A Bayesian Hierarchical Model for Educational Infrastructure Quality using School Effectiveness, and other Spatial/Regional Information

Zheng Li, Chamari Idamawatta, Sulalitha Bowala, Micah Thornton [DropBox]

December 28, 2018

1 Introduction

The school effectiveness index (SEI) is a measure of a schools ability to impart relevant and useful information to students. It is a score that is computed by the Dallas Independent School District [1] such that they may assess individual performance of schools. The SEI is computed for each student and at the institutional level for each school. To compute the SEI, analysts compile data regarding standardized testing, and other relevant information to arrive at the score. The full details of computation of the SEI including a listing of the utilized standardized tests such as the STAAR and the TerraNova reading assessments. The score is not calculated uniformly, for instance, in the case of Grade 1 - Grade 2 students, depending on the dominant language of the student, different tests are considered. For Spanish speaking students, instead of the TerraNova test, scores from the SUPERA test are considered. This is in an attempt to control for the possible confounding associated with language-dominance, but could potentially introduce a Bimodal distribution if the tests were not norm-referenced which according to the manual they are. [2] Schools in the Dallas Metroplex are arranged into districts. In the supplied SEI data set, 9 trustee districts are named, and is reported for each of the 149 institutions. The modeling procedures that are discussed and applied in this paper are run within districts to produce specific estimates for parameter values within each district, which then may be compared across districts. A Similar study was done in California in 2004. [7].

Also reported in the dataset are percentages indicating the ratio of students enrolled at the institution which fit into a specific category of ethnicity (the file also considers students who personally identify with two or more ethnic origins). In this study one item of interest is the manner in which the diversity of the school affects the probability of an SEI score higher than 50 (a nominal cutoff decided on as it differentiates schools well). To this ends, the demographic information for each school was used to generate a diversity score which is computed as the residual distance from the respective ethnic percentages to those which would be expected in the Dallas population. In other words, we imagine that the ideal diversity to be exhibited within a Dallas school would perfectly reflect the diversity of the population of Dallas. For the n ethnic frequencies reported for each of the schools$(f_1, \ldots, f_n)$ and the corresponding population frequencies of Dallas $(f_1^*, \ldots, f_n^*)$, the diversity score D may be computed as follows:

$$\text{let } D^* = \sum_{i=1}^{n} (f_i - f_i^*)^2$$

D^* simply represents the sum of the squared errors from the ideal population values, with higher scores indicating a lower diversity. The score is scaled by the maximal observed distance (as to scale it between 0 and 1) and the resulting scaled D^* is subtracted from one, in order to create a diversity score D which is a value between zero and one indicating a higher diversity when the score is closer to one.

Table 1: Model Covariates

Variable of Interest	**Description**
School Type	Elementary School, Middle School,High School
Normalized Diversity Score	Score based on the ethnicity of children
Choice Magnet School	Schools for which enrollment is based on student choice.
Post Secondary Distinctions	Number of awards for college readiness.
Average Distance	Average distances between school i to others within same district
Number of Rail Stations	Number of rail stations within 1.5 km from school i

Along with demographic information, the total number of students attending each institution is listed, and any distinctions the institution won during the academic year 2016/17 (for which the SEI is reported). The schools included offer a variety of different grade levels of education, and are divided by a school type which refers to the primary set of grades offered by the institution. The designations for school type are 'ES' for elementary schools (typically grades 1-5) 'MS' for middle schools (typically grades 6-8), and 'HS' for high school (grades 9-12). These designations may be represented as an ordinal categorical variable, and the reassignments of 'ES' to 1, 'MS' to 2, and 'HS' to 3 are made prior to the analysis. Choice/Magnet schools are schools that offer special accelerated learning courses in a variety of subjects (an example is the Booker T. Washington school, which is an arts magnet that focuses on education in music, acting, and the visual arts). A binary variable concerning the schools status as a magnet or choice school is also considered as a predictor of the schools SEI, as it is postulated that schools which are attractive to high-performing students by offering a variety of courses aligned with their interests.

Spatial statistics of interest are also considered in this Bayesian model, as previous studies have suggested [9]. In this study, two spatial variables, average distances from school i to others and number of rail stations within 1.5km from school i were considered in order to identify their effects on SEI score. These variable are computed using supplied ArcGIS data on locations of schools that were processed using Python. [8]

The full listing of variables that are included as predictors in the logistic regression model for the probability of having a high SEI is available in Table 1.

2 Modeling and Justification

The work outlined in this paper documents the application of a hierarchical Bayesian model for the Bernoulli outcome of whether a school, given the relevant covariates, has an SEI score higher than 50. The model links the probability of the SEI being greater than a specified nominal level (50 in the case of this study), using a logit link.

$$SEI_{ij} = \text{SEI score of } i^{\text{th}} \text{ school in } j^{\text{th}} \text{ school district;}$$

$$\text{where } i = 1, 2, ..., n_j,\ j = 1, 2, ..., J$$

Let us assume j^{th} school district has n_j number of schools in total:

$$y_{ij} = \mathbb{I}(\text{School } i \text{ has SEI score above given level } \alpha)$$

$$y_{ij} \sim Bernoulli(\theta_{ij})$$

The Logistic regression is defined as:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \boldsymbol{\beta_j^T X_{ij}} = \beta_j^0 + \beta_j^1 x_{ij}^1 + \beta_j^2 x_{ij}^2 + ... + \beta_j^p x_{ij}^p \ ;$$

where;

$$i = \{1, 2, ..., n_j\},\ j = \{1, 2, ..., J\},$$
$$p = \text{Number of included covariates},$$
$$\pi_{ij} = \frac{e^{\boldsymbol{\beta}_j^T \boldsymbol{X}_{ij}}}{1 + e^{\boldsymbol{\beta}_j^T \boldsymbol{X}_{ij}}}$$

The likelihood function can be written as:

$$p(y \mid \boldsymbol{\beta}, \boldsymbol{X}) = \prod_{j=1}^{J}\prod_{i=1}^{n_j} Bern(\pi_{ij}) = \prod_{j=1}^{J}\prod_{i=1}^{n_j}\left(\frac{e^{\boldsymbol{\beta}_j^T \boldsymbol{X}_{ij}}}{1 + e^{\boldsymbol{\beta}_j^T \boldsymbol{X}_{ij}}}\right)^{y_{ij}}\left(\frac{1}{1 + e^{\boldsymbol{\beta}_j^T \boldsymbol{X}_{ij}}}\right)^{1-y_{ij}}$$
$$= exp\left(\sum_{j=1}^{J}\sum_{i=1}^{n_j} y_{ij}\boldsymbol{\beta}_j^T \boldsymbol{X}_{ij} - log(1 + e^{\boldsymbol{\beta}_j^T \boldsymbol{X}_{ij}})\right)$$

3 Probability Model

The above model links the probability of interest with covariates that are contained within the design matrix $\boldsymbol{X}_{n*(p+1)}$, via the parameters $\boldsymbol{\beta}_{(p+1)*1}$.

The Zellner g-prior [10] distribution is placed on logistic regression parameters. σ^2 has a gamma hyper-prior distribution with the hyper-parameters α and λ. Mathematically, we may express the model as follows:

$$P(y_{ij}=1)=\frac{1}{1+e^{-\boldsymbol{\beta}_j^T x_i}};\qquad P(\boldsymbol{\beta}_j=\boldsymbol{\beta}_j^*)=\frac{\exp\left(-\frac{1}{2}(\boldsymbol{\beta}_j^{*T})(g\cdot\sigma^2\cdot\boldsymbol{\Sigma})^{-1}(\boldsymbol{\beta}_j^*)\right)}{\sqrt{(2\pi)^p\cdot|g\cdot\sigma^2\cdot\boldsymbol{\Sigma}|^{-1}}};\qquad \boldsymbol{\Sigma}\propto I^{-1}(\boldsymbol{\beta}_j)=(\boldsymbol{X}^T\boldsymbol{X})^{-1};$$

Where y_{ij} represents a binary outcome indicating whether the i^{th} school in the j^{th} district had a SEI score higher than 50. $\boldsymbol{\beta}_j$ represents a **column** vector of the coefficients for the logistic regression model within a given district, and x_i is a **column** vector containing the explanatory variables for the i^{th} school. The g-prior distribution placed on the coefficients is a multivariate normal distribution with a mean vector of zeros, and covariance structure proportional to the inverse information matrix of the parameters of the logistic regression model. [3] The parameter σ further has a gamma hyper-prior distribution placed on it, with hyper parameters α and λ. The variable g of the prior-g distribution is taken as the number of observed data records. In this case, the SEI and corresponding covariate values are reported for a total of 149 schools, meaning that the value of g taken in this case is 149. The R-package used in this analysis was LaplacesDemon [5, 6]

4 Posterior and conditional distributions

Prior distributions:
Set up a g-prior on parameter coefficient of the logistic model, $\boldsymbol{\beta}$, and a noninformative and appropriate prior on σ.

$$\begin{aligned}
P(\boldsymbol{\beta}\mid g,\sigma,\boldsymbol{X}) &= MVN(\boldsymbol{\mu}, g\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})\\
P(\boldsymbol{\beta}\mid g,\sigma,\boldsymbol{X}) &= \frac{1}{(2\pi)^{P/2}|(g\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}|^{1/2}} exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})^T(g\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})\right)\\
P(\sigma) &\sim Gamma(\alpha,\lambda)
\end{aligned}$$

Joint posterior distribution:

$$\begin{aligned}
&P(\boldsymbol{\beta},g,\sigma\mid y,\boldsymbol{X})\propto P(y|\boldsymbol{\beta},\boldsymbol{X})P(\boldsymbol{\beta}|g,\sigma,\boldsymbol{X})P(\sigma)\\
&P(\boldsymbol{\beta},g,\sigma\mid y,\boldsymbol{X})\propto exp\left(\sum_{j=1}^{J}\sum_{i=1}^{n_j} y_{ij}\boldsymbol{\beta}_j^T\boldsymbol{X}_{ij}-log(1+e^{\boldsymbol{\beta}_j^T\boldsymbol{X}_{ij}})\right) MVN\left(\boldsymbol{\mu}, g\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)Gamma(\alpha,\lambda)\\
&P(\boldsymbol{\beta},g,\sigma\mid y,\boldsymbol{X})\propto exp\left(\sum_{j=1}^{J}\sum_{i=1}^{n_j} y_{ij}\boldsymbol{\beta}_j^T\boldsymbol{X}_{ij}-log(1+e^{\boldsymbol{\beta}_j^T\boldsymbol{X}_{ij}})\right)\\
&\qquad\qquad \frac{1}{(2\pi)^{P/2}|(g\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}|^{1/2}} exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})^T(g\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})\right)\sigma^{\alpha-1}e^{-\lambda\sigma}
\end{aligned}$$

Conditionals derive from the joint density:

$$\begin{aligned}
&P(\sigma\mid .)\propto\frac{1}{|(g\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}|^{1/2}} exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})^T(g\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})\right)\sigma^{\alpha-1}e^{-\lambda\sigma}\\
&P(\boldsymbol{\beta}\mid .)\propto\prod_{j=1}^{J}\prod_{i=1}^{n_j}\left(\frac{e^{\boldsymbol{\beta}_j^T\boldsymbol{X}_{ij}}}{1+e^{\boldsymbol{\beta}_j^T\boldsymbol{X}_{ij}}}\right)^{y_{ij}}\left(\frac{1}{1+e^{\boldsymbol{\beta}_j^T\boldsymbol{X}_{ij}}}\right)^{1-y_{ij}} exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})^T(g\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})\right)
\end{aligned}$$

5 Simulation Results and Sensitivity Analysis

The simulation we are conducting in this study is to evaluate the conditional posterior distributions of the parameters of the logistic regression model as well as conditional posterior distribution of σ. First, we will set a basic model considering all the data while not taking the district level into account. Later we will introduce the district level to our Bayesian hierarchical structure and deduce more realistic results. The simulations performed in this study were done using the R-language [4].

The full model was simulated using the Markov-Chain Monte-Carlo techniques of Gibb's sampling, and the Metropolis Hastings (MH) algorithm. As suggested in our class notes and many studies, g is equal to the sample size. In the Hyperprior-g Prior setup σ is the residual standard deviation and it can vary from 0 to ∞. Thus, matching the scale of σ we have introduced a prior on σ which follows a *Gamma* distribution with parameters α and λ. Since the exact information is not available us, we have set up several priors for σ considering several combinations of αs' and λs'. We have evaluated these hyper-priors and how it effects the parameter estimates under a sensitivity analysis and hyper-parameters that are examined in the sensitivity analysis are: ($\alpha = 8$, $\lambda = 3$), ($\alpha = 9$, $\lambda = 3$), and ($\alpha = 9$, $\lambda = 4$).

Simulation will evaluate conditionals starting from $P(\sigma \mid .)$. We have simulated 2000 iterations followed by 500 burn-in iterations. Once conditional posterior is estimated for σ simulation evaluate the conditional posterior density of the parameter vector $\boldsymbol{\beta}$.

5.1 Full Model

First we set up the full model. Here the district level is omitted and Figure 1 summarize the simulation results for σ and 8 parameters we have included in the logistic regression model including intercept. For the three hyper-priors we have considered, estimated σ values are close. However, when $\alpha = 9$ and $\lambda = 4$, estimated σ is low with smaller variance compared to the other two cases. With 2000 iterations, except β_2, β_3, and β_6, parameters of the logistic regression model others have converged to closer values. Convergence of the algorithm can be confirmed with the ACF plots provided in the appendix.

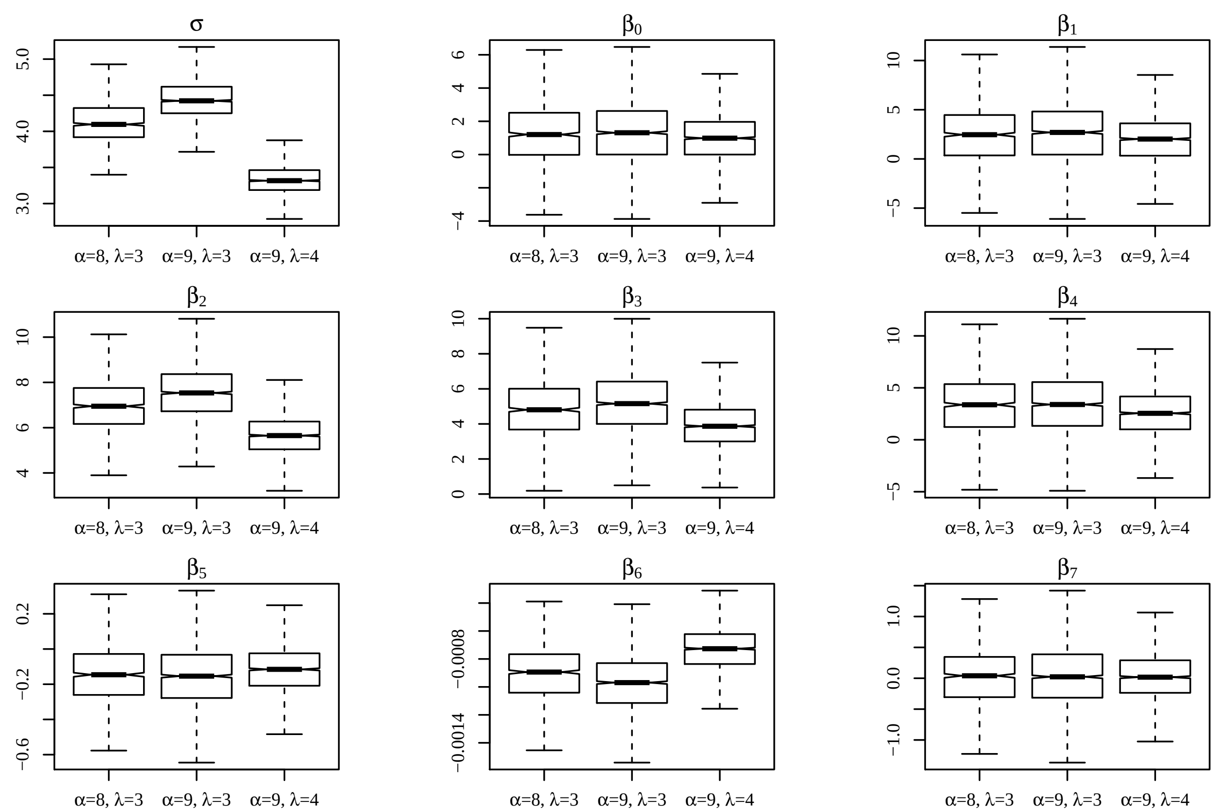

Figure 1. Sensitivity Analysis with $(\alpha, \lambda) = (8, 3), (9, 3)$ and $(9, 4)$.

5.2 District Level Model

In order to identify district effects on SEI, we have introduced district level into the model and the simulation separately done for each of 9 districts. Because of the matrix singularity issues encountered during the simulation (discussed in section 7), we have to regroup the districts into three District Groups(DG) in order to allow MCMC to converge. The $\boldsymbol{\beta}$ posterior estimates among three district groups is presented in the Figure 2. Detailed DG setting and the posterior mode and central posterior interval(CPI) for each $\boldsymbol{\beta}$ along with districts are given in the Table 2, 3 in Appendix.

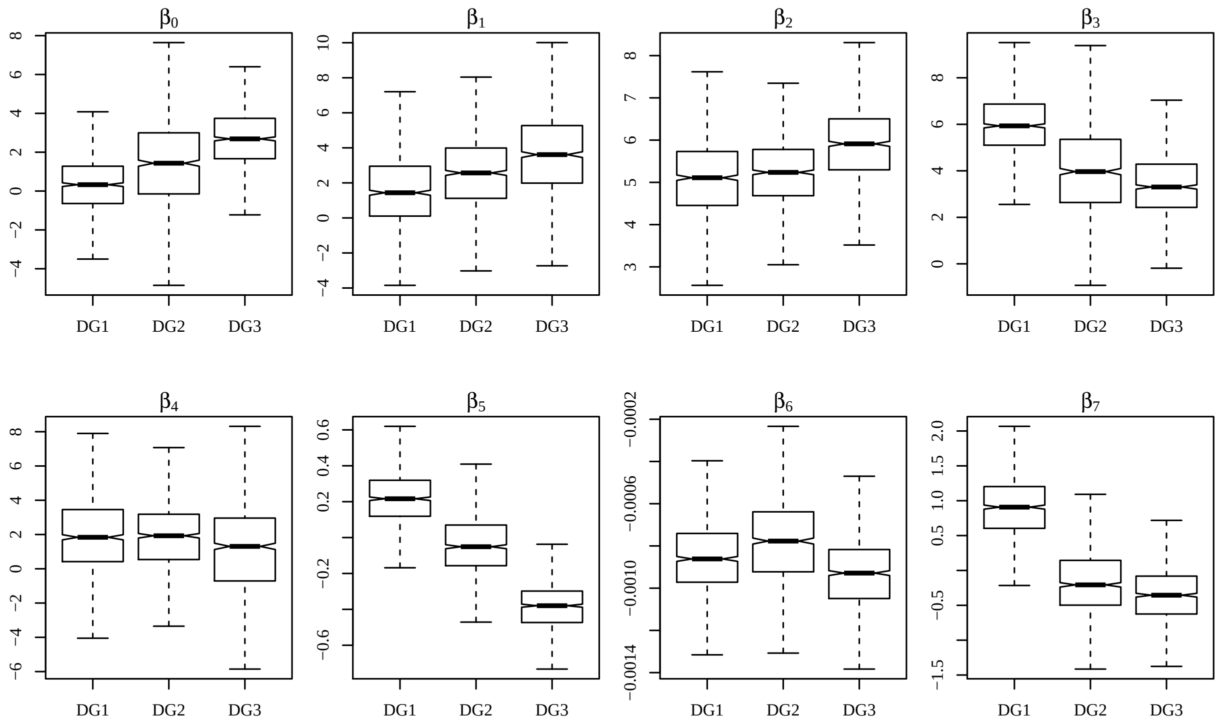

Figure 2. $\boldsymbol{\beta}$ posterior distribution estimates comparison among district groups (DG1, DG2, DG3) with fixed $\sigma = 3.317$.

6 Conclusion

The suggested logistic model was implemented to predict SEI of Dallas Independent School Districts, and determine the relative effects of model parameters on the probability of having a high SEI, for a given set of covariates. In this model we include variables related to the diversity of school, as well as spatial variables and other information. The convergence of algorithm and sensitivity analysis were performed for the logistic model. Since we didn't have SEI for all schools one can use this model to predict SEI of other schools in Dallas ISD and model checking can be done with later released SEI. Moreover, we can improve this by considering more variables and can identify latent variables which affects a schools SEI. Further model comparison can be done in order to identify best school districts by considering district level models in future works. The work in this paper suggests that there are ways to determine the effect of certain attributes of a school on the overall probability of the school being a high-achiever (in terms of SEI). A more extensive sensitivity analysis in which many different values for hyper parameters are chosen, and the sensitivity is numerically quantified may be relevant. It was noted in this work that σ^2 was quite sensitive to choice of different hyper-parameters, when using a gamma hyper-prior, this suggests a potentially different choice of hyper-prior may be necessary. Using a standard MVN distribution with fixed covariance structure, and hyper prior distribution for the mean vector may also be a relevant pursuit, however the use of MVN characterized as a g-prior in this work may outperform a standard MVN with fixed covariance structure. The covariance structure could be estimated by building 149 leave-one-out logistic models then computing empirical correlation.

7 Discussion and Future Direction

1. Grouping of Trustee Districts

   The information matricies for the logistic regression parameters within districts were too ill-conditioned to effectively compute the inverse required for computation in the case of several trustee districts. This problem is solved by grouping spatially adjacent trustee districts into 3 groups, and using the data for all schools within the grouped districts. Other approaches were also discussed and implemented, but the final solution was grouping districts. Alternatives tried included the Moore-Penrose Generalized pseudo-inverse, and the dampening of the matricies by adding small values to diagonal elements.

2. Prior selection issues

   We also observed a wide variance observed for estimates of sigma across the sensitivity analysis for the hyper-prior gamma distribution on the g-priors σ^2 parameter. This shows that the σ^2 parameter of the g-prior MVN distribution is quite sensitive to the selection of parameters for the hyper-prior gamma (α and λ). Given the time limitation, we only experimented three possible prior combinations and as suggested in the results, the prior used for the final model may not be the optimal prior for the model. Although we picked the prior with the least variance of its posterior distribution, it is still difficult to clarify what is the optimal prior given the limited experiments we conducted.

3. Computational issues

   The model contains one hyper-parameter and eight $\boldsymbol{\beta}^i_j, i = 1, ...8$ to be estimated for each district. As each of the parameter is estimated using Gibbs sampler with MH algorithm, it becomes inevitably computational intensive. During the sensitivity analysis, It took $\sim$ 2 hours for one full MCMC chain with pre-defined σ (defined by $\alpha,\ \lambda$) to complete (2000 Iterations). After introducing the district layer, despite the number of data for each district is reduced compared to full data, it is still a massive undertaking to compute all the posterior distributions within R. Combined with the singularity and prior issues, we advocate that extra effort need to be put in the aspect of prior selection and code parallelization in Bayesian computing.

References

[1] Dallas independent school district (isd). *https://www.dallasisd.org/Page/29132*.

[2] Dallas Independent School District. *School Effectiveness Index Manual*, 2016/2017 edition.

[3] Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.

[4] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[5] Statisticat and LLC. *LaplacesDemon Examples*, 2018. R package version 16.1.1.

[6] Statisticat and LLC. *LaplacesDemon Tutorial*, 2018. R package version 16.1.1.

[7] Justin L Tobias. Assessing assessments of school performance: The case of california. *The American Statistician*, 58(1):55–63, 2004.

[8] Guido van Rossum and Fred L. Drake. *The Python Language Reference Manual*. Network Theory Ltd., 2011.

[9] Lance A Waller. Bayesian thinking in spatial statistics. *Handbook of Statistics*, 25:589–622, 2005.

[10] Arnold Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*, 1986.

8 Appendix

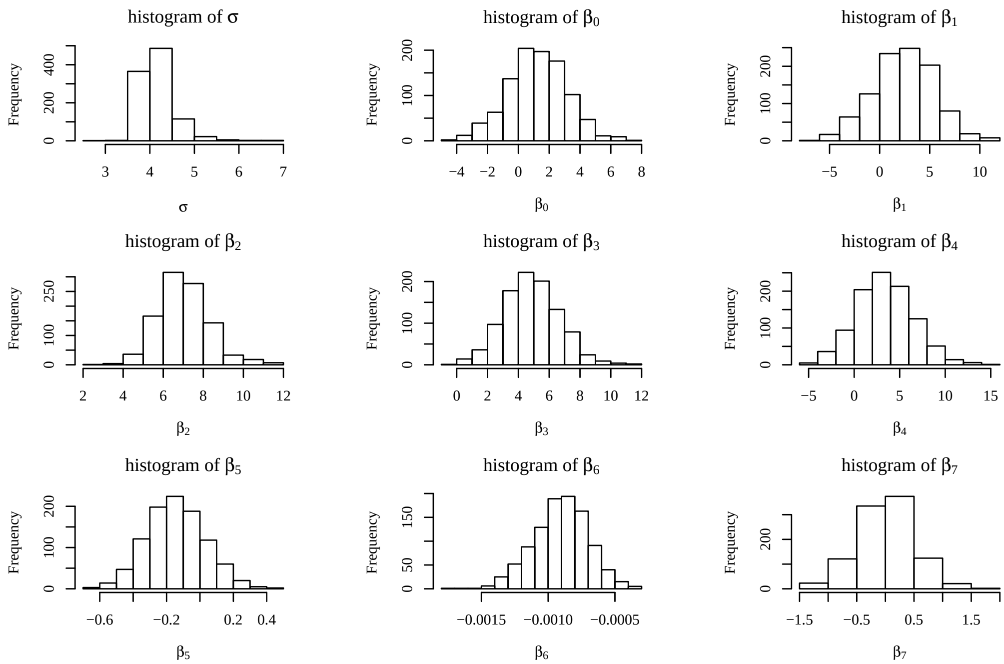

Figure 3. Estimated posterior distributions with $\alpha = 8, \lambda = 3$

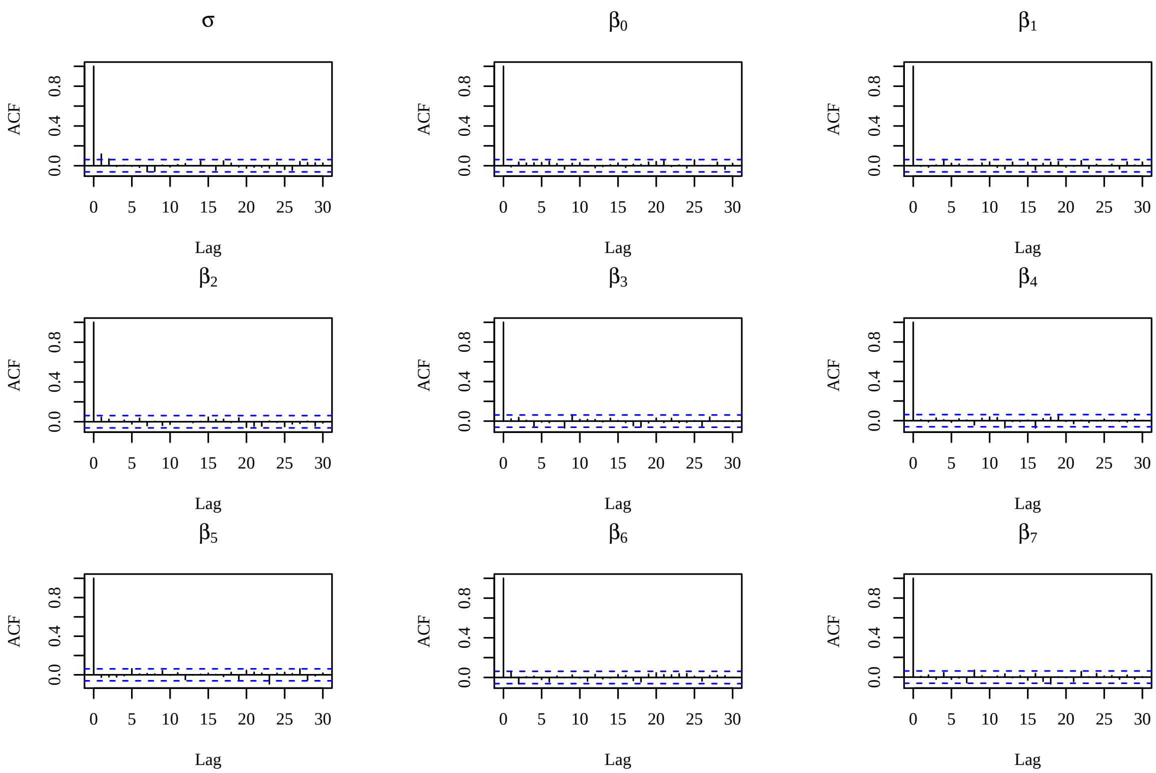

Figure 4. ACF plot with $\alpha = 8, \lambda = 3$

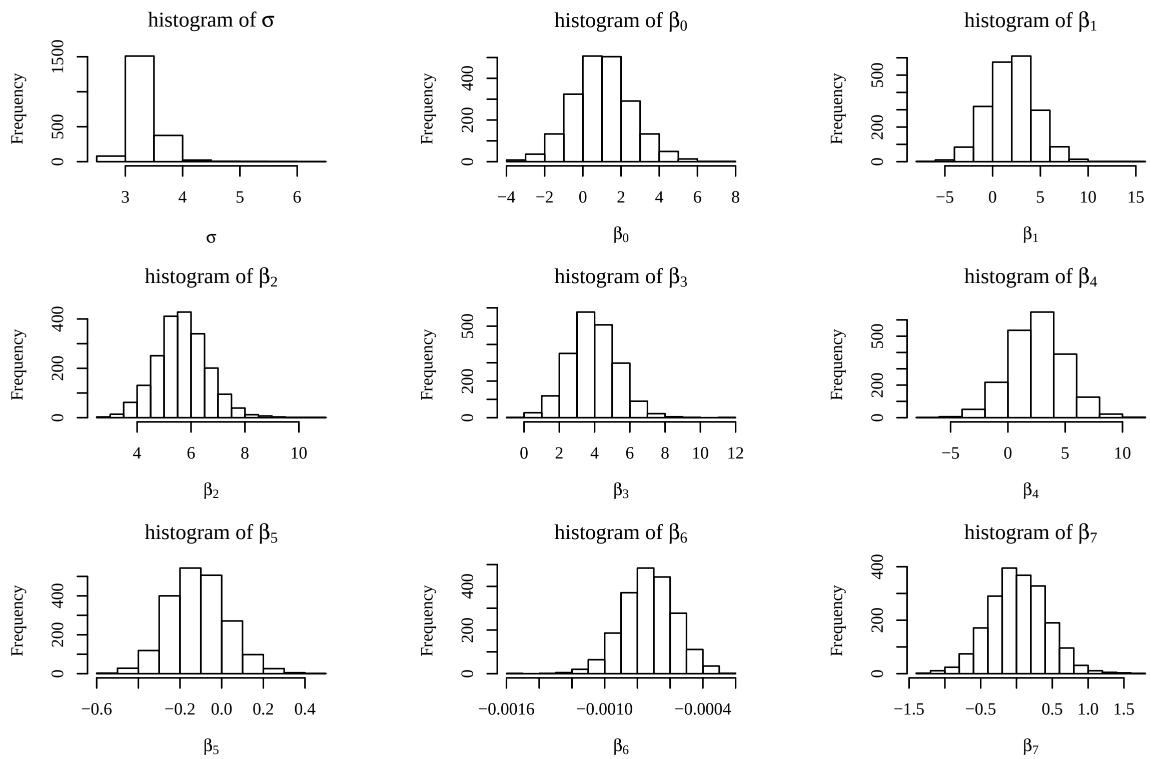

Figure 5. Estimated posterior distributions with $\alpha = 9, \lambda = 4$

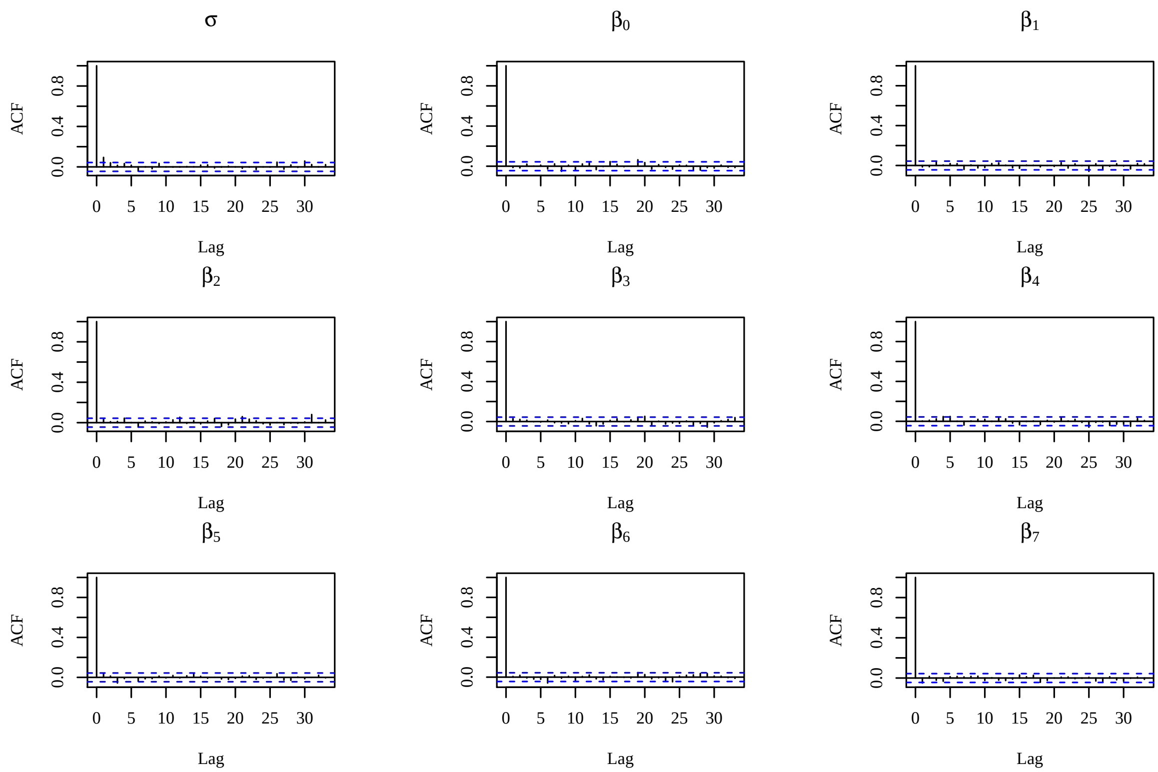

Figure 6. ACF plot with $\alpha = 9, \lambda = 4$

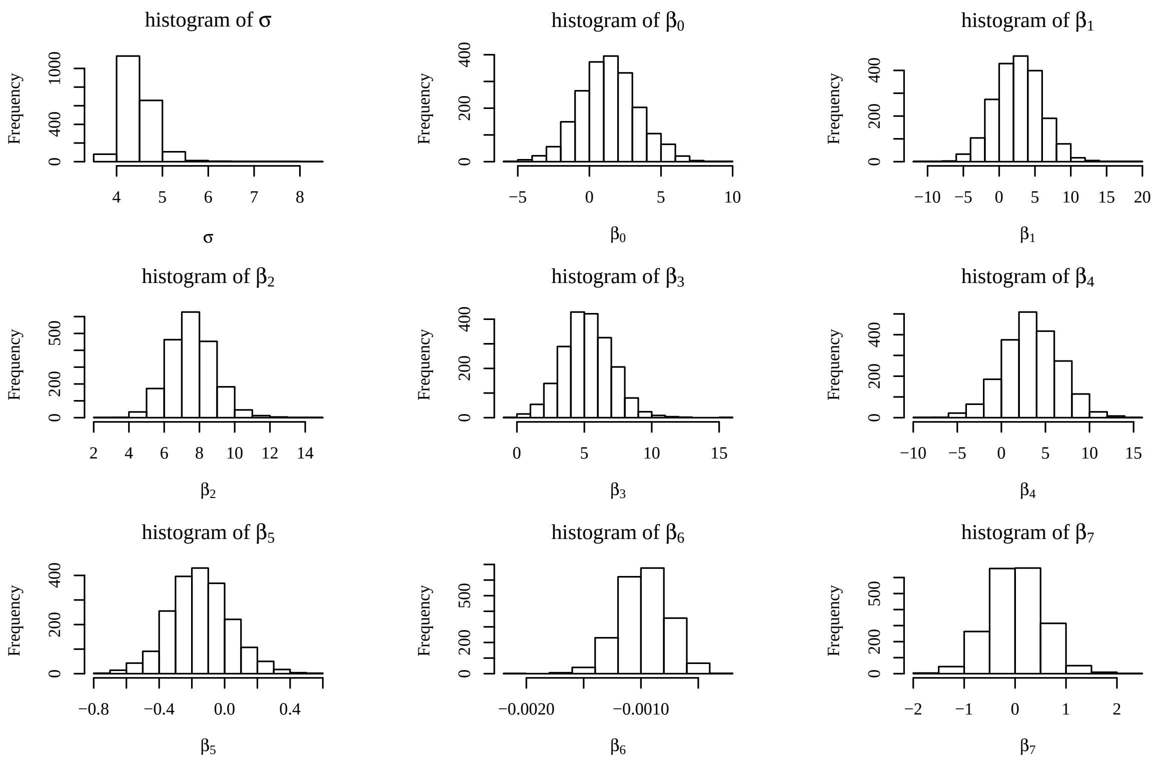

Figure 7. Estimated posterior distributions with $\alpha = 9, \lambda = 3$

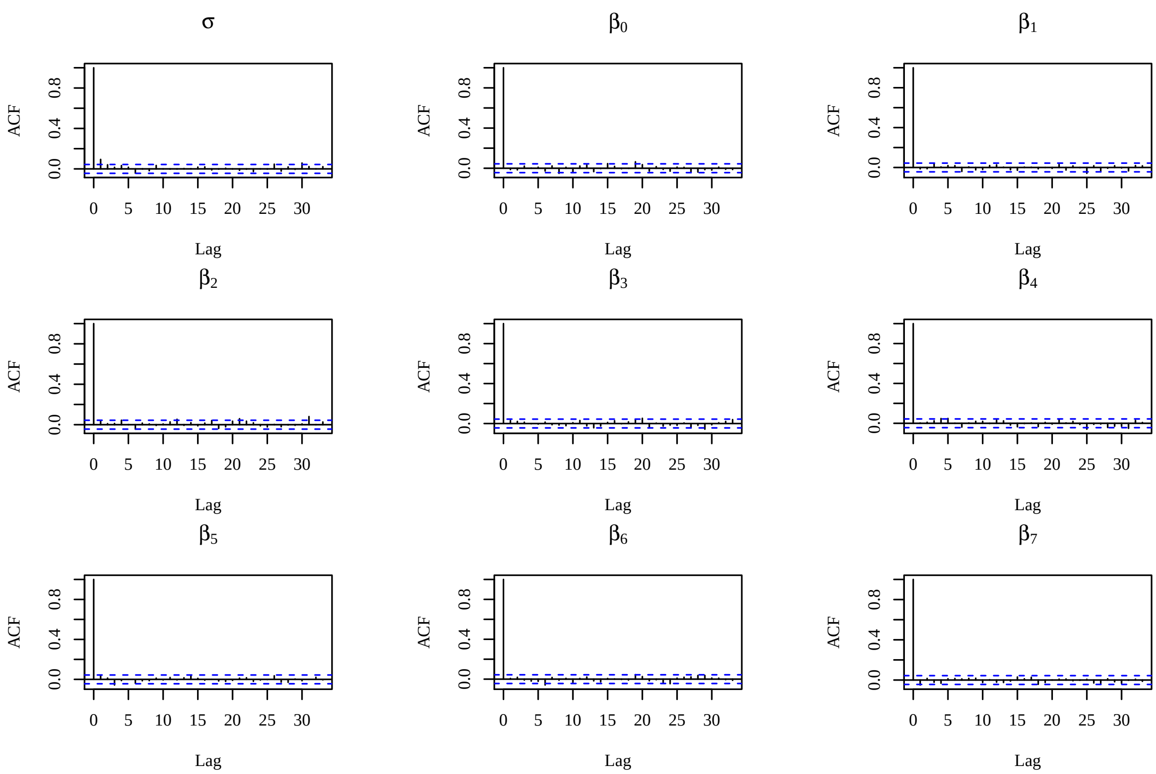

Figure 8. ACF plot with $\alpha = 9, \lambda = 3$

Table 2: District Group setting (* d: original district ID)

DG1	DG2	DG3
d1, d2, d3, d8	d4, d9	d2, d5, d6, d7

Table 3: Central Posterior Intervals for each Parameters along with Districts Level

Parameter	**DG1**			**DG2**			**DG3**		
	2.5%	**50%**	**97.5%**	**2.5%**	**50%**	**97.5%**	**2.5%**	**50%**	**97.5%**
β_0	-2.60604	0.327643	3.260982	-3.16431	1.433597	5.950008	-0.37607	2.685686	5.72512
β_1	-2.58809	1.43674	5.683269	-1.65515	2.568087	6.897953	-1.28385	3.612476	8.508606
β_2	3.305736	5.109657	6.885841	3.644324	5.236603	6.831538	4.287315	5.911711	7.624567
β_3	3.484183	5.932772	8.601214	0.359999	3.965393	8.271048	0.850726	3.303734	6.05528
β_4	-2.60509	1.838651	6.362735	-2.04357	1.92657	5.605023	-3.98468	1.307704	6.184294
β_5	-0.0878	0.216193	0.510327	-0.3516	-0.051	0.338561	-0.67297	-0.37971	-0.12936
β_6	-0.00118	-0.00086	-0.00051	-0.00117	-0.00078	-0.00036	-0.00127	-0.00093	-0.0006
β_7	0.034281	0.90818	1.692285	-1.15437	-0.20675	0.704066	-1.09523	-0.35513	0.412821