**Reviewer's report**

**Title:** Agricultural, socioeconomic and environmental risk factors for Verotoxigenic Escherichia coli (VTEC) infection in humans in Finland

**Version:** 1  **Date:** 27 January 2011

**Reviewer:** Michael Höhle

**Reviewer's report:**

**Summary:**

The authors investigate the effect of environmental, agricultural and socioeconomic variables on the risk of VTEC in the 416 municipalities of Finland. Statistically, this ecological regression is performed using a disease mapping model taking the large number of municipalities with zero cases into account while addressing spatial dependence between the municipalities. Altogether, this constitutes a very advanced but necessary model where inference is done in a Bayesian framework using Markov Chain Monte Carlo methods. A number of the risk factors are then identified as being significant using model selection, of which “Fresh water per total surface area” is the most interesting result. The paper is from a statistical perspective sound and advanced, but regrettably the epidemiological perspective of the paper is currently underrepresented.

**Major Compulsory Revisions:**

Currently, the writing style of the manuscript is too statistical for the intended audience of BMC Infectious diseases. This already starts in the background section on p.2 where no epidemiological research questions and hypotheses are formulated. The statistical model description should be removed here. The same applies to the conclusions: also here the epidemiological findings should be dominant. Altogether the paper needs substantial rewriting to better address the epidemiological objectives and findings of the work. As an example: bulls are found to be a risk factor while milk cows are not. Is there a difference in Finland on how bulls and milk cows are kept, e.g. bulls spend more time outside than milk cows? Such substance matter interpretations are regrettably missing.

As another major point: The statistical model should be introduced as an “ecological regression approach” where the modelling unit is the municipality. Currently, descriptive results on the individual cases are mixed with results for municipalities - this makes the overall objective of the manuscript unclear. Furthermore, in the modelling, an entire raster of possible covariates is investigated. Here, a more biological plausible strategy of which covariates to take is missed. Especially, since many of the covariates appear to greatly correlate with each other. The manuscript says, that of the correlated variables
the most significant one was included in the model, but I am unsure exactly what strategy was done to investigate this? Was it the result from the univariate analysis or was from a multivariable approach including both variables?

To more specifically put light on the correlation issue, please report correlation coefficients between the variables and then a selection of which variable to include based on epidemiological reasoning, possibly substantiated by the results from univariate modelling.

The performed modelling is inspired by disease mapping for chronical diseases, with the tweak that zero-inflation needs to be handled as a result of VTEC being a very rare disease. Because the disease is so rare, it is also ultimate to honour the count data nature (please make the count data aspect clearer before talking about the zero inflation). Currently, the clog-clog hurdle model is more or less taken as given as the solution to the zero-inflation aspect. Thus, I miss a discussion of important issues such as:

1) Why is it necessary to consider the exact distribution of >0 areas, i.e. why not just have a binary model for the indicator variable \( z_i = I(y_i>0) \) (i.e. the zero part) and no additional part. Are the additional gains worth the effort?

2) Why use a cloglog & Poisson truncation? Other approaches exist to catch zero-inflation. What are the specific advantages? Purely availability of an implementation using WinBugs? This comment especially applies to the motivation on p.5.

3) How exactly are the model parameter estimates to be interpreted? This might be easy/standard for the zero-component (standard cloglog, but in an epidemiological model used to logit modelling possibly less obvious) but not so easy in the truncated Poisson. I miss an epidemiological statement like “for every additional bull per XXX persons in the municipality, the risk increases by factor YYY. One problem is presumably, that such an interpretation in the hurdle model requires some kind of conditioning when talking about parameters in the truncated Poisson part. Like: IF there are case, then every additional bull per XXX leads to…. Furthermore, is it possible to combine the parameter values for bulls from the two components into one interpretable value? I think this would be of interest in the target audience.

4) What about interactions between the covariates. Why are none of these investigated?

5) VTEC cases are not chronic but infectious. Those additional clustering of the disease is to be expected. However, the authors use the Poisson distribution without overdispersion. Thus, clustering due to the infectious character becomes an issue which currently is not handled. This could be addressed in several ways: a) The spatial model contains an additional unstructured random effect (c.f. BYM model) and thus the marginal distribution allows for greater variation for s municipalities. B) use a truncated NegativeBinomial response instead of Poisson (more or less the same as (a)). C) include only the first case of each cluster. At least one of the solutions should be investigated and the issue discussed as it represents one fundamental difference between chronic disease modelling and infectious disease modelling using purely spatial models.
A recommendation in the conclusions on p.3 is that a “full likelihood approach” is recommended. However, the analysis performed is Bayesian and thus I think the recommendation should be that a “Bayesian framework” is recommended.

Minor Essential Revisions:

p.2, Results: The average annual … was 4.8 cases per million…
was estimated to be 4.8 (95% CI XX.X – YY.Y)
I expect that this is an overall raw estimate not adjusting for any covariates?

p.8, end of 3rd paragraph: missingness in the data..
What exactly is missing (covariate information for some of the municipalities?). Possibly, table 1 could state how many missings there are for each covariate.

p.10, 3rd paragraph
Here, descriptive results of the questionnaires are given. However, as the regression model is of ecological nature, the answer of the questionnaires do not play any role in the analysis. Thus a motivation why these data are stated lacks. Has this been done in order to identify possible risk factors? If so, what is the background rate for some of the risk factors (e.g. swimming in natural water) in the background populations. A clearer presentation is needed here.

p.11, 3rd paragraph: In the multivariate model
What exactly is meant by multivariate? If it is the inclusion of several risk factors/covariates, I would prefer the term “multivariable”. Where are the results of the univariable modelling stated?

P.11, line -2
The two influential municipalities: are these outliers in the covariates or in the response (large number of cases)?

p.16, lines 1-3
Such data would be needed not for the cases, but for the municipality, since an ecological regression is performed. I’m not sure such data actually exist?

p.16, 2nd paragraph, line 2: models designed for scarce data. The seasonal trends…
In my opinion it was also compensated by performing a purely SPATIAL analysis and no space-time analysis which could have revealed trends in the data. Seasonal trends would be a further complication as this in my understanding of the word “seasonality” happens within the year (thus a monthly or weekly analysis would be needed).

Appendix: Models of \( \lambda_i \) and \( \pi_{0i} \)
With the used representation \( \log(E_i) \) should enter the linear predictor, not \( E_i \) itself. Furthermore, in the text after the two equations, one should first explain the index \( I \), then \( E_i \).

Appendix: Standardisation

Currently, standardization is internal by population size. Since mainly kids are affected by VTEC, I wonder if the Finish population structure necessitates an age standardization?

Appendix: MCMC convergence

The appendix contains valuable information about the statistical aspects of the model. Since MCMC was used for the inference I miss a statement about convergence diagnostics. As part of this, it would be interesting to see a plot in the appendix of \( (\alpha_k^j, \beta_k^j) \) over the samples \( j \) for a covariate like bulls which enters in both components. This would help to gain insight about the identifiability of effects in the two components. Also the correlation between \( b_i \) and \( c_i \) would be of interest to obtain a statements about the identifiability of the model. Furthermore, it should be written somewhere that \( b_i \) and \( c_i \) are assumed to be independent (the vector \( b \) or \( c \) is spatially dependent).

Table 2

I expect the intervals to be 95% HPD intervals? Please state so.

Discretionary Revisions:

p.2, line -4, ...per human population
...per human population in the municipality

p.9 statistical packages
I guess ArcGIS was used for data pre-processing not for the actual statistical analysis?

p.9, line -5: 4.8 (1.9-12.0)
4.8 (95% CI 1.9-12.0). Or is it an HPD?

Fig. 1 & Fig.2
In the map I miss a northing arrow, a scale bar to get some idea about distance and a unit for the incidence rate (yearly incidence rate per 1,000,000?). Furthermore, for those not too familiar with the Finnish geography it might be of help to add e.g. the 5 major cities to the map. The same comments apply to Figure 2.

Appendix: Upload a PDF file, not a DVI file.

Appendix: I’m not a fan of having to email authors to get code. In case of
publication make the code available as part of an web appendix. Even better: would it be possible to also provide the data for download? There could be confidentially issues here, but would greaten strength a) availability of some benchmark data sets for spatial analysis in infectious disease epidemiology and b) reproducible research.

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests