Investigating Dissimilarity in Spatial Area Data Using Bayesian Inference: The Case of Voter Participation in the Philippine National and Local Elections of 2016

Francisco N. de los Reyes
School of Statistics
University of the Philippines Diliman

Abstract

A commonly studied characteristic of area data is the assessment of similarity (or absence thereof) among neighboring areal units. However, most methodologies do not measure uncertainties which are likely outcomes of sampling variation and do not consider spatial autocorrelation. This paper explores the ability of Bayesian modeling to address the said situation. It applies this modeling technique to the voting participation statistics in the Philippine National and Local Elections of 2016.

Keywords: *conditional autoregressive (CAR), proximity matrix, dissimilarity, voter turnout*

1. Introduction

Many inquiries in statistics are interested in determining heterogeneity in some population. Dissimilarity is one such measure. It is the extent to which two or more groups are integrated or isolated. The most popular metric is the Dissimilarity Index. However, the Dissimilarity Index has the following inadequacies in spatial data: (1) it does not measure uncertainties which could potentially be a result of random sampling variation, and (2) it does not consider spatial autocorrelation which could be present in the data.

This paper aims to detect dissimilarity in a specific spatial area data: voter participation. In the Philippines, voter turnout is intuitively spatially autocorrelated. There are strong bailiwicks in various corridors in Philippine geography especially in Northern Luzon and Bicol region. There are also strong solid votes in Panay and Negros Islands, another in Cebu and the Davao region. Voter turnout in nearby barangays (the Philippine's basic geopolitical unit) tend to be similar. The same may be opined for larger units like cities and municipalities and even up to the level of the province or region. This paper shall first present a classical method in establishing dissimilarity. However, in consideration of the spatial nature of voter turnout, a Bayesian model shall be used to introduce smoothing in the presence of spatial autocorrelation.

2. The Dissimilarity Index

Let the data be denoted by $\underline{Y} = (Y_1, . . . , Y_n)$ and $\underline{N} = (N_1, . . . , N_n)$, which respectively denote the number of people who voted and the number of registered voters for each of the *n* areal units. Here, the areal units are the provinces, both regular and special provinces as determined by the Commission on Elections (COMELEC), as well as the districts in the National Capital Region. Define voter turnout as the proportion of registered voters who actually voted. Let $\underline{p} = (p_1, . . . , p_n)$ denote the true voter turnout in each areal unit. The **Dissimilarity index** is given by (Lee et al, 2015)

$$D = \sum_{k=1}^{n} \frac{N_k |p_k - p|}{2Np(1-p)}$$

where $N = \sum_{k=1}^{n} N_k$ and p are the total population of registered voters and overall voter turnout in 2016 for the entire Philippines. The value of *D* lies in the interval [0, 1], where zero conveys parity and one means full disparity (or segregation). The unknown true proportions are typically estimated by their sample equivalents, that is $\hat{p}_k = Y_k / N_k$ and $\hat{p} = (\sum_{k=1}^{n} Y_k) / \sum_{k=1}^{n} N_k$. Sampling variation is clearly present if (Y_k, N_k) emanates from a survey, since they are based on a random sample in areal unit *k*. It should be emphasized that elections data is practically census data, which is not obtained from a survey, so that "variation" is essentially due to measurement error. Other variation may be alluded to misreporting, misrecording or computation as in the case of manual tallying.

3. Bayesian Modelling

The estimator $\hat{p}_k$ is both the method of moments estimator and the maximum likelihood estimator under the model $Y_k \sim Binomial(N_k, p_k)$. However, this model assumes that data among areal units are independent, something which is not valid in the presence of spatial autocorrelation. To accommodate this dependence, a Conditional Autoregressive (CAR) model will be used to model the spatial autocorrelation in the data. In this study, the methodology proposed by Lee, Minton and Pryce (2015) was followed. Lee, et al. proposed a global smoothing model for spatially autocorrelated data using a binomial generalized linear mixed model (GLMM), where the random effects are spatially autocorrelated. The full model is given by (Lee et al, 2015):

$$Y_k \sim Binomial(N_k, p_k)$$

$$ln\left(\frac{p_k}{1-p_k}\right) = \beta_0 + \varphi_k \,;\; \underline{\varphi} \sim N(\underline{0}, \tau^2 Q(\rho, W)^{-1})$$

$$\beta_0 \sim N(0, C),\; C\; constant$$

$$\tau^2 \sim Inverse\; Gamma(a, b)$$

$$\rho \sim Uniform(0,1)$$

The random effects $\underline{\varphi} = (\varphi_1, \ldots, \varphi_n)$ shall account for the spatial dependence in the data, and are represented by a CAR prior distribution. Despite the fact that voter turnout areal data is deemed constituted by the totality of all areal units, the randomness of spatial effects is asserted mainly due to possible inaccuracies and unaccounted variation in the reported turnout, and to some extent, the selection of the proximity matrix that accounts for the contiguity structure. Thus, these effects are safer assumed to vary in a range governed by stochastic behavior and not as fixed effects. Moran's Index was used to confirm if spatial autocorrelation exists. The CAR priors shall induce the spatial autocorrelation by a binary $n \times n$ proximity matrix $W = (w_{ki})$, which is computed from the contiguity structure of the n areal units. Based on W, the CAR priors take the form of a zero-mean multivariate Gaussian distribution, where spatial autocorrelation is induced via the precision matrix that depends on W. Leroux et al. (1999) proposed that the strength of the autocorrelation be estimated from the data. The precision matrix for this model involves an autocorrelation parameter and the proximity matrix and is given by

$$Q(\rho, W) = \rho(diag(W\underline{1}) - W) + (1 - \rho)I,$$

where I is an $n \times n$ identity matrix, $\underline{1}$ is an nx1 vector of ones, and $diag(W\underline{1})$ is a diagonal matrix with elements equal to the row sums of W. The matrix $Q(\rho, W)$= $\rho(diag(W\underline{1}) - W) + (1 - \rho)I$ is proper if $\rho \in [0, 1)$, and the spatial structure amongst φ can be observed more clearly from the univariate full conditional distributions

$$\varphi_k | \underline{\varphi}_{-k} \sim Normal\left(\frac{\rho \sum_{i=1}^{n} w_{ki}\varphi_i}{\rho \sum_{i=1}^{n} w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{i=1}^{n} w_{ki} + 1 - \rho}\right)$$

where $\underline{\varphi}_{-k}$ denotes the vector of random effects except for φ_k. Note that by nature and construction of W above, $w_{ki} = 0$ whenever $k = i$, thus φ_k has weight zero in the location parameter of its full conditional distribution. The parameter ρ controls the spatial autocorrelation structure, with $\rho = 1$ corresponding to strong spatial autocorrelation, while $\rho = 0$ corresponds to independent random effects (Besag et al, 1991). The effects have constant mean and variance. Weakly informative prior distributions are assigned to the other hyperparameters so as to allow for estimates of these parameters to be determined from the observed data and not to skew the analysis. The intercept coefficient in the logit function was assigned a univariate Normal distribution with mean zero and homoskedastic. The hyperparameter τ^2, accounting for the variation in the spatial effects are assigned the inverse-gamma. The spatial autoregression parameter is assumed to be uniform over (0,1) since it is over this support that the precision matrix is also deemed proper.

The posterior distribution for the dissimilarity index D (Lee et al, 2015) can be computed using M Markov chain Monte Carlo (MCMC) samples from the posterior distribution

$$\left\{\boldsymbol{\Theta}^{(j)}\right\}_{j=1}^{M} \text{where } \boldsymbol{\Theta}^{(j)} = \left(\phi^{(j)}, \beta_0^{(j)}, \tau^{2(j)}, \rho^{(j)}\right).$$

In the analysis, three values of M were used: 20 thousand, 30 thousand, and 40 thousand all with 50 percent burn-in. The posterior samples are then used to construct samples $\boldsymbol{p}^{(j)} = (p_1^{(j)},\ldots,p_n^{(j)})$, using the inverse logit transform

$$p_k^{(j)} = exp\left(\beta_0^{(j)} + \varphi_k^{(j)}\right)/[1 + exp\left(\beta_0^{(j)} + \varphi_k^{(j)}\right)].$$

The jth sample from the posterior distribution of D is constructed as

$$D^{(j)} = \sum_{k=1}^{n} \frac{N_k|p_k^{(j)} - p^{(j)}|}{2Np^{(j)}(1-p^{(j)})},\ j = 1, \ldots, M \text{ where } p^{(j)} = (\textstyle\sum_{k=1}^{n} N_k p_k^{(j)})/(\textstyle\sum_{k=1}^{n} N_k).$$

Finally, D can be estimated by the median of $\{D^{(1)},\ldots, D^{(M)}\}$, while a 95 percent credible interval is obtained from the 2.5th and 97.5th quantiles of $\{D^{(1)},\ldots, D^{(M)}\}$.

4. Modelling Voter Participation in the Philippine National and Local Elections (NLE) of 2016

Official data from the COMELEC was used in the research. Since voter turnout is viewed here in a spatial data analysis paradigm, and therefore contiguity-sensitive, turnout from overseas voting was not included. Provincial level information on number of registered voters and actual voter turnout for 86 areal units comprise the entirety of the dataset. This includes special readings for the cities of Isabela and Cotabato, which are labeled special provinces, and the four districts of Manila. Proximity due to common-border cannot be used since there are 15 island provinces, which are as follows: Batanes, Biliran, Bohol, Camiguin, Catanduanes, Cebu, Dinagat, Guimaras, Marinduque, Masbate, Palawan, Romblon, Siquijor, Sulu, and Tawi-Tawi. Here, connectivity was based on a nominated critical distance. Special consideration arose for the island of Palawan since a large critical distance was needed for it to have just one neighbor. Thus, for this specific province, indication of geographic integration like presence of boat routes and trade with a nearby province was used. This led to Iloilo being set as Palawan's neighbor. The proximity matrix was then revised to force a neighbor for Palawan.

Areal centroids were identified via the Universal Transverse Mercator (UTM) coordinate system. Inter-unit distance was computed via these coordinates. Created a proximity matrix W based on L_1 distance of at most a nominated δ.

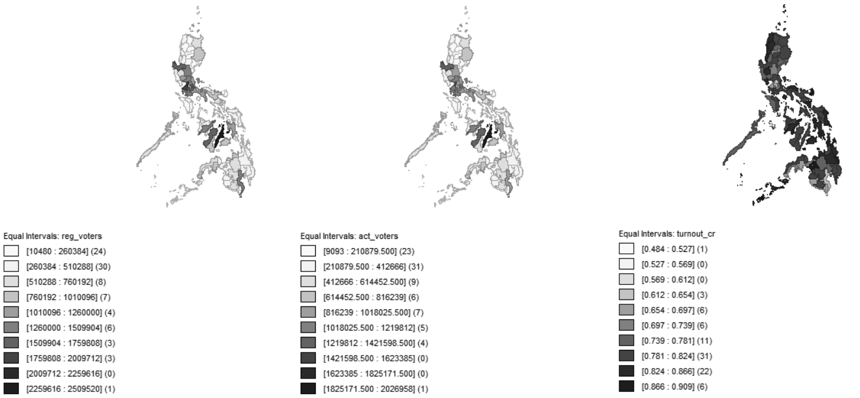
$$W = \{w_{ik}\} \text{ has } w_{ik} = \begin{cases} 1, & d_{L_1}(A_i, A_j) \leq \delta \\ 0, & d_{L_1}(A_i, A_j) > \delta \end{cases} \quad ; \quad w_{ii} = 0$$

Several values were tried for δ but only in $\delta = 500$ UTM units did all areal units, except Palawan, that had at least one neighbor. Iloilo was forced to become Palawan's neighbor by virtue of transportation and trade relations. A distance decay term was introduced to the spatial variance matrix in accordance to Tobler's principle. The term is $g(d_{ij}) = e^{-3d_{ij}}$ where d_{ij} is the inter-centroid distance between areal units i and j. A piecewise mean function was also generated for two clusters: voter turnout greater than 80 percent where the mean turnout is 83 percent and voter turnout less than 80 percent where the mean turnout is 73 percent. The logit transform of the proportions are generated from a multivariate Gaussian distribution with a mean of 0.99 (logit corresponding to the mean turnout of 73 percent) or 1.73 (logit corresponding to the mean turnout of 83 percent). Low, moderate, and high (on account of Cotabato City) spatial variation scenarios were investigated. The different spatial variation scenarios were included to see how the dissimilarity index and the parameter estimates of the hierarchical Bayesian model behave. This was done mainly to assess the performance of the model in three situations. Low spatial variation assumes that nearby areas share generally similar turnout while large variation assumes that nearby areas tend to have widely dissimilar turnout. Moderate spatial variation assumed a middle ground. The Bayesian model was fitted at ***M***=20,000 ,30,000 *and* 40,000 MCMC samples. Parameter estimation proceeded after 50 percent burn-in. Model fit was assessed by the width of the credible intervals, variability of residuals and the Deviance Information Criterion (DIC). The DIC was particularly chosen since it performs like an Akaike Information Criterion (AIC) when comparing and selecting among several hierarchical Bayesian models whose posterior distributions emanate from MCMC samples (Spiegelhalter et al, 2002; Berg et al, 2004; Ando, 2007). The DIC also promotes simplicity (i.e. parsimony) in resulting models since it incorporates a penalty for model complexity (i.e. with more parameters). Note that since the values of DIC are not normed; models with relatively smaller DIC are favored. Narrower credible intervals, lower standard deviation of residuals, and lower DIC signify good model fit. Data integration, computation of proximity matrix, and testing for spatial autocorrelation were done in Geoda. Modelling was done in R CARBayes package with extensive use of the **S.CARleroux** function.

5. Results

Voter turnout was generally high in the NLE of 2016 as indicated by an average of 79 percent across the 86 areal units (Figure 1). The special province of Cotabato City was outlying with a turnout of only 48 percent. There was a significant positive spatial autocorrelation in voter turnout (Moran's I = 0.22, p = 0.0067). This indicates that areas with relatively higher voter turnout are spatially close. A similar conclusion can be said of areas with relatively lower voter turnout. There is a suggestion of parity in voter participation across provinces as evidenced by a dissimilarity index of D=0.15. The 95 percent confidence interval is (0.113, 0.175) based on 10,000 bootstrap samples.

Figure 1. Registered Voters, Actual Voters, and Voter Turnout in the Philippine National and Local Elections of 2016



The hierarchical model had poor fit under the assumption of high spatial variation scenario within clusters given the official election voter turnout data (Table 1). Here, the standard deviation of residuals and DIC are highest within tiers of MCMC sample. The width of the 95 percent credible interval for the dissimilarity index, intercept term for the logit expression, τ^2 and ρ are generally largest.

The hierarchical model had relatively better fit under the assumption of moderate spatial variation scenario as compared to the model, given a high spatial variation assumption. In this scenario, the standard deviation of residuals and DIC are lower within tiers of MCMC sample. The width of the 95 percent credible interval for the dissimilarity index, intercept term for the logit expression, τ^2 and ρ tend to be narrower. The hierarchical model showed best fit under low spatial variation scenario within clusters, given the official election turnout data. The 95 percent confidence intervals are narrowest within each group of MCMC samples. Residual variability is at its least and so is the DIC. Estimates seem to have good precision at M=30,000 MCMC samples (50 percent burn-in). Here, the spatial autocorrelation parameter can reasonably be expected to fall in the interval (0.001, 0.181).

The dissimilarity indices generated across all scenarios are relatively small. These values signify that variability in voter participation is indeed small across provinces. There is generally high turnout nationwide with provincial rates which are not far from this general average.

Table 1. Results of Bayes Modelling of Voter Turnout in the Philippine National and Local Elections of 2016

POSTERIOR QUANTITIES AND MODEL FIT												
	Scenario A: Low spatial variation											
	M=20,000 (50% Burn In)				M=30,000 (50% Burn in)				M=40,000 (50% Burn in)			
	Median	L95	U95	width	Median	L95	U95	width	Median	L95	U95	width
Dissimilarity, D	0.103	0.097	0.121	0.024	0.141	0.137	0.142	0.005	0.120	0.119	0.120	0.001
Intercept	1.097	1.095	1.109	0.014	1.099	1.098	1.101	0.003	1.114	1.113	1.115	0.002
tau-square	0.191	0.113	0.619	0.506	0.221	0.118	0.875	0.757	0.194	0.114	0.874	0.760
rho	0.014	0.000	0.109	0.109	0.022	0.001	0.181	0.180	0.015	0.001	0.171	0.170
SD residuals	184.67				168.22				181.97			
DIC	3011698				2515216				2914018			
	Scenario B: Moderate spatial variation											
	M=20,000 (50% Burn In)				M=30,000 (50% Burn in)				M=40,000 (50% Burn in)			
	Median	L95	U95	width	Median	L95	U95	width	Median	L95	U95	width
Dissimilarity, D	0.098	0.097	0.099	0.002	0.092	0.092	0.098	0.006	0.103	0.102	0.103	0.001
Intercept	1.292	1.291	1.293	0.002	1.286	1.283	1.287	0.004	1.305	1.304	1.306	0.002
tau-square	0.250	0.139	1.424	1.285	0.223	0.124	1.069	0.945	0.253	0.143	1.165	1.022
rho	0.017	0.001	0.240	0.239	0.019	0.000	0.205	0.205	0.018	0.001	0.184	0.183
SD residuals	222.96				204.17				216.36			
DIC	4313926				3629979				4053158			
	Scenario C: High spatial variation											
	M=20,000 (50% Burn In)				M=30,000 (50% Burn in)				M=40,000 (50% Burn in)			
	Median	L95	U95	width	Median	L95	U95	width	Median	L95	U95	width
Dissimilarity, D	0.223	0.222	0.228	0.006	0.207	0.206	0.208	0.002	0.209	0.203	0.213	0.010
Intercept	1.210	1.208	1.212	0.004	1.223	1.214	1.225	0.011	1.217	1.215	1.218	0.003
tau-square	1.406	0.449	7.197	6.748	1.784	0.545	6.326	5.781	0.864	0.327	4.425	4.098
rho	0.113	0.014	0.743	0.729	0.169	0.026	0.669	0.643	0.063	0.005	0.452	0.447
SD residuals	238.28				229.23				219.91			
DIC	4884536				4514759				4165975			

6. Conclusions, Recommendations and Learnings

Both Bayesian and non-Bayesian models reveal that there is low dissimilarity in voter turnout among the 86 areal units contained in the official COMELEC dataset. When spatial variation is taken into account, there is sufficient basis to say that the spatial variation is low. Thus, it is clear that Filipinos participated well in the National and Local Elections of 2016 and quite consistently homogeneous in pattern if taken spatially. As to the statistical specification of the model, the case of the Philippines requires critical distance of 500,000 UTM units to assure that areal units have at least one neighbor based on inter-centroid. A proximity matrix can still be constructed for a critical distance lower than this value but the algorithm fails to converge due to provinces without neighbors. For the case of Palawan, one needs to override the generated proximity matrix to force a neighbor under some special criterion (here, transportation and trade relation). Localized smoothing is beyond the scope of this study and is a suggested improvement moving forward. In addition, the dissimilarity index in voter turnout should be tracked over time to validate if the Philippine electorate is indeed participative. The technique presented here may be also be applied to other areal information with inherent spatial variation like poverty and health statistics where strong spatial components are expectedly inherent.

References:

Ando, Tomohiro., 2007. Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. Biometrika. 94 (2): 443–458.

Bailey, Trevor C. and Gatrell, Anthony C., 1995. Interactive Spatial Data Analysis. Longman Group Limited. New York.

Berg, Andreas; Meyer, Renate; and Yu, Jun. Deviance Information Criterion for Comparing Stochastic Volatility Models, 2004. Journal of Business and Economic Statistics. 22, (1), 107-120.

Besag,J.,York,J.,Mollie,A., 1991. Bayesian image restoration with two applications in spatial statistics. Ann.Inst.Statist.Math. 43,1–59.

Bivand, Roger S., Pebesma, Edzer J. and Gomez-Rubio, Virgilio, 2103. Applied Spatial Data Analysis with R. Springer. New York.

Cressie, Noel, A., 1993. Statistics for Spatial Data. John Wiley & Sons, Inc. New York.

Duncan Lee, Jon Minton, Gwilym Pryce, 2015. Bayesian inference for the dissimilarity index in the presence of spatial autocorrelation. Spatial Statistics 11 (2015) 81–95.

Leroux, B., Lei, X., Breslow, N., 1999. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: Halloran, M., Berry, D. (Eds.), Statistical Models in Epidemiology, the Environment and Clinical Trials. Springer-Verlag, New York, pp. 135–178. (Chapter).

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL: http://www.R-project.org.

Spiegelhalter, David J.; Best, Nicola G.; Carlin, Bradley P.; van der Linde, Angelika. 2002. "Bayesian measures of model complexity and fit (with discussion)". Journal of the Royal Statistical Society, Series B. 64 (4): 583–639.