Smarter Than Your Average Model - Bayesian Model Averaging as a Spatial Analysis Tool

Chris Brunsdon
National Centre for Geocomputation, Maynooth University, Ireland

Paul Harris
Rothamsted Research, Harpenden, UK

Alexis Comber
School of Geography, University of Leeds, UK

Abstract

Bayesian modelling averaging (BMA) allows the results of analysing competing data models to be combined, and the relative plausibility of the models to be assessed. Here, the potential to apply this approach to spatial statistical models is considered, using an example of spatially varying coefficient modelling applied to data from the 2016 UK referendum on leaving the EU.

2012 ACM Subject Classification Mathematics of computing → Bayesian nonparametric models

Keywords and phrases Bayesian, Varying coefficient regression, Spatial statistics

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.17

Category Short Paper

Supplementary Material *Text*: `https://drive.google.com/file/d/1HHb5uEsGX0Qu61hIwWd-8BgjTrGzn6R-/view?usp=share_link`

1 Overview

Imagine that you are waiting for a taxi and it is already slightly late. You are concerned that you will miss a train, and want to estimate how long you will need to wait. A number of scenarios could cause the delay. For example: The taxi is stuck in traffic; There was an administrative error and the booking service gave the taxi driver the wrong time; The taxi was involved in a road accident; and so on. In each case a number of factors effect the expected delay - but the factors are not the same in each scenario. However your main concern is the delay time, regardless of the scenario. This is a similar problem to those which Bayesian Model Averaging (BMA) may be used to address.

If you had models encompassing k scenarios based on past data D - say $\{M_1 \cdots M_k\}$ intended to predict the delay time T, and posterior beliefs in each scenario being correct: $\{\Pr(M_1|D) \cdots \Pr(M_k|D)\}$ you could obtain the predictive distribution of T given D as a weighted average of the individual predictive distributions obtained from each model as

$$\Pr(T|D) = \sum_{i=1,k} \Pr(T|M_k, D)\Pr(M_k|D).$$

This in essence is Bayesian Model Averageing (BMA) – if we have a number of competing models with at least one quantity of interest that all have in common, and relative likelihoods of each of them being the correct model, we can obtain a posterior distribution of the quantity of interest by averaging them using the likelihoods as weights.

Up to this point, there is nothing exclusively spatial about this process, but it can be a powerful tool for assessing and utilising spatial models. For example, the competing models could be:

1. Spatial regression models using different spatial weight matrices.
2. Spatially Varying coefficient regression models where different parameters have fixed or spatially varying coefficents in each model.
3. Spatial trend models with differing map projections (eg. a cartogram vs. national grid coordinates)

In general, this approach can be used for any parameter that is common to all models, or a predicted dependent variable – so if one were interested a particular regression coefficient, its posterior distributioncould be considered in terms of various models containing this coefficient. A key advantage of this approach is that while many other approaches (eg stepwise regression, best AIC, best cross validation score) have a workflow to select a single "best" model, this averages over all possibilities on the basis of relative evidence. In particular when several models all perform similarly well, this approach makes use of information from all of them, rather than discarding all but one.

2 A Brief Description of Computational Methodology

The approach to computing $\Pr(M_i|D)$ – a crucial stage in BMA - is to firstly compute $\Pr(D|M_i)$ – then, via Bayes' Theorem, we have

$$\Pr(M_i|D) = \frac{\Pr(D|M_i)\Pr(M_i)}{\sum_j \Pr(D|M_j)\Pr(M_j)}.$$

Each model M_i will have its own parameter vector Θ_i - although the respective Θ_i may differ in length and form between models. Standard statistical models typically specify $\Pr(D|\Theta_i, M_i)$ - but here we are interested in the *marginal* probability of the observed data D across all possible Θ_i values for each M_i, weighted by their prior probabilities. That is

$$\Pr(D|M_i) = \int_{\Theta_i} \Pr(D|\Theta_i, M_i)\Pr(\Theta_i|M_i)\ d\Theta_i.$$

This is sometimes referred to as the marginal posterior probability of D given M_i. Although the right hand side expression cannot usually be derived analytically, two broad approaches may be taken:

1. Approximation.
2. Monte Carlo Simulation.

Approximation is generally quicker and less "resource hungry" to evaluate, but less accurate. A usual strategy for approximation is based on the Bayesian Information Criterion (BIC) [4] for model M_i. If $\widehat{\Theta}_i$ is the maximum likelihood estimate for Θ_i for M_i, and $\widehat{L}$ is the value of the likelihood corresponding to $\widehat{\Theta}_i$, n is the sample size, and k is the dimension of Θ_i then

$$\text{BIC} = k\log(n) - 2\log(\widehat{L})$$

and for larger n it can be shown that

$$\Pr(D|M_i) \approx \exp\left(-\frac{\text{BIC}}{2}\right).$$

Finally, for the parameter(s) of interest, say $\theta_i \subset \Theta_i$ for a given model M_i the posterior distribution can be approximated via Laplace's approximation [1]. The posterior distribution for θ_i may be approximated as having a multivariate normal distribution with a variance-covariance matrix equal to the Hessian of the posterior likelihood function, with the maximum

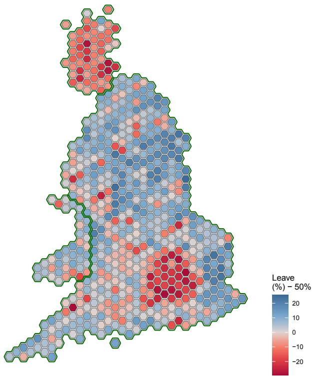

Figure 1 Leave Vote (%) by Parliamentary Constituency.

Table 1 Variables Used in Referendum Outcome Modelling.

Variable	Description
`Leave`	Percentage of "Leave" votes for each constituency (Dependent Variable).
`Born_uk`	Percentage of electorate born in the UK.
`Age_65_plus`	Percentage of electorate aged 65 or older.
`Turnout`	Percentage of electorate who voted in the 2015 general election.
`Christian`	Percentage of electorate stating their religion as "Christian".

likelihood estimators of θ_i as mean values. For a scalar θ_i this suggests that the marginal posterior distribution may be estimated as a normal distribution with the maximum likelihood estimate $\hat{\theta}_i$ as its posterior mean, and $\mathrm{SE}(\hat{\theta}_i)$ as its posterior standard deviation. The BMA may then be approximated as a mixture of the k Normal distributions with $\Pr(M_i|D)$ as the weight for M_i.

In this study, the example will use the BIC-based approach, and so attention will be focused on this method.

3 Example: The UK's 2016 Referendum on Leaving the EU

On June 23rd 2016, the United Kingdom held a referendum regarding its then membership of the European Union. Voters were offered two choices: "Leave the European Union" (Leave) or " Remain a member of the European Union" (Remain). The outcome was a 51.9% majority in favour of "Leave", although a hexagonal cartogram map of voting by Parliamentary Constituencies in England, Scotland and Wales (Figure 1) suggests this overall figure conceals notable regional patterns. This leads to a further question: if the voting patterns themselves show strong regional patterns, do the *drivers* of these outcomes also vary geographically?

To investigate this, a number of variables were obtained (from the `parlitools` R package) [3], recorded at at the Parliamentary Constituency geographical unit – listed in Table 1. The UK census-based variables (`born_uk`, `age 65+`, and `Christian`) were recorded in the 2011 UK Census – this being the latest Census held in the UK prior to the referendum.

The key questions for each variable are whether they influence the leave vote; and if so then does the direction and magnitude of this influence vary geographically? To investigate this, for each variable it is possible to include it in a model with a fixed linear coefficient $\beta \times$ Variable or a geographically varying coefficient $\beta(u, v) \times$ Variable where (u, v) is the centroid of each parliamentary constituency, or to omit it from the model.

To investigate this, the R package `mgcv` was used to fit every permutation of these kinds of model. For each of the four predictor variables there were three possibilities - omit the variable from a regression model, include with a fixed linear coefficient, or include with a spatially varying coefficient. In the latter case a thin-plate spline approach was used (although other options could be chosen). In the `R` formula notation, an example of a model might be

```
Leave ~ s(u,v,by=Born_uk,bs='tp') + Turnout
```

suggesting a model where the coefficient for `Turnout` was fixed, that for `Born_uk` varied, and the other variables were omitted. This yields $3^4 = 81$ models. In addition to this, each model was fitted with both fixed and varying intercept terms, and with the coordinates (u, v) based on location on the cartogram and physical (UK National Grid) location. Thus there are 4 variants on each model, resulting in $81 \times 4 = 324$ possible models altogether. In the marginal likelihood approach, there is no requirement that models be nested, so all 324 models can be considered. Here the `mgcv` package offers Bayesian Information Criterion methods (BIC) for `gam` model objects, and so the `BIC` based approximation will be used here. Using this approach, all models with a posterior probability ≥ 0.01 are listed in Table 2.

The most likely model includes all variables with the intercept and the `Born_uk` coefficient being modelled as thin plate splines, and the remaining variables having fixed linear coefficients. The geographical coordinates for the splines are based on the cartogram, rather than physical space. However, reading the $\Pr(M|D)$ column in the table suggests that this model is the correct one is a little under two thirds. The possibility of one of the "runners up" being correct is non-trivial. In the next model in the table (probability around one in five) `Born_uk` has a fixed coefficient - but also although the intercept term is varying, the coordinates are now based in *physical* space.

The `Intercept` term has a spatially varying coefficient in all of the three most probable models. These three models dominate the posterior marginal probabilities, totalling around 0.95 of all possibilities. These surfaces are shown in Figure 2. The Bayesian model average surface (over all possible models) for intercept is given by

$$\beta_{0*}(u, v) = \sum_{i=1\cdots 324} \Pr(M_i|D)\beta_{0i}(u, v)$$

where $\beta_{0i}(u, v)$ is the intercept coeffient for model i. For models where the intercept is constant, $\beta_{0i}(u, v)$ is a constant w.r.t. (u, v). This is shown as the fourth map in Figure 2 on the LHS map quartet.

In these models all variables - as listed in Table 1 - are standardised to have mean zero and standard deviation 1 prior to analysis. For the intercept term, this gives the standardised value for the `Leave` variable assuming all other predictors are at their mean value. It is not a direct measure of overall tendency to vote "Leave" or otherwise - more of a measure of geographical effects not accounted for by current variables in the model. On this basis, there seems to be among other things a "Scotland effect" and a "West London effect" (although this is not apparent in the second most probable model, which uses physical coordinates rather than cartogram). Once the models are averaged the West London effect remains, although muted.

Table 2 Models with Highest Posterior Probabilities.

Intercept	Born_uk	Age 65+	Turnout	Christian	Coords	Pr(M|D)
Spline	Spline	Fixed	Fixed	Fixed	Cartogram	0.637
Spline	Fixed	Fixed	Fixed	–	Physical	0.205
Spline	Fixed	Fixed	Fixed	Fixed	Cartogram	0.109
Spline	Fixed	Fixed	Fixed	Fixed	Physical	0.045

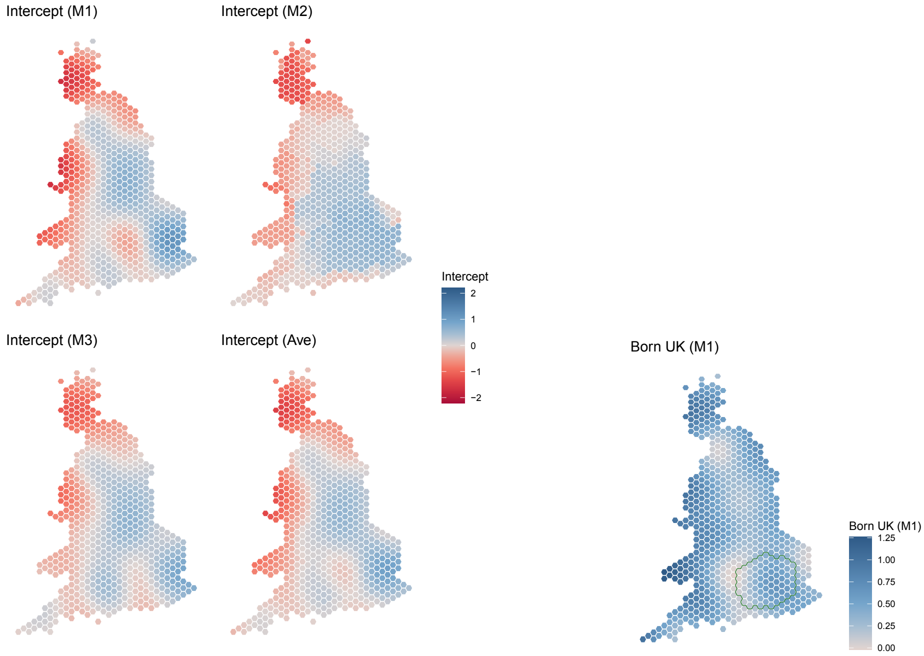

Figure 2 The `intercept` and `born_uk` terms by parliamentary constituency (Great Britian).

The coefficient for `born_uk` can also be mapped. This is shown in Figure 2 (RHS). The values are calculated using the formulae above. Of note here is perhaps that in a region to the west of london, `born_uk` seems to have little infuence on the outcome than in much of the country where higher values suggest a `Leave` majority is more likely.

It is also possible to map the marginal posterior standard deviation for these parameters, after model averaging. These are computed using the formula

$$[\mathrm{PSD}\,(\beta_*(u,v))]^2 = \sum_{i=1\cdots 324} \Pr(M_i|D)\,[\mathrm{PSD}\,(\beta_i(u,v))]^2$$

and are shown in Figure 3. Notable in both cases is the "edge effect" where the PSD is high near to the coastal areas. Also of note is the raised PSD in the London area.

4 Discussion

The BMA approach provides a number of useful tools. It provides a means of assessing the viability of competing models, by providing posterior probabilities of each being the correct model. This can be thought of as similar to hypothesis testing, but it treats hypotheses symmetrically, and can handle more than two competing hypothesis. It also a provides means

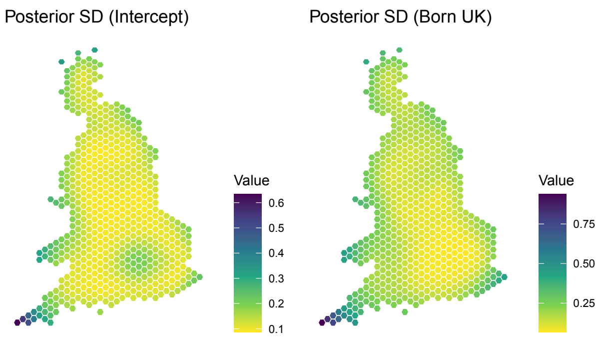

Figure 3 Posterior Standard Errors for `Intercept` and `born_uk`.

of combining competing models to investigate parameters common to all models, in the presence of uncertainty as to which model is correct. The example here used an approximate approach that is convenient, as it can be achieved using standard R tools. More accurate approaches are also possible via techniques such as *Bridge Sampling* – see [2] for example.

References

1 Robert E. Kass, Luke Tierney, and Joseph B. Kadane. Laplace's method in Bayesian analysis. *Contemporary Mathematics*, 115:89–100, 1991. `doi:10.1090/conm/115/07`.

2 X. Meng and S. Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586, 2002. `doi:10.1198/106186002457`.

3 Evan Odell. *parlitools: Tools for analysing UK politics in R*, 2017. R package version 0.4.1. `doi:10.5281/zenodo.591586`.

4 Gideon Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6(2):461–464, 1978. `doi:10.1214/aos/1176344136`.