Modeling Spatial Effects of the Spread of Malaria in Gambia using SPDE's and R-INLA

Jucoen Yeater
Department of Mathematics
Purdue University Fort Wayne
Fort Wayne, IN 46805
yeatjr01@pfw.edu

Abstract

This study will look to model the different spatial locations of the prevalence of Malaria in Gambia to see if the values differ across those locations. This was approached by use of stochastic partial differential equations (SPDEs). These SPDEs will be modeled using the R-INLA package. The integrated nested Laplace approximation (INLA) method provides faster computation time than the Markov chain Monte Carlo (MCMC) methods. This is vital when using large or complex datasets as they can cause long computational times. It was found that neither the greenness of a particular village nor if the village was in a primary health care system had a significant affect on the prevalence of malaria within a village. Another model was also ran without the spatial effect which returned a higher Deviance information criterion (DIC) from which it can be determined that the data supports the presence of a spatial effect.

1 Introduction

Being able to model the spread or to test for spatial effects of a disease such as Malaria can help show what locations need to be improved so that less individuals will become infected. The model can also be used to predict further spread of the disease. The approach used in this study will use stochasitc partial differential equations (SPDEs) to estimate the model. The model used will describe the spatial variation in the prevalence of malaria in Gambia. The dataset used is *gambia* from the *geoR* package.

The study will use the R package R-INLA and is based on the integrated nested Laplace approximation (INLA) method. One reason to use INLA is because the INLA method can provide faster computation than Markov chain Monte Carlo (MCMC) methods. INLA relies on a combination of analytical approximations and effecient numerical integration schemes to achieve highly accurate deterministic approximations to posterior quantities in interest [1]. INLA will be further expanded in the following sections.

1.1 Background/Literature Review

The MCMC method was one of the first Bayesian methods used for applications in real case studies. Before the devolopment of MCMC, Bayesian methods were mainly used for theortical models. A potentially huge problem, depending on the size of the dataset, is that MCMC can have large computational times. Recent years have shown that data collection is big business and as such data is being collected all of the time. This means that faster computational methods need to be developed

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

and evaluated to keep with the growing demand. One such method is INLA.

One such study that used SPDE with the INLA package was to predict the occurrence of spinetail devil rays [5]. The goal of the study was to be able to obtain an understanding of their spatiotemporal distributions so as to be able to protect them from further harm. The reasoning for using the SPDE approach was because the dataset was complex and using conventional simulation-based approaches were often found to be computationally intensive. This leads to one of the main reasons for the INLA package and in particular the building a model using SPDEs. It can save computational time with large or complex datasets. Although this can be at the cost of accuracy as explained later when building the mesh construction.

Gaussian random fields have become increasingly popular in the field of epidemiology. Data pertaining to epidemiology can be spatial or spatio-temporal [3]. Spatial data contains information about some location. Spatio-temporal data has characteristics of both location and time. To use the SPDE approach, as needed in this study, a discretely indexed spatial random process represents a continuous spatial process, a Gaussian field (GF). Starting with a SPDE as shown

$$(\kappa^2 - \Delta)^(\alpha/2)(\tau\xi(s)) = W(s)$$

with the solution to the SPDE being a Gaussian Field with a Matérn Covariance. Since the solution to the SPDE is a GF the INLA method can be applied.

1.2 Research Question

The study aims to test the spatial effects of the prevalence of Malaria in Gambia by using SPDEs and the INLA method. The study aims to answer if the spatial effects are significant when building the model to test for the prevalence of Malaria. The spatial characteristics of the data comes from the location data provided by the dataset. This *point-referenced data* comes in the form of a two-dimensional vector represented by a village location in Gambia.

2 Methods

There are eight variables accounting for 2035 children living in 65 villages in Gambia. The first two variables are the location coordinates of the village. There are three covariates contained at the child level which are the age of the child in days (age), the usage of bed nets (netuse), and whether the bed nets are treated with insecticide which can help to kill insects that may have malaria (treated). There are two covariatess at the village level which are the vegetation index of the village and immediate area surrounding the village (green) and whether or not the village is within a primary health care system (phc). The response variable is whether or not a child tests positive for malarial parasites in a blood sample and is given by a 1 if positive and a 0 if negative (pos).

The R-INLA package will be used to build a model that will describe the spatial variation of the confirmed cases of malaria. This allows for a Matérn GF and as such the SPDE approach can be used. The GF is approximated by the triangulation of the spatial domain which results in a mesh as shown in figure 1. The resutling mesh is dependent upon the number of triangles. A larger number of triangles used to build the mesh results in a more accurate approximation for the GF but at the risk of increased computation time. So there is an acceptable tradeoff that must be found between accuracy and computation time which can be dependent upon how large the dataset is.

The deviance information criterion (DIC) value will be used to measure the best for for the model. This will be done by comparing a model buillt with spatial effects and a model built without spatial effects. In particular, the model with the smaller DIC are better supported by the data. The DIC is calcuated by taking the sum of posterior expectation for the deviance and the complexity of

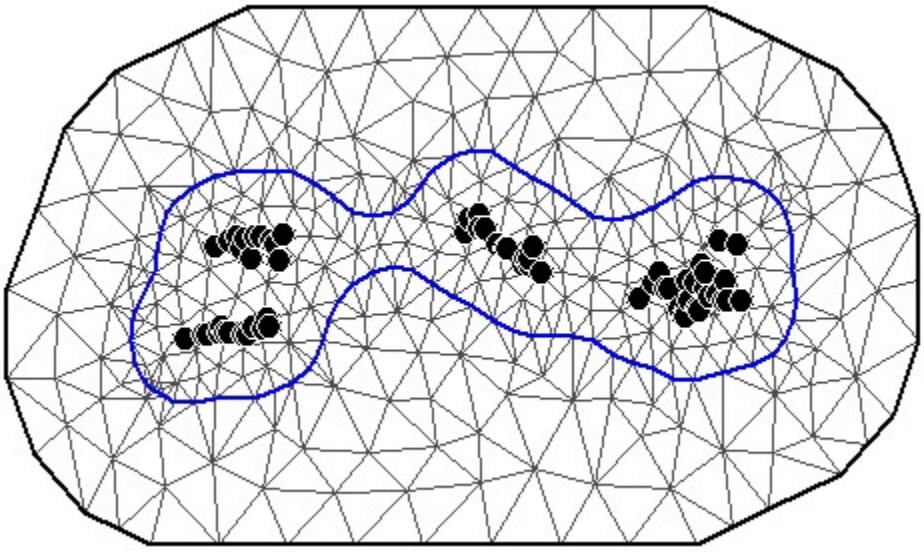

Figure 1: Triangulated mesh for villages in Gambia with internal line defining the nonconvex hull

the model which is derived from the effective number of parameters. This is further expanded upon in chapter 5 of reference [2].

3 Results

First, the posterior estimates can be evaluated to see what, if any, effects the covariates have on the response variable. Table 1 shows that as age increases the child is more likely to have the presence of malarial parasites in their blood. However, the use of bed nets, especially if the bed nets are treated with insectice, will reduce the likelihood of the child testing positive for malaria. At the village level the vegetation, or greenness of village and its immediate surroundings don't seem to have a significant effect on the presence of malaria in a child. If the village is in a primary health care system does seem to have a positive effect, but it is a somewhat limited effect.

To test for the spatial effects another model was ran without the spatial effects included. The Deviance information criterion (DIC) was then compared between the two models. The DIC value for the model without spatial effects is 2334.371 while the DIC value for the model with spatial effects is 2326.605. The DIC value for the model with spatial effects is a smaller value, thought not by much, which means that it is a better fit for the data. Therefore, the conclusion is that the data supports the presence of a spatial effect.

4 Discussion

The strengths will be that the spatial effects can be tested for to see if the location of a village within Gambia effects the prevalence of a patient having malaria. The weakness with the model is that the spread of the disease can't be measured over time, just the spatial effects. This can be further expanded upon when working on the COVID-19 dataset. Including a temporal component in the SPDE framework to model effects like the spread of the disease over time and how the spatial effects factor into that spread of the virus.

Another strength in relation to the Gambia dataset is the relatively small size of the dataset. There were only 2035 children that were included in this dataset. This means that computation time wasn't a real factor and allowed for the mesh to be build with a higher accuracy when compared to

Table 1: Posterior estimates for Gambia model with spatial effects included.

Parameter	Mean	SD	2.5%	50%	97.5%
b_0	-1.325	1.235	-3.695	-1.346	1.157
$\beta_{treated}$	-0.360	0.202	-0.759	-0.359	0.035
β_{netuse}	-0.374	0.158	-0.684	-0.374	-0.064
β_{age}	0.246	0.044	0.160	0.246	0.333
β_{green}	0.012	0.025	-0.038	0.012	0.059
β_{phc}	-0.310	0.231	-0.767	-0.310	0.142

what will be able to be done with a much larger dataset such as the COVID-19 dataset which could lead to a weakness. When using the COVID-19 dataset the mesh construction may need to be less fine at the cost of accuracy so that the computational costs aren't as high. This can be accounted for if looking at only one state such as Indiana, but if multiple states were to be looked at then the dataset can become large quite quickly. This is because dense matrix operations scale cubically with the matrix size because of the increase in the number of locations [2]. One possible way to overcome this could be to build a separate model for each state although it would be interesting to see if it would be possible to build a single model for the contiguous United States.

Now that results for the Gambia dataset have been successfully obtained research on the COVID-19 dataset can be continued using the SPDE approach. The first steps will be to clean the dataset so that a binary indicator of the presence of COVID-19 in a patient is included as the response variable. Then the ZIP code will need to be converted to longitude and latitude so that can then be converted to location coordinates to be able to obtain a euclidean distance between the counties of interest. Finally, the code used for the Gambia dataset can be adapted so that the spatial effects of the COVID-19 dataset can be tested for by looking at the posterior estimates as well as the DIC values for the two models. One including spatial effects and another not including them. This will be done after the final report is turned in for the Gambia dataset.

5 References and Citations

[1] Martino, S. & Riebler, A (2019) Integrated Nested Laplace Approximations. arXiv:1907:01248

[2] Blangiardo, M. & Cameletti, M. (2015) *Spatial and Spatio-temporal Bayesian Models with R-INLA* Wiley

[3] Blangiardo, M., Cameletti, M., Baio, G. & Rue, H (2013) Spatial and Spatio-Temporal Models with R-INLA. *Spatial and Spatio-Temporal Epidemiology 7*, pp. 39-55.

[4] Lindgren, F., Rue, H. & Lindström (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach (with discussion). *Journal of Royal Statistical Society Series B*, 73(4), pp. 423-498

[5] Lesama-Ochoa, N., Pennino, M., Hall, M., Lopez, J., & Murua, H. (2020) Using a Bayesian modelling approach (INLA-SPDE) to predict the occurrence of Spinetail Devil Ray (Mobular mobular). *Scientific Reports*, 10, Article Number: 18822

6 Work Division

Jucoen Yeater - All Work