Final Exam: STA465

The final exam is due on Monday, May 2nd at 11:59 p.m. EST. The final exam is worth 25 points in total.

Fitting Models

Fit the following five models to the malaria prevalence in The Gambia data set in INLA using the default priors + penalized complexity priors for those models that include the spatial random effect (a total of 7 models):

- Complete pooling and altitude (no random effects)
- Hierarchical random effect (iid) - (intercept only)
- Hierarchical random effect (iid) + altitude covariate
- Spatial + iid random effect
- Spatial + iid random effect + altitude covariate

Include all INLA code. For each model, compute the CPO, PIT values and create maps of mean predicted prevalence along with upper and lower limits of predicted prevalence. These maps should include predictions of prevalence across the entire country (as has been done in Homework 4). Comment on any major differences in predicted prevalence across models. Explain your choice of penalized complexity prior.

Spatial Residuals

For each model, compute (observed prevalence - mean predicted prevalence) and plot these values (on a map). There should be a total of 7 maps. Comment on any patterns you observe in and across the maps. For models that do not include a spatial component, are there any patterns you observe that would indicate the need for a spatial random effects term?

Results + PIT histograms

Organize the results of the estimates, 95% credible intervals, sum log(CPO) for each model in a table. Plot a histogram of the PIT values. Which model has the best predictive performance as measured by sum log(CPO)? Should we be concerned if we select a 'best predictive performance' model via CPO?

Spatial K-fold CV

Measure the predictive capacity of each model via a 4-fold CV that takes spatial dependence into account.

- Include all R and INLA code.
- Make a map that shows how the data are partitioned into the 4 folds.

- Compute $\frac{1}{S}\sum(y_i - \hat{y}_i)^2$. Here S is the sample size and $\hat{y}_i = N_i \cdot \hat{p}_i$, for $\hat{p}_i$ equal to the mean (out-of-sample) predicted prevalence.

Compare models and select the one with the best predictive capacity. Does it match your results when using CPO? Are there any major observed differences in your assessment of predictive capacity when comparing CPO and 4-fold CV across models?

Choice of Model

What model would you say explains/predicts Malaria prevalence in The Gambia the best? Are there any concerns you have with your choice of selected model?