Staubach, Schmid, Ziller, Knorr-Held:

A Bayesian Model for Spatial Disease Prevalence Data

Sonderforschungsbereich 386, Paper 254 (2001)

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

A Bayesian Model for Spatial Disease Prevalence Data

Staubach, C.[a,*], Schmid, V.[b], Ziller, M.[a], Knorr-Held, L.[b,c]

[a]Federal Research Centre for Virus Diseases of Animals, Institute of Epidemiology; Seestr. 55, D-16868 Wusterhausen/Dosse, Germany

[b]Institute of Statistics, Ludwig-Maximilians-University Munich; Ludwigstr. 33, D-80539 Munich, Germany

[c]Department of Mathematics and Statistics, Lancaster University, Fylde College, Lancaster LA1 4YF, UK

*Corresponding author. Tel. +49-22979-80142; fax: +49-33979-80200; e-mail: staubach@wus.bfav.de

Abstract

The analysis of the geographical distribution of disease on the scale of geographic areas such as administrative boundaries plays an important role in veterinary epidemiology. Prevalence estimates of wildlife population surveys are often based on regional count data generated by sampling animals shot by hunters. The observed disease rate per spatial unit is not a useful estimate of the underlying disease prevalence due to different sample sizes and spatial dependencies between neighbouring areas. Therefore, it is necessary to account for extra-sample variation and and spatial correlation in the data to produce more accurate maps of disease incidence. For this purpose a hierarchical Bayesian model in which structured and un-structured overdispersion is modelled explicitly in terms of spatial and non-spatial components was implemented by Markov Chain Monte Carlo methods. The model was empirically compared with the results of the non-spatial beta-binomial model using surveillance data of Pseudorabies virus infections of wildboars in the Federal State of Brandenburg, Germany.

Key words: spatial; binomial data; full Bayesian model; Markov Chain Monte Carlo methods; beta-binomial model; areal data

1 Introduction

In geographical epidemiology spatial data are frequently mapped on the basis of artificial boundaries such as administrative units. In wildlife population surveys, administrative structures are often the only feasible way to map samples. Just in a few studies each case and control is connected to exact coordinates (Kitron *et al.*, 1991; Staubach *et al.*, 2001). One of the most popular map types in geographical epidemiology is the choroplethic map. Choroplethic maps assign a shading or colour to geographic areas (defined by their boundaries) to visualize the variable of interest. The assigned hatching pattern or colour is based on a class interval or continuous scale deduced from a descriptive statistic of the aggregated data. Especially in situations where the sampling size per spatial unit is small, dot maps and proportional symbolic maps may be more helpful for visualization. Nevertheless, all manipulation and analysis relies fundamentally on the given set of zonal units, and this cannot be overcome without access to individual data records (Gatrell, 1994).

The simple classification into different prevalence ranges that is frequently applied causes some problems: (i) Spatial boundaries are artificially chosen and not relevant to disease spread. (ii) The sample size is often not taken into consideration when spatial data are presented. Therefore, confidence limits may overlap with those of neighbouring prevalence ranges and prevalence differences between neighbouring units could thus be random. (iii) When data are stratified, sample sizes in some units or strata may be too low to obtain reliable prevalence estimates. (iv) Mapping surveillance data in this manner may also lead to false interpretations of disease clusters or disease-free areas. Furthermore, all relationships observed between variables will only hold for this particular aggregation of the data, a phenomenon which is well-known as ecological fallacy (Fotheringham and Wong, 1991; Fotheringham and Rogerson, 1993; Pfeiffer and Morris, 1994; Smans and Estéve, 1996;

Haining, 1998).

Particularly in medical epidemiology, different spatial filters, smoothing methods and parametric regression techniques have been suggested for the solution of these problems (Pfeiffer and Morris, 1994; Elliott et al., 1995; Haining, 1998). They do frequently not regard all sampled informations as parameters of a hypergeometric distribution (Elliott et al., 1995; Smans and Estéve, 1996). A full Bayesian model for a spatial analysis of disease prevalence data based on a spatial smoothing prior was implemented using modern Markov Chain Monte Carlo (MCMC) methods. By using surveillance data of Pseudorabies virus (PRV) infections of wildboars we compared the diagnostic results with the prevalence estimates of the MCMC and the non-spatial beta-binomial model.

2 Material and Methods

2.1 Disease data

The study area comprised of the Federal State Brandenburg in the eastern part of Germany and covers approximately 29,530 km^2. The Federal State of Brandenburg is divided into 1700 administrative units (municipalities) with an average area of 17.4 km^2. Municipalities may have enclaves with the same identification number which are not directly spatially linked to the main administrative unit. The topographical map consists of a total of 1902 geographic areas.

Disease data consist of the numbers of diagnosed positive and negative results directly linked in the GIS to the spatially defined administrative units. If enclaves of municipalities existed, the positive and negative results were subdivided proportionally to the area of each unit.

The data base, sampling frame and investigation procedure of the disease data is described

elsewhere (Müller *et al.*, 1998) . We used summarized surveillance data of PRV infections of wildboar based on a serological survey of the year 1993 as example. In 370 spatial units 1364 shot wildboars were examined (Figure 1). 119 animals were serologically positive in the diagnostic test (full-antigen ELISA) resulting in prevalences ranging between 0.0 and 1.0 for each spatial unit.

2.2 Statistical Models

Within a map of N spatial units, let n_i denote the sample size and y_i the number of positive results in the ith region. The observed prevalence for each spatial unit is given by the ratio $p_i = y_i/n_i$. However, note that often no prevalence estimate is available due to the lack of data ($n_i = 0$).

Throughout we assume that the number of cases in each spatial unit is binomial distributed with $y_i \sim Binomial(n_i, \pi_i)$, where π_i represents the unknown true disease prevalence.

2.2.1 Beta-binomial Model

To model only the unstructured overdispersion of the data we use an empirical Bayesian approach based on the beta-binomial model. This model is often fitted to binary response data which display a larger variance than that expected under a binomial model (Williams, 1975; Gelman *et al.*, 2000). The prevalence in each spatial unit is assumed to be independently beta distributed, i.e. $\pi_i \sim Beta(\alpha, \beta)$, where α and β are fixed, but unknown hyperparameters. These hyperparameters were estimated by maximum likelihood techniques (Smith, 1983).

The beta distribution has mean $\alpha/(\alpha+\beta)$ and density

$$p(\pi_i) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi_i^{\alpha-1}(1-\pi_i)^{\beta-1}. \tag{1}$$

Applying Bayes' theorem, i.e. multiplying the prior distribution (1) with the binomial likelihood

$$p(y_i|\pi_i) = \binom{n_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i} \tag{2}$$

yields the posterior distribution, which is (due to conjugacy) again beta distributed: $\pi_i|y_i \sim Beta(\alpha+y_i, \beta+n_i-y_i)$. In particular, the posterior mean of the true prevalence is $(\alpha+y_i)/(\alpha+\beta+n_i)$. We will use the posterior mean as a point estimate of the true disease prevalence, alternative choices include the posterior mode $(\alpha+y_i-1)/(\alpha+\beta+n_i-2)$ and the posterior median, which can easily be calculated numerically.

The posterior variance

$$var(\pi_i|y_i) = \frac{(\alpha+y_i)(\beta+n_i-y_i)}{(\alpha+\beta+n_i)^2(\alpha+\beta+n_i+1)}. \tag{3}$$

can be used to calculate approximate credible intervals for the underlying true prevalence in each spatial unit. However, exact credible intervals, based on the quantiles of the beta distribution, are easily available and should be preferred. Similarly, the posterior probability that the prevalence π_i in the ith unit exceeds the overall prevalence $\sum y_i / \sum n_i$ can be computed and mapped in order to visualize the "significance" of the point estimate.

2.2.2 Hierarchical Bayesian Model

For the full Bayesian approach we adopt a commonly used model from spatial epidemiology (Besag *et al.*, 1991) to the binomial observation model. In contrast to the beta-binomial model, this formulation incorporates the spatial structure of the data and essentially smoothes the observed prevalences. Furthermore, for spatial units without any data,

the model is able to interpolate the prevalence surface. An underlying spatial smoothing parameter is treated as unknown and estimated from the data.

In the spirit of Besag *et al.* (1991) (replacing the log-linear Poisson with a binomial logistic model), we decompose the log-odds of π_i into an intercept μ and two heterogeneity parameter, one displaying spatial structure (u_i) the other displaying unstructured heterogeneity (v_i) a priori:

$$\log \frac{\pi_i}{1-\pi_i} = \eta_i = \mu + u_i + v_i \tag{4}$$

For the spatially structured parameter u_i we adopt a so-called Gaussian Markov random field distribution, also known under the name "intrinsic autoregression". More specifically, we assume that the conditional distribution of u_i, given $u_j, j \neq i$, is normal with mean equal to the average of the u_j's in neighbouring units and variance equal to an unknown variance parameter σ_u^2 divided by the number of neighbouring regions of unit i. Spatial units are considered as adjacent if they share a common boundary.

The unstructured effects v_i are assumed to be a priori independent with mean zero and variance σ_v^2. Note that, due to prior independence of the $v_i's$, districts without any data ($n_i = 0$) can provide no information about the unstructured component. Let I be the number of spatial units with data, and assume that the units are ordered with respect to n_i (in decreasing order). To avoid potential identifiability problems, we set the non-spatial heterogeneity parameter for the $N - I$ districts without any data to zero: $v_i = 0$, $i = I + 1, \ldots, N$.

For both the spatial and the non-spatial heterogeneity variance parameters σ_u^2 and σ_v^2 we will use priors which favour a nearly constant prevalence pattern, with a high prior mass on very small values. However, the (inverse gamma) priors used are highly dispersed, hence the formulation will be flexible enough to capture heterogeneity if there is evidence in the

data for it. Note that σ_u^2 has the role of a spatial smoothing parameter, determining the variability of the spatial heterogeneity component u.

Statistical inference in this model is only feasible using modern MCMC simulation techniques, see for example Gelman *et al.* (2000). It is computationally convenient to reparameterize the model from v_i to $\eta_i = \mu + u_i + v_i$ (Besag *et al.*, 1995). This has the advantage that the full conditional distribution for u is multivariate Gaussian. We can therefore block update the vector u efficiently using an algorithm described in Rue (2001). For sampling η_i we use a Metropolis step with a Gaussian proposal, which was tuned in order to get acceptance rates between 35% to 45%. The full conditional distributions of the precision parameters are again inverse gamma distributed and can be therefore sampled in a Gibbs step. More details on computational issues in hierarchical models with underlying Markov random field components can be found in Knorr-Held and Rue (2001).

MCMC techniques generate samples from the posterior distribution of the π_i's, from which posterior characteristics such as quantiles can be estimated. Posterior probabilities of an exceedence prevalence can be calculated, similar as in the beta-binomial model.

3 Application

We now present an empirical comparison of the two methods for data on Pseudorabies virus infections in Brandenburg. Figure 2 displays the observed prevalences y_i/n_i while Figure 3 gives posterior mean estimates based on the non-spatial beta-binomial model. A considerable shrinkage of the observed disease prevalences towards the overall prevalence can be seen. In fact, estimates based on the beta-binomial model lie always between the sample proportion, y_i/n_i, and the prior mean, $\alpha/(\alpha + \beta)$ (Gelman *et al.*, 2000). Note that no estimates are available for units without any data. Finally, results from the full Bayesian model are

displayed in Figure 4. The spatial structure is now more apparent with a rather constant disease prevalence throughout the area considered with slightly increased prevalence in the east of Brandenburg. Most notably, the full Bayesian formulation does find considerably less evidence for spatial variability compared to the naive estimates and the empirical Bayes estimates in Figure 2 and 3. It seems that the incorporation of the spatial structure of the data, together with the large uncertainty about disease prevalence (due to many areal units without any data) implies a rather strong smoothing effect of the observed prevalences. Figure 5 and 6 map the posterior probabilities for a prevalence above the overall prevalence from the two different models. The maps show that the greatest evidence for an increased disease prevalence can be found in the eastern part of Brandenburg. This may support the hypothesis that the epidemic in Brandenburg starts at locations near to the border to Poland (Müller *et al.*, 1998).

Incidentally, although the beta-binomial model is non-spatial, there is some agreement of the posterior probabilities from that model (for the 370 regions with data) with the corresponding ones obtained from the hierarchical Bayesian model, with an empirical correlation of 0.549. See also Figure 7, which compares the two estimates in a scatter plot. The difference between the estimates from the two models can be explained by the additional spatial component in the hierarchical Bayesian formulation.

4 Discussion

A large number of epidemiological studies utilize explorative spatial data analysis to describe geographical distributions of very different types (e.g. virological and serological prevalences, incidences, biological marker proportions). Health data are often collected at the scale of geographic areas, because even the step from the smallest administrative unit to exact

coordinates of the sampled animal is enormous and only in a few cases necessary, (e.g. for spatial analysis of habitat and microclimate; Kitron *et al.*, 1991; Staubach *et al.*, 2001). Therefore, the data are often mapped independently for each spatial unit. In spite of efforts to reach a sample size as large as possible, the sampling sizes per spatial unit are often very low (Figure 1).

The possibility to calculate and map Bayesian posterior probabilities is important to assess the significance of the prevalence estimates on a small-area level and to judge the geographical variation of the disease.

An advantage of the full Bayesian model in comparison to the beta-binomial or mixture models is the possibility to estimate prevalences also for spatial units with missing data. To aggregate the data on a higher spatial level (e.g. districts or countries) is often not feasible, because this reduces the sampled spatial information and may lead to false interpretations of disease clusters or disease-free areas (Schlüter and Müller, 1995; Tackmann *et al.*, 1998). On the lower spatial level (e.g. municipalities, counties) regions without any sample are often recorded due to logistic and administrative problems.

Time-consuming large scale simulation studies are the only feasible way to further explore the statistical properties of the different models for spatial binomial data in detail (Lawson *et al.*, 2000). Nevertheless, the examination of different data sets from field with different statistical properties illustrates that the full Bayesian model may be useful as a first step of descriptive and explorative spatial data analysis. The final map displays a more adequate data representation then the raw prevalence estimate, especially if the sampling frame is sparse - as it is often the case in data sets from wildlife population surveys. Of course, for the final interpretation of the maps and hyperparameter values, it is necessary to consider the character of the diseases, as e.g. contagious, environmental, vector-borne and parasitic.

From a methodological point of view, efficient MCMC algorithms for the analysis of

disease prevalence data can be challenging to implement, due to the sparseness with many spatial units without any data. We were able to avoid some of the problems using a block update of the underlying Markov random field u. However, Knorr-Held and Rue (2001) show in the disease mapping context that a joint update of u and v, preferably together with the corresponding variance parameters, is possible and might in fact be necessary to decrease the simulation error of the estimates. We are currently investigating the applicability of such algorithms to the binomial setup.

A potential disadvantage of the Markov random field approach for spatial smoothing is that it assumes a priori, that the degree of spatial smoothness is constant over the whole study area. More adaptive smoothing methods have recently been proposed in Knorr-Held and Raßer (2000) and Denison and Holmes (2001). However, the applicability of such models to the binomial set-up is beyond the scope of this paper.

Acknowledgments

We thank Thomas Müller for making available the data sets utilized by this study and Hartmut Schlüter for support.

References

Besag, J.E., Green, P.J., Higdon, D.M. and Mengersen, K.L., 1995. Bayesian computation and stochastic systems (with discussion). *Statistical Science*, **10**, 3-66.

Besag, J.E., York, J.C. and Mollié, A., 1991. Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1-59.

Denison, D. and Holmes, C., 2001. Bayesian partitioning for estimating disease risk. *Biometrics*, **57**, 143-149.

Elliott, P., Martuzzi, M. and Shaddick, G., 1995. Spatial statistical methods in environmental epidemiology: a critique. *Statistical Methods in Medical Research*, **4**, 137-159.

Fotheringham, A.S., Wong, D.W.S., 1991. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, **23**, 1025-1034.

Fotheringham, A.S., Rogerson, P.A., 1993. GIS and spatial analytical problems. *Int. J. Geographical Information Systems*, **7**, 3-19.

Gatrell, A.C., 1994. Density estimation and the visualization of point patterns. In: Hernshaw, H.M., Unwin, D.J. (Eds.), *Visualization in Geographical Information Systems*, Wiley, Chichester, pp. 66-75.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2000. *Bayesian data analysis.* Boca Raton, Chapman and Hall/CRC, 526p.

Haining, R., 1998. Spatial statistics and the analysis of health data. In: Gattrell, A.C., Löytönen, M. (Eds.), *GIS and Health, GISDATA VI*, London, Taylor and Francis, 29-47.

Kitron, U., Bouseman, J.K. and Jones, C.J., 1991. Use of ARC/INFO GIS to study the distribution of Lyme Disease ticks in an Illinois county. *Prev. Vet. Med.*, **11**, 243-248.

Knorr–Held, L. and Raßer, G., 2000. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, **56**, 13-21.

Knorr-Held, L. and Rue, H., 2001. On block updating in Markov random field models for disease mapping. Revised for *Scandinavian Journal of Statistics.*

Lawson, A.B., Biggeri, A.B., Boehning, D., Lesaffre, E., Viel. J.-F., Clark. A., Schlattmann, P., Divino, F., 2000. Disease mapping models: an empirical evaluation. *Statist. Med.*, **19**, 2217-2241.

Müller T., Teuffert J., Ziedler K., Possardt C., Kramer M., Staubach C., Conraths F.J., 1997. Pseudorabies virus infections of the European Wildboar from Eastern Germany. *Journal of Wildlife Diseases*, **34**, 251-258.

Pfeiffer, D.U., Morris, R.S., 1994. Spatial analysis techniques in veterinary epidemiology. *The Kenya Veterinarian*, **18**, 483-485.

Rue, H., 2001. Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society*, Series B, **63**, 325-338.

Smans, M., Estéve, J., 1996. Practical approaches to disease mapping. In: Elliot P., Cuzick, J., English, D., Stern, R. (Eds.), *Geographical and Environmental epidemiology*, Oxford, Oxford University Press, 141-150.

Schlüter, H., Müller, T., 1995. Tollwutbekämpfung in Deutschland. Ergebnisse und Schlußfolgerungen aus über 10jähriger Bekämpfung. *Tierärztl. Umschau*, **50**, 748-758.

Smith, D.M., 1983. Maximum likelihood estimation of the parameters of the beta binomial distribution - algorithm AS 189. *Applied Statistics*, **32**, 196-204.

Staubach, C., Tackmann, K., Thulke, H.-H., Hugh-Jones, H. Conraths, F.J., 2001. Geographical information system-aided analysis of factors potentially influencing the spatial distribution of Echinococcus multilocularis infections of foxes. *The American Journal of Tropical Medicine and Hygiene* (in press).

Tackmann, K., Löschner, U., Mix, H., Staubach, C., Thulke, H.-H., Conraths, F.J., 1998. Spatial distribution patterns of *Echinococcus multilocularis* (Leuckart 1863) (Cestoda: Cyclophyllidea: Taniidae) among red foxes in an endemic focus in Brandenburg (Germany). *Epidemiology and Infection*, **120**, 101-109.

Williams, D.A., 1975. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, **31**, 949-952.

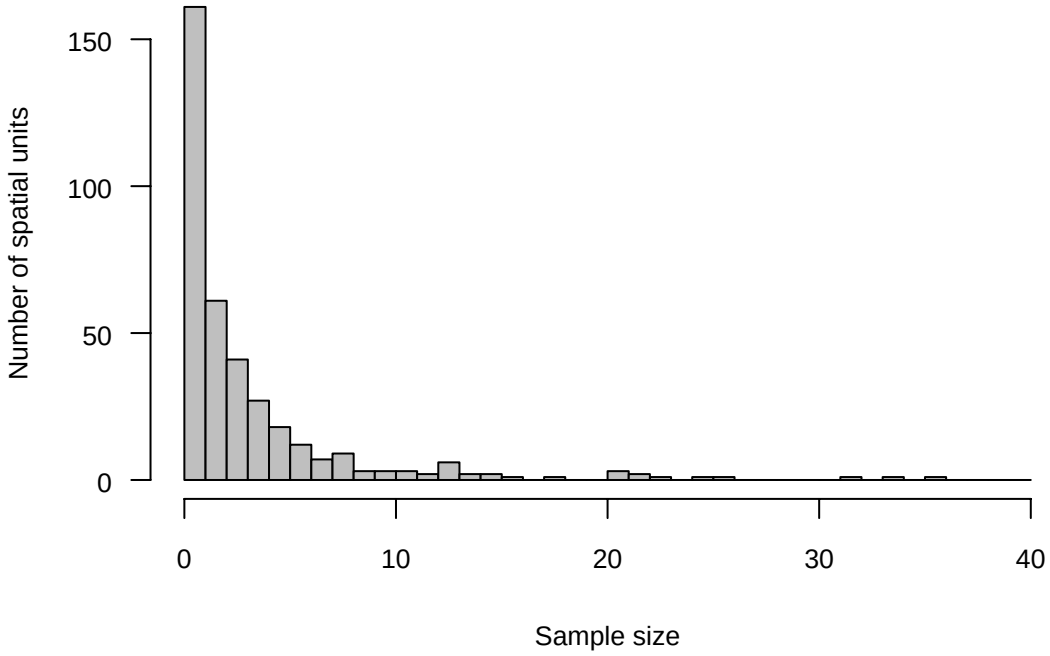

Figure 1: Distribution of the sample size of examined wildboars for PRV per spatial unit

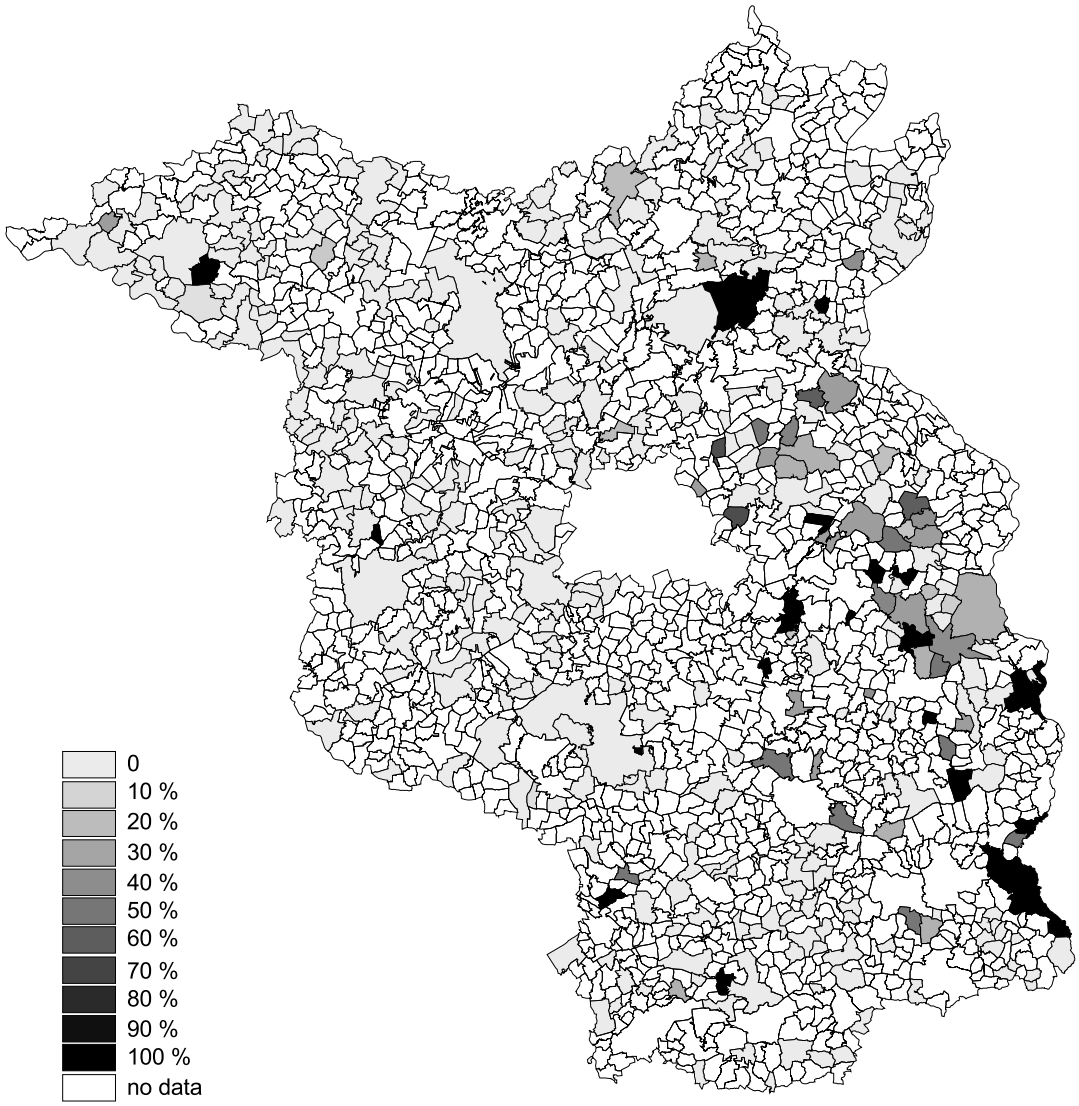

Figure 2: Observed prevalences of the PRV infections of wild boars in Brandenburg based on a serological survey of the year 1993

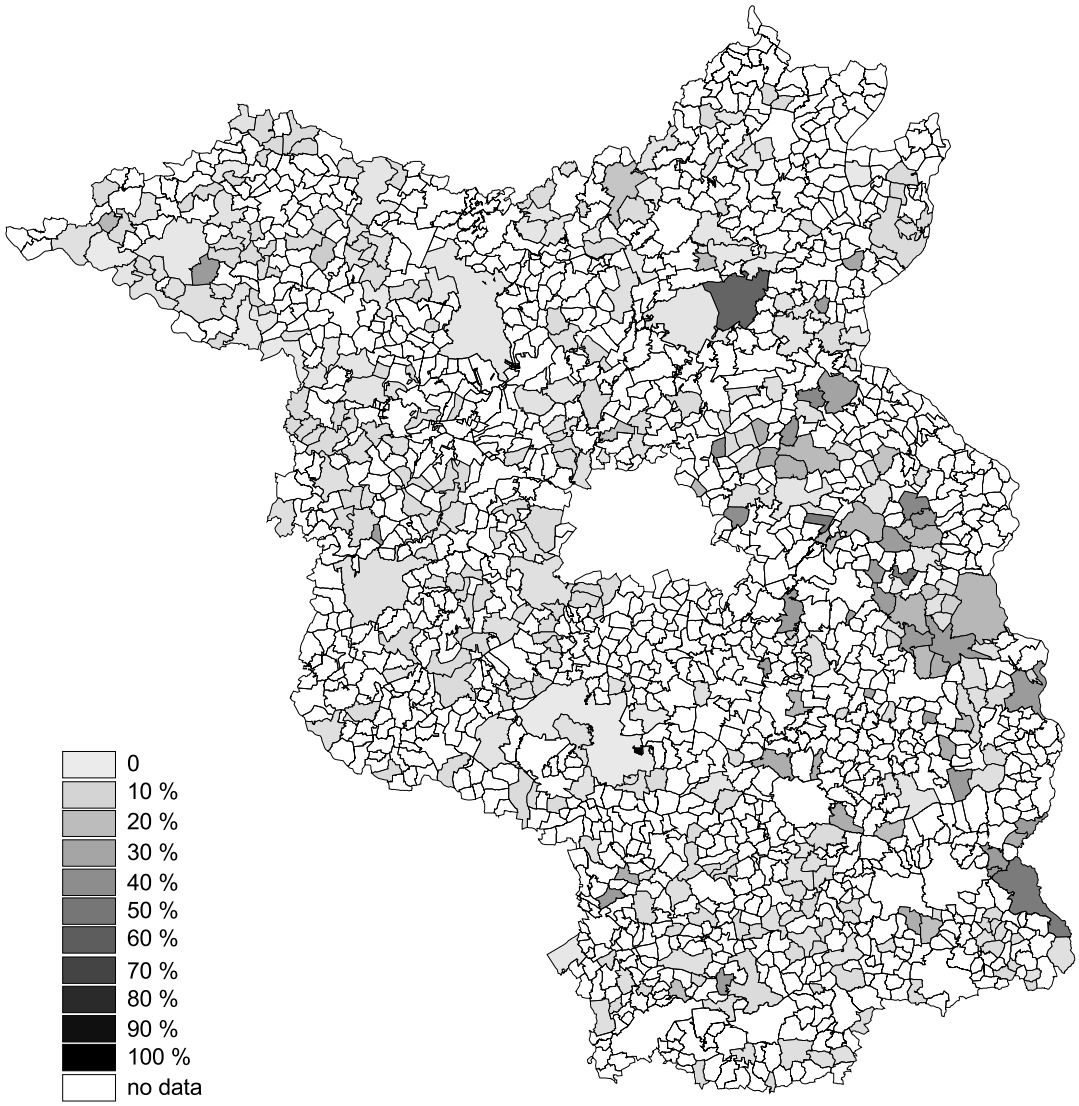

Figure 3: Prevalence estimate of the PRV infections of wild boars using the beta-binomial model

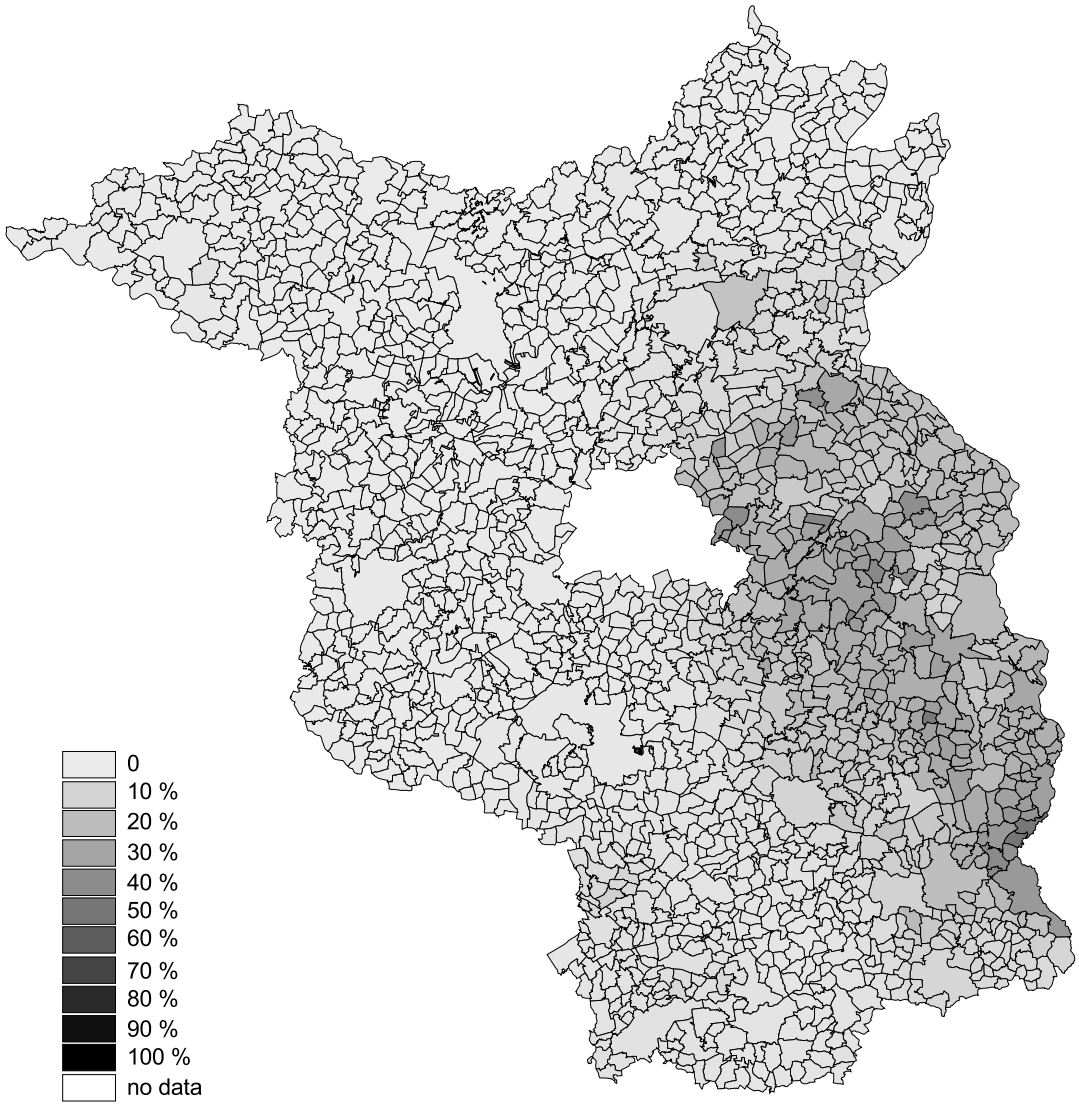

Figure 4: Estimated median prevalences for PRV infections of wild boars using the full Bayesian model

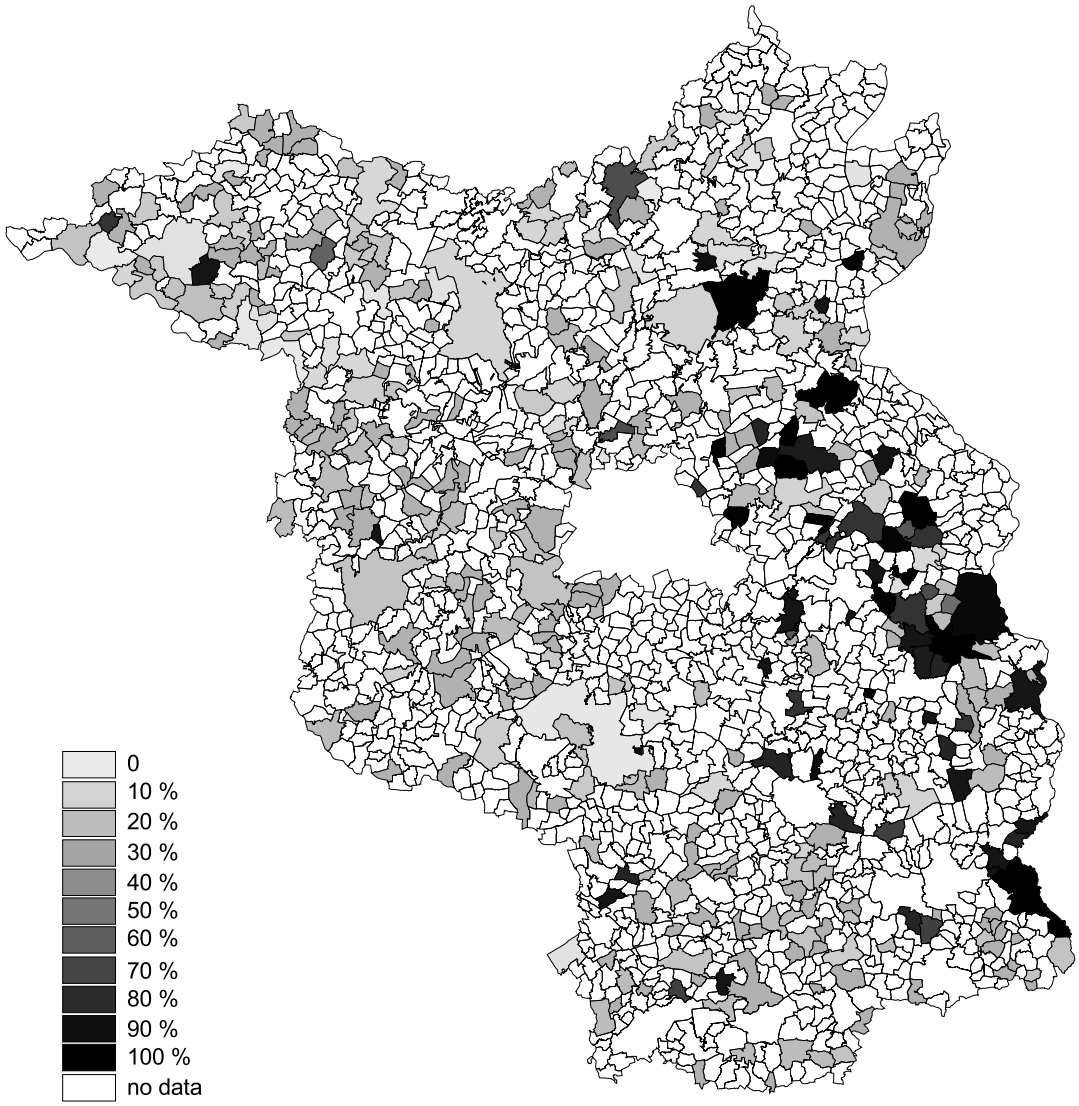

Figure 5: Posterior probablilities of a prevalence above the overall prevalence for PRV infections of wild boars using the beta-binomial model

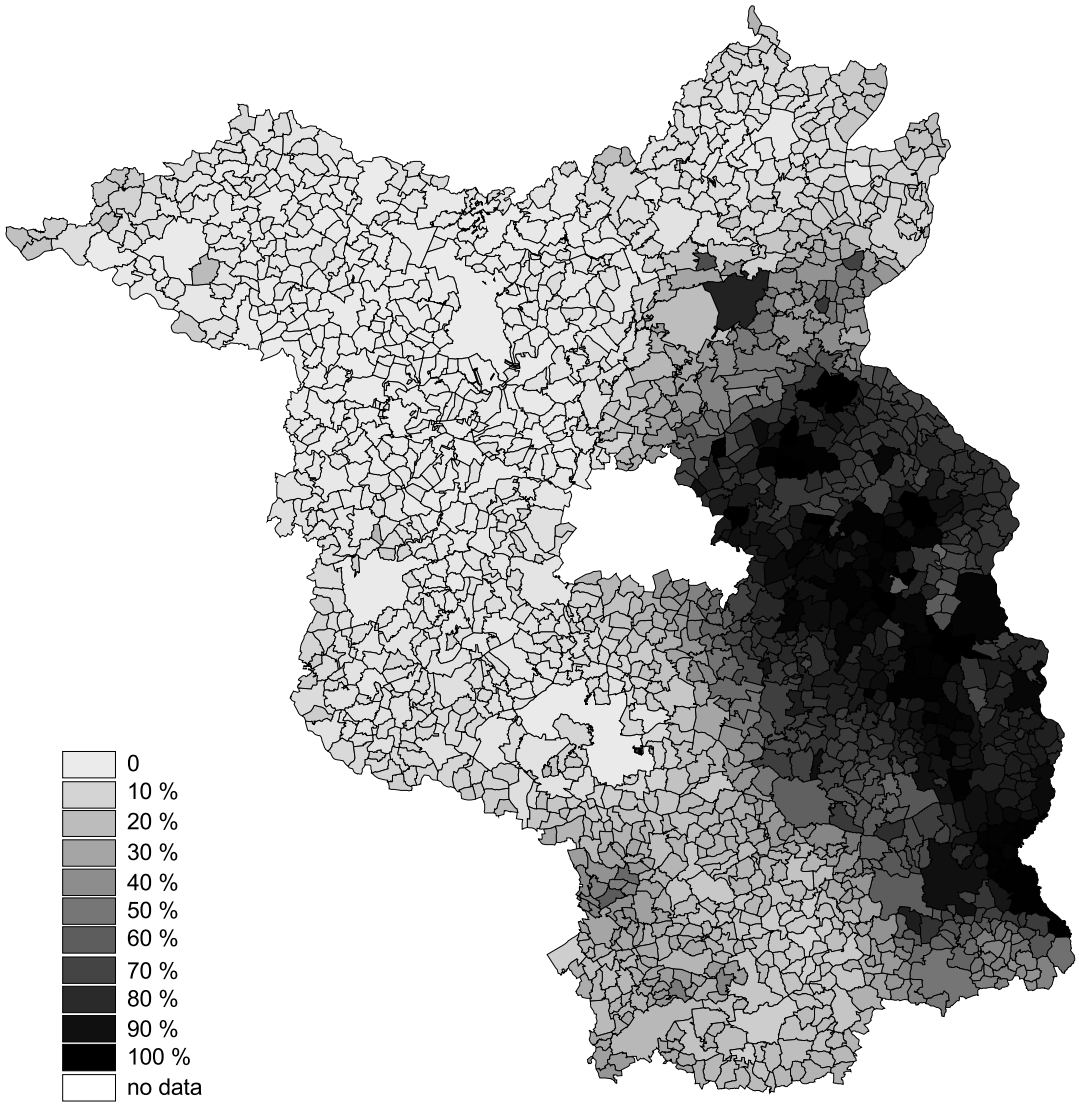

Figure 6: Posterior probablilities of a prevalence above the overall prevalence for PRV infections of wild boars using the full Bayesian model

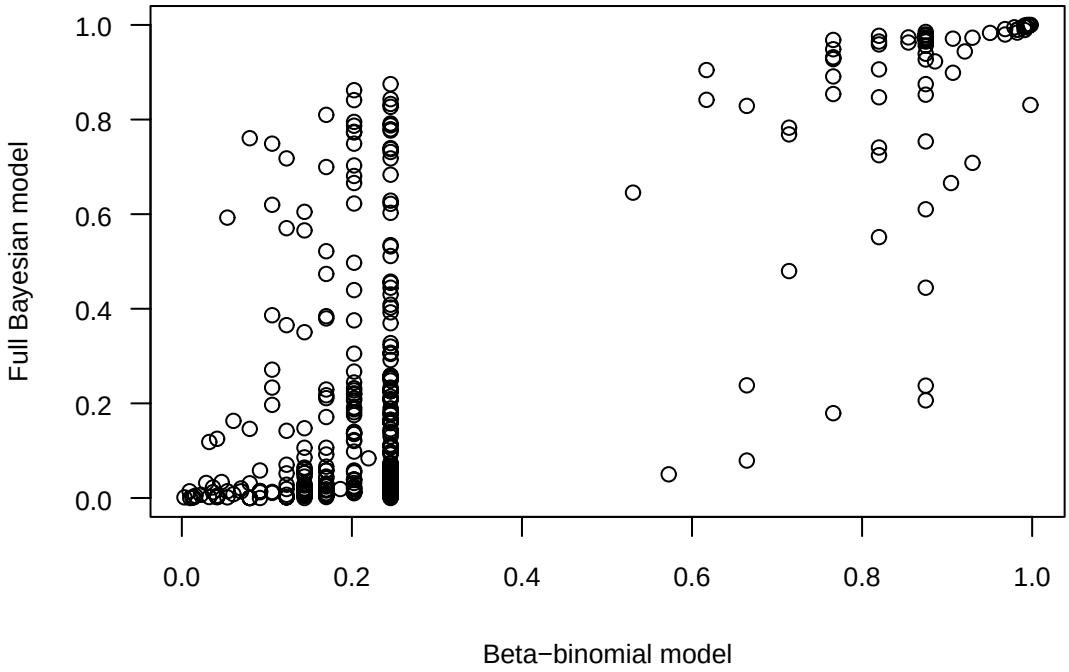

Figure 7: Graphical comparison of the estimated posterior probabilities obtained from the two models