Variable selection for Small Area Prediction

We propose a comparison of different model selection methods for small area estimation in the situation in which the number of areas with full data is a small fraction of the total areas for which estimates are required. However, there are abundant auxiliary data for the non-sampled areas. Three procedures are studied: our preferred systematic Bayesian shrinkage method; a stepwise approach that uses sampling adjusted standard model selection criteria (e.g. AIC); a machine learning prediction approach.

This work is motivated by Ghanaian data available from the sixth Living Standard Survey (GLSS) and the 2010 Population and Housing Census. We aim to estimate the areal mean household log consumption using both datasets. The dependent variable is measured only in the GLSS for 3% of all the areas (136 out of 5037). More than 170 potential covariates are available from both datasets. Modelling is carried out within a linear mixed model framework and we utilize state-of-the-art Bayesian computation techniques.