{6}$$

The dispersion parameter for the Poisson distribution is $\varphi=1$. For the mean and variance it holds

$$E(O)=b'(\theta)=\lambda,\quad \mathrm{var}(O)=b''(\theta)\varphi=\lambda. \tag{7}$$

The estimation of parameters is based on the maximum likelihood method when substituting (6) for the parameter λ into the likelihood function. The estimates are obtained iteratively using the IWLS (iterative weighted least squares regression) described in McCullagh and Nelder [5].

In the classical linear regression, the residual sum of squares is a measure of the goodness of fit, in GLM models residual deviance D^R is used instead (in the following bold letters are used for vectors, e.g. $\boldsymbol{o}=(o_i)_{i=1}^n$),

$$D^R=2l(\boldsymbol{o})-2l(\hat{\boldsymbol{\lambda}},\boldsymbol{o}). \tag{8}$$

Here $l(\boldsymbol{o})$ is the maximum likelihood function obtained by the absolute fit of observed data with the model, that is $\lambda=o$. Similary $l(\hat{\boldsymbol{\lambda}},\boldsymbol{o})$ is the likelihood function with substituted ML estimates. D^R has approximately χ^2_{n-q} distribution, where q is the number of parameters. The residual deviance D^R is also a relative measure to compare two models. Denote $D^0=2l(\boldsymbol{o})-2l^0$ the null deviance where l^0 is the null model of likelihood function, i.e. no relationship is assumed and the predicted values of $\hat{\boldsymbol{\lambda}}$ are substituted by the overall mean $\hat{\lambda}=\sum_{i=1}^{n}o_i/n$.

The fit of our model can be quantified relatively when we substract the residual deviance D^R from null deviance D^0,

$$D=D^0-D^R=2l(\hat{\boldsymbol{\lambda}},\boldsymbol{o})-2l^0; \tag{9}$$

D has approximately χ^2_1 distribution.

Accordingly, the contribution of explanatory variables added into the model can be tested. For comparison between the model G and F the deviance is obtained as

$D(\hat{\boldsymbol{\lambda}}^{\boldsymbol{G}}, \hat{\boldsymbol{\lambda}}^{\boldsymbol{F}}, \boldsymbol{o}) = 2l(\hat{\boldsymbol{\lambda}}^{\boldsymbol{G}}, \boldsymbol{o}) - 2l(\hat{\boldsymbol{\lambda}}^{\boldsymbol{F}}, \boldsymbol{o})$ where $D(\hat{\boldsymbol{\lambda}}^{\boldsymbol{G}}, \hat{\boldsymbol{\lambda}}^{\boldsymbol{F}}, \boldsymbol{o}) \sim \chi^2_{t-s}$, t is the number of parameters of the model G and s is the number of parameters of the model F. In classical regression analysis, the residual sum of squares is reduced by adding any explanatory variable. The same applies for residual deviance. The AIC criterion (Akaike Information Criterion) takes into account the number of estimated parameters

$$\text{AIC} = D^R + 2p\hat{\varphi} \tag{10}$$

where p is the number of parameters, and $\hat{\varphi}$ is the estimate of the dispersion parameter from the sample

$$\hat{\varphi} = \frac{1}{n-p} \sum \frac{(o_i - \hat{\lambda})^2}{\hat{\lambda}}. \tag{11}$$

The criterion for the choice among all possible models is the minimum of AIC.

3.3. Overdispersion

For the Poisson distributed response variable, the variance (7) and the dispersion parameter $\varphi = 1$ are given. High values of residual deviance, which is here the measure of the fit, can cause the rejection of the model. This phenomenon is called 'overdispersion', which means that the model has higher variability than assumed. The indicator that the overdispersion is present in the model can be also the estimated dispersion parameter φ. A value greater than 1 makes the variance of the model for Poisson data greater than in (7). To avoid the problem of overdispersed data we can replace the assumption of Poisson distributed response variable by negative binomial distribution which is more flexible due to an additional parameter τ. More details are described in Venables and Ripley [8]. The probability function of the negative binomial distribution in the form of (4) can be written as

$$P(O = o) = \exp\left\{ o \underbrace{\log\left(\frac{\lambda}{\tau + \lambda}\right)}_{\theta} - \underbrace{\tau \log\left(\frac{\tau + \lambda}{\tau}\right)}_{b(\theta)} + c(o, \theta) \right\}. \tag{12}$$

In (12), the natural link function is θ, and the dispersion parameter is $\varphi = 1$. It holds $E(O) = \lambda$, $\text{var}(O) = \lambda + \lambda^2/\tau$, so the variance is greater than the mean, which may be a better reflection of reality.

4. Numerical results

Here we deal with the squares where we have complete (or almost complete) information, i.e. with the squares in Fig. 1 having the overlap with the CB more than 90 %. Since it is assumed that the explanatory variables affect the response variable by a linear relationship after transformation of the response variable, the simple scatterplots between response and explanatory variables can be the first indicators of what the dependency might be.

We distinguish between the random variable O_i and its observed value o_i. In Fig. 2 the relationship between the number of infected persons o_i and the variables $X3$ and $X4$ is obvious. The variable $X2$ does not show any dependency.

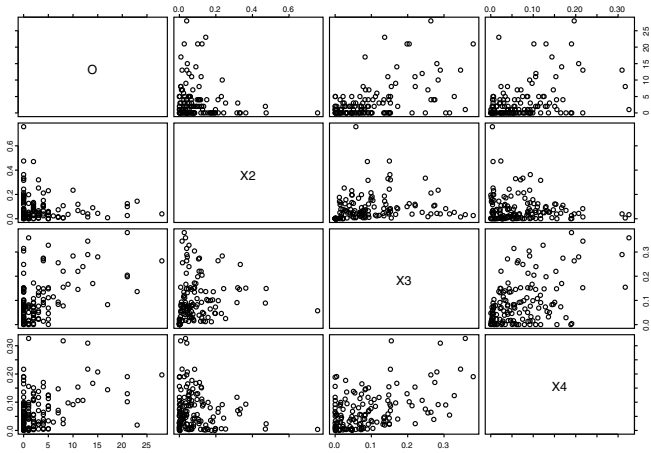

Figure 2. Scatterplots of variables o, $X2$–$X4$.

In Fig. 3, the variable $V2$ shows the best correlation with the number of infected, and so do the variables $L2$ and $L3$ in Fig. 4. The simple correlation coefficients are shown in the Tabs. 1a, 1b.

	$X2$	$X3$	$X4$	$V1$	$V2$	$V3$	$V4$
o	-0.02	0.56	0.40	-0.40	0.43	0.03	-0.08

Table 1a. Correlation coefficients between variables o and X, V, respectively.

	$L1$	$L2$	$L3$	$L4$	$L5$	$L6$
o	0.37	0.62	0.61	0.24	-0.02	-0.07

Table 1b. Correlation coefficients between variables o and L.

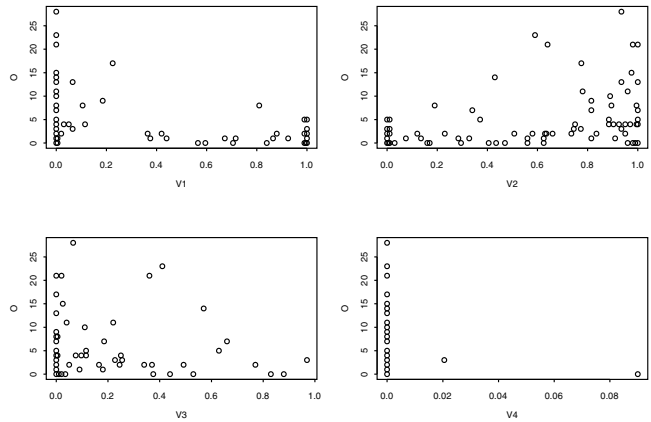

Figure 3. Scatterplots of variables o, $V1$–$V4$.

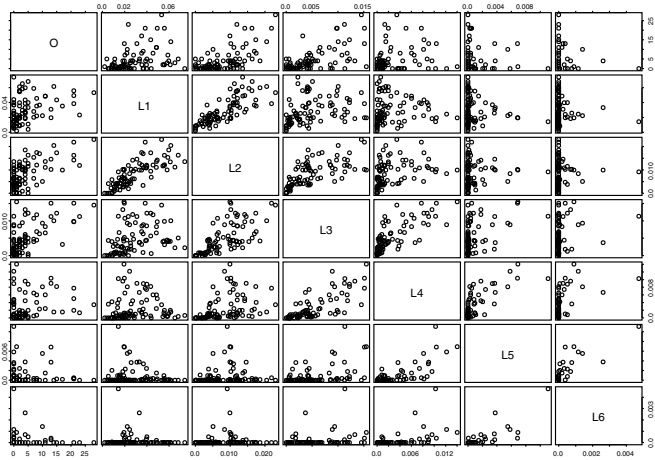

Figure 4. Scatterplots of variables o, $L1$–$L6$.

4.1. The choice of the model

In the first step, we have to find an appropriate model for the distribution of the response variable and the form of the link function. We deal with the variable $X3$ at first, which exhibited positive correlation with the number of infected, and assume the number of infected in the square area to be Poisson distributed. The estimation of the parameters using the natural link function coming from (5) has been carried out in the programming language S supported by the statistical software S-PLUS

$$\log \hat{\lambda}_i = 0.62 + 6.14(X3)_i.$$

The residual deviance exceeded the quantile of its approximative distribution $D^R = 388.3 > \chi^2_{0.95;85} = 107.5$. The estimated dispersion parameter from (11) is $\hat{\varphi} = 5$. The data show higher variability than expected by Poisson model. We try to estimate the parameters from the likelihood function computed for negative binomial distribution of the response variable. The program does not use the natural link function as in (12) and works with the same link function as in the Poisson model

$$\log \hat{\lambda}_i = 0.40 + 7.73(X3)_i, \quad \hat{\tau} = 1.01, \quad \hat{\varphi} = 1.02.$$

This time the residual deviance did not exceed the quantile of its approximative distribution $D^R = 96.6 < \chi^2_{0.95;85} = 107.5$ and the estimation of the dispersion parameter corresponds to its expected value 1. The goodness of fit with the model using the variable $X3$ is confirmed since in (9) it holds

$$D = D^0 - D^R = 136.0 - 96.6 = 39.4 > 3.84 = \chi^2_{0.95;1}.$$

The plot of the observed values o_i against the values $\hat{\lambda}_i$ predicted by the model with the explanatory variable $X3$ shows the goodness of fit in Fig. 5.

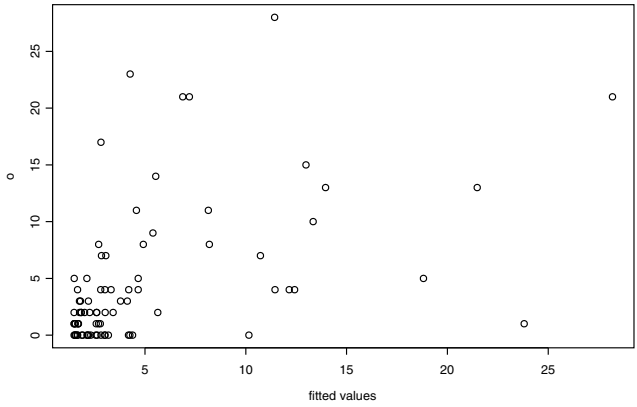

Figure 5. Observed cases o_i plotted against the fitted $\hat{\lambda}_i$.

The fitted values should in the ideal case lie on the diagonal. The correlation coefficient is $\mathrm{cor}(o_i, \hat{\lambda}_i) = 0.49$. There is still a lack of fit using only the variable $X3$ and we keep looking for a set of available explanatory variables to reach a better fit of the data. The model for the negative binomial distribution of the response variable O is used for that. The estimate $\hat{\tau} \doteq 1$ is used fixed for better comparison among the models with other explanatory variables.

4.2. The choice of the set of the explanatory variables

In the first column of Tab. 2, there are six explanatory variables that are included into the model separately and ordered by the size of the residual deviance R^D. In the third column of Tab. 2 there is the value of the correlation coefficient between the observed and the fitted values $\mathrm{cor}(o_i, \hat{\lambda}_i)$. The value R^D that exceeds the quantile $\chi^2_{0.95;85} = 107.5$ is marked with the star. The rest of the variables showed poor fit to the data.

	R^D	$\mathrm{cor}(o_i, \hat{\lambda}_i)$
$L3$	90.1	0.58
$L2$	91.9	0.69
$X3$	95.8	0.49
$V2$	103.1	0.42
$V1$	103.1	0.40
$X4$	115.8∗	0.32

Table 2. Residual deviance R^D and correlation coefficient $\mathrm{cor}(o_i, \hat{\lambda}_i)$ for explanatory variables included into the model separately.

If all variables are included into the model the algorithm of stepwise regression using the minimization of the AIC results in the final model with the variables $L2$, $L3$ (AIC= 90.0418). We should prefer a simpler model with different types of explanatory variables because the information they contain is less dependent. The variables $L2$ and $L3$ are highly correlated therefore including the variable $X3$ instead of $L3$ could improve the model (AIC $= D^R + 2p = 91.9 + 4 = 95.9$). The correlation coefficient between the observed and the fitted values has the value $\mathrm{cor}(o_i, \hat{\lambda}_i) = 0.70$. The variable $V2$ did not show bad fit to the data but its inclusion into the model with the variables $L2$ and $X3$ would not improve the fit significantly.

The size of the forest 10–50 ha ($L2$) and the ratio of the mixed forest ($X3$) seem to be satisfactory explanatory variables explaining the number of infected and represent the required prior information in the study. The final model used in the next section to get the predicted values is

$$\log \hat{\lambda}_i = -0.04 + 3.93(X3)_i + 91.89(L2)_i, \tag{13}$$

where $D^R = 91.9 < \chi^2_{0.95;84} = 107.5$, $\hat{\varphi} = 0.95$, $D = D^0 - D^R = 134.8 - 91.9 = 42.9 > \chi^2_{0.95;2} = 6.0$.

4.3. Estimation of the risk

Here we deal with all squares including the border squares. The explanatory variables are in these squares corrected by multiplication by the variable *Overlap*.

The empirical Bayesian estimation is used. We substitute the result (13) of the regression $\pi_i = \hat{\lambda}_i/N_i$ into the formula (2) to get the posterior mean estimator as the estimate of the risk. In Fig. 6 the trivial estimates of the risk are compared with the posterior ones. The values are ordered by the size of the trivial estimates of the risk (in ascending order). The highest values are shown in Tab. 3.

Id	o	o/N	$E(r_i\|o_i)$
155	4	117.10	119.80
119	17	155.27	127.14
168	11	165.66	154.56
157	5	183.62	170.58
156	4	197.92	205.15
139	8	204.55	206.52
136	15	232.67	232.19
41	9	310.45	295.23
140	23	371.21	313.53
101	13	393.82	380.65
183	14	429.18	424.34
100	21	639.07	637.56
167	21	748.13	728.62
182	21	848.48	795.11
152	28	1132.23	1113.71

Table 3. The number of the square (Id) in Fig. 1 with the number of the infected persons (o_i), trivial estimate (o_i/N_i) and Bayesian estimate $E(r_i|o_i)$ of the risk [$\times 10^5$] for the squares with the highest estimated risk (in ascending order).

The posterior mean is very close to the trivial estimate. If the trivial estimate equals zero the posterior mean tends to be higher, on the contrary by very high trivial estimate the posterior mean is lower. There are, though, some remarkable outliers that are marked in Fig. 6 by the numbers. The numbers 1, 29, 54, 59 in Fig. 6 represent the squares with the number 11, 81, 195, 204 in Fig. 1, respectively. These squares have the overlap with the area of the CB less than 9 %. The correction of the explanatory variables in the border squares is unsatisfactory.

The posterior mean shows higher departure from the trivial estimate in further squares with numbers 109, 116, 126, 140. As for the squares 109, 126 in Fig. 1 the low posterior mean is caused by extremely low value of one or both explanatory variables. $X3$ is on the contrary very high in the square 116 and $L2$ is very low in the square 140.

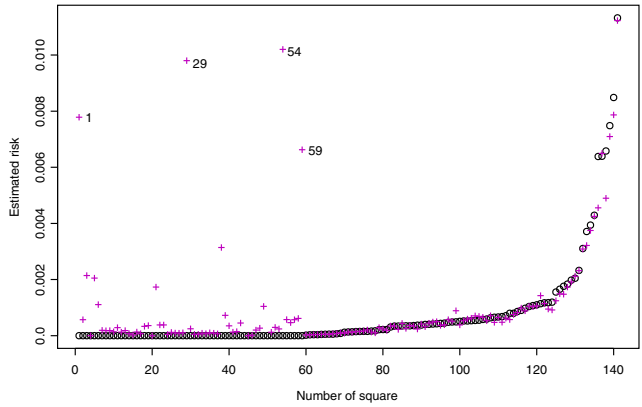

Figure 6. Trivial (o) and Bayesian estimates (+), mean value (in ascending order—the number of square is renumbered).

4.4. Properties of the risk estimators

Mean square error MSE is an appropriate measure of the quality of the estimation of the parameters comprising both bias and variability

$$\text{MSE} = \text{var}(\hat{r}) + (E(\hat{r}) - r)^2 \tag{14}$$

where $\hat{r}$ is the estimation of the unobservable risk r. Denote the mean value of the trivial (Bayesian) estimate by r_{BI} (r_{AP}), respectively.

In the Introduction, the possible bias of the trivial estimate r_{BI} was discussed. Intuitively, the zero risk obtained by the trivial estimate is less likely, and hence the Bayesian estimate never predicting zero value of the risk might be closer to the true risk and therefore less biased. If we do not consider the error due to regression in the estimation of the posterior mean we can compare the variability of the trivial and the Bayesian estimates in the range of the real data set.

The trivial estimate p, assuming that O_i are binomially distributed, has the mean r_{BI} and variance $r_{BI}(1 - r_{BI})/N$. The Bayesian estimate $\hat{r}_{AP}$ has the gamma distribution with the parameters (η, σ), see (1), with the mean (2) $r_{AP} = \eta\sigma = O + 1(N + 1/\pi)^{-1}$ and variance $\text{var}(\hat{r}_{AP}) = \eta\sigma^2 = O + 1(N + 1/\pi)^{-2}$.

In Fig. 7, the interquartile range $\widetilde{X}_{0.75} - \widetilde{X}_{0.25}$ of both distributions as a function of the mean value is shown. N is in both cases taken as $N = 2000$. The higher N the less different are both estimates. The vertical line can be understood as the risk, the maximum value 0.015 in Fig. 7 corresponds approximately to the maximum observed value in the square 152.

From the formula (2), it is obvious that the prior information makes always the denominator higher. To compare the variability of the trivial and the Bayesian

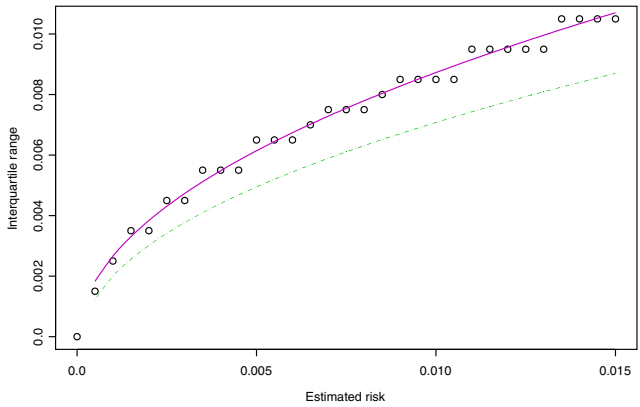

Figure 7. The comparison of the variability of the risk estimates as a function of $E(\hat{r})$. The points correspond to the binomial distribution with the mean $E(\hat{r}) = O/N$, the line represents the interquartile range of the gamma distribution with the parameters $\eta = O$, $\sigma = 1/N$ with the mean $E(\hat{r}) = \eta\sigma$, the dot-and-dash line corresponds the interquartile range of the gamma distribution with the $\sigma^* = (N + 1000)^{-1}$.

estimate for the same mean value the second parameter σ in gamma distribution has to be decreased. The interquartile range of the gamma distribution with the $\sigma^* = (N + 1000)^{-1}$ is drawn in Fig. 7 by dot-and-dash line. If both estimates were unbiased or equally biased the Bayesian one is that of lower variability for low level of risk $r < 0.015$ and hence of lower MSE.

If the trivial estimate underestimates the true risk there is no guaranty of lower MSE until the squared bias in the formula (14) exceeds the variance. This happens already for a bias of 3 infected from 2000 people in the square. Such an analysis applies for higher N values, too. It is worth mentioning that the variability of the trivial estimate is lower compared to the Bayesian estimate for the higher level of risk approximately $r > 0.1$, but these values do not appear in practice.

The pure comparison of credibility intervals without regard to the bias should be taken with caution but can still bring some information on reliability. The true bias can hardly be estimated. The 95 % credibility (confidence) intervals for the estimated risk in each square area based on the Bayesian (trivial) method are shown in Fig. 8. These intervals correspond to the squares with overlap greater than 90 % with the area of the CB, which are in ascending order according to mean estimated risk.

In 25 squares, there is no observed case of the infection and therefore the trivial estimates and the confidence intervals for the risk are equal to zero and are closer than the credibility intervals obtained from the Bayesian method. In the rest of the squares, 70 % of the credibility intervales obtained from the Bayesian method

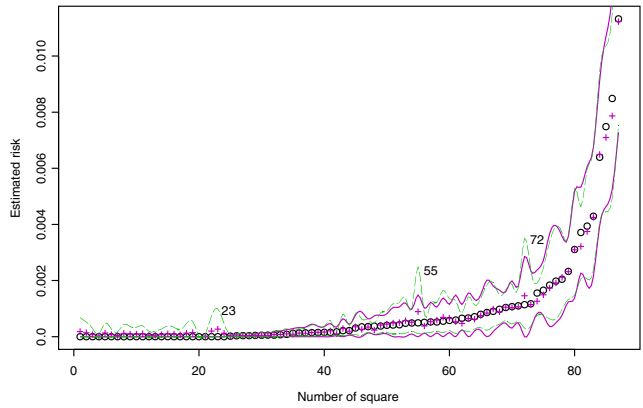

Figure 8. 95 % intervals for trivial (full line) and Bayes risk estimates (broken line).

are closer than the trivial ones. The squares with high estimates of the risk by the Bayesian method are marked in Fig. 8 by the numbers 23, 55, 72. These numbers correspond to the squares 169, 116, 98 in Fig. 1. In all cases, the number of observed cases is zero or low and the Bayesian estimate is increased by prior information.

5. Conclusions

The whole paper is written in an area level data setting, an alternative approach discussed in Diggle [3] and Best et. al. [1] makes use of (marked) point process modelling of cases where intensity function is a crucial concept in the risk model. Both approaches make it possible to construct models with varying complexity, e.g. the ratio kernel estimator (Bithell [2]) of intensity function applied in Zeman [9] to the pattern of cases studied here is simpler than a hierarchical Bayesian model in Stern, Cressie [7] in an area level approach. Generally the point process approach may give a finer resolution of disease cases and it is both theoretically and computationally more demanding. The area level approach is finite-dimensional and useful in situations where a natural (geographical) division into subregions is desired or a fine resolution is not needed. E.g. for our pattern of 446 cases the area level approach with subregions in Fig. 1 must lead to satisfactory interpretations.

A substantial part of the paper is devoted to the choice of explanatory variables which influence most the risk of infection. The AIC criterion seems to be natural and enables to reduce large geographical information available to the most important factors.

For the risk estimation, a simple approach has been applied which misses modelling of spatial dependence and association, see e.g. Stern and Cressie [7]. We can compare

our approach with that of Stern and Cressie [7] as applied to our dataset in Mašata [4]. Mašata [4] used only explanatory variables X (type of forest) and a coarser division of CB into 41 subregions. Under the same conditions we obtain comparable results using our approach, see Fig. 9.

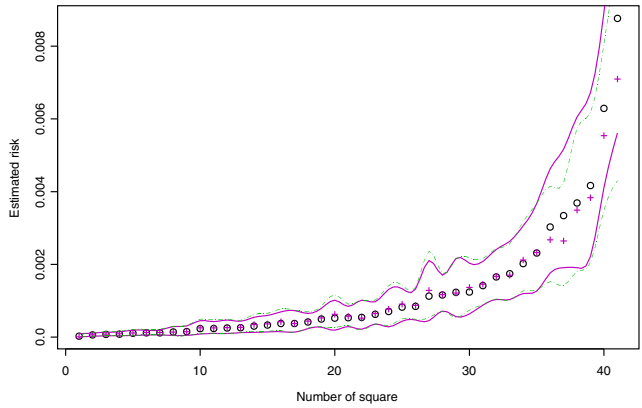

Figure 9. Comparison of the present method and that of Stern and Cressie [7]. Estimated risk is on the vertical axis, the number of the subarea is on the horizontal axis. Empirical Bayesian estimates are denoted by (o) and estimates obtained by Mašata [4] by (+). 95 % credibility intervals for empirical Bayesian estimates (full line) and Mašata [4] estimates (broken line) are drawn.

In Fig. 9, we observe that in six subregions with highest mean posterior risk the values obtained by us are about 10 % higher. This may be caused by the fact that in the model with spatial dependence the extreme risk estimates are dropped by the values in neighbouring subregions. But this does not lead to a conclusion that the more complex model is better. Unfortunately, we are not able to distinguish clustering and spatial variation from a single point pattern of cases. Since clustering is a departure from the assumption of independent cases and spatial variation is a departure from the assumption of equal risk (see Diggle [3]), spatial variation is expected in our case and not dependence. Concerning the credibility intervals in Fig. 9 the more complex model applied by Mašata [4] does not lead to a substantial precision improvement. Summarizing these observations we conclude that the presented approach yields an appropriate method for evaluation the tick-borne encephalitis infection risk.

References

[1] N. G. Best, K. Ickstadt, and R. L. Wolpert Spatial Poisson regression for health and exposure data measured at disparate resolutions. Journal of the American Statistical Association *95* (2000), 1076–1088.

[2] *J. F. Bithell*: An application of density estimation to geographical epidemiology. Statistics in Medicine *9* (1980), 691–701.

[3] *P. Diggle*: Overview of statistical methods for disease mapping and its relationship to cluster detection. In: Spatial Epidemiology: Methods and Applications (P. Elliott et al., eds.). Oxford University Press, Oxford, 2000, pp. 87–103.

[4] *M. Mašata*: Assessment of risk of infection by means of a Bayesian method. In: Proceedings S ^{4}G International Conference on Stereology, Spatial Statistics and Stochastic Geometry (V. Beneš, J. Janáček, and I. Saxl, eds.). JČMF, Praha, 1999, pp. 197–202.

[5] *P. McCullagh, J. A. Nelder*: Generalized Linear Models. Chapman & Hall, London, 1992, pp. 26–43, 193–200.

[6] *A. Mollie, S. Richardson*: Empirical Bayes estimates of cancer mortality rates using spatial models. Statistics in Medicine *10* (1991), 95–112.

[7] *S. H. Stern, N. Cressie*: Inference for extremes in disease mapping. Methods of Disease Mapping and Risk Assessment for Public Health Decision Making (A. Lawson et al., eds.). Wiley, New York, 1999, pp. 63–84.

[8] *W. N. Venables, B. D. Ripley*: Modern Applied Statistics with S-PLUS. Springer, New York, 1997, pp. 242–243.

[9] *P. Zeman*: Objective assessment of risk maps of tick-borne encephalitis and lyme borreliosis based on spatial patterns of located cases. International Journal of Epidemiology *26* (1997), 1121–1130.

Authors' addresses: *M. Jiruše*, University of Economics, Department of Probability and Statistics, Nám. Winstona Churchilla 4, 130 00 Praha 3, Czech Republic, e-mail: `yjiruse@vse.cz`; *J. Machek*, Charles University, Department of Probability and Statistics, Sokolovská 83, 186 75 Praha 8, Czech Republic, e-mail: josef.machek@mff.cuni.cz; *V. Beneš*, Charles University, Department of Probability and Statistics, Sokolovská 83, 186 75 Praha 8, Czech Republic, e-mail: `Viktor.Benes@mff.cuni.cz`; *P. Zeman*, Regional Centre of Hygiene, Dittrichova 14, 120 07 Praha 2, Czech Republic.