Proceedings of The National Conference
On Undergraduate Research (NCUR) 2015
Eastern Washington University, Cheney, WA
April 16-18, 2015

An Analysis of Bullying and Suicide in the United States using a Non-Gaussian Multivariate Spatial Model

Mary Ryan
Department of Statistics
University of Missouri
Columbia, MO 65211

Faculty Advisors: Jonathan R. Bradley, Trevor Oswald
Christopher K. Wikle, and Scott H. Holan

Abstract

Bullying affects thousands of students across the United States (US) each year, which can lead to mental health problems, and in some cases suicide. Intuitively, rates of bullying and suicide may be correlated, and this relationship may change based on region. Thus, a spatial analysis of bullying and suicide rates could help identify regions where more attention is needed to prevent both bullying and suicide. As such, we develop a non-Gaussian multivariate spatial model to analyze bullying and suicide rates in the US. This model incorporates the right-skewed nature of bullying and suicide rates, and leverages multivariate spatial dependence to improve spatial predictions. We apply our statistical model to data obtained from the Centers for Disease Control and Prevention's (CDC) Youth Risk Behavior Surveillance System (YRBSS). In particular, we consider YRBSS estimates of attempted suicide and (self-reported) bullying rates (per 100 thousand) over the 48 contiguous states. Our model provides accurate spatial predictions for suicide and bullying rates, while providing spatial prediction variances. These results indicate regions that have higher rates of bullying and suicide, which can have implications on policy decisions.

Keywords: Bayesian, Spatial, Youths

1. Introduction

Bullying impacts thousands of students across the United States (US) each year, and leads to lower self-esteem, higher prevalence of mental health issues, and cases of suicide[6, 9]. In fact, in 2013 19.6% of high school-aged youths have been bullied on school property (based on data from the Centers for Disease Control and Prevention (CDC)), and suicide is the third leading cause of death among 12 to 19 year olds[8]. Each of these studies provide global measurements (i.e., percent over the entire US), but the (US) regional patterns have been given less attention in the literature. Intuitively, rates of bullying and suicide may be correlated, and this relationship may change based on region due to cultural factors[6,13].

A spatial analysis of bullying and suicide rates could help identify regions where more attention is needed to prevent both bullying and suicide, or regions with particularly low rates that may have policies worth studying and implementing (e.g., see American Foundation for Suicide Prevention, 2015, for current policies on bullying and suicide prevention). To learn about this relationship, we apply a Bayesian model to 2013 data obtained from the CDC's Youth Risk Behavior Surveillance System (YRBSS). Within the dataset of interest, we look at the estimated percent of self-reported bullying on school property and the estimated percent of attempted suicide over the 48 contiguous states among high school-aged youths. We chose to model attempted suicide in this study as opposed to other YRBSS suicide-related variables (e.g., successful suicide, seriously considered attempting suicide, felt sad or hopeless) because the population is larger and, thus, more straightforward to model[2]. Furthermore, those who attempted suicide are likely to attempt again[9], and any regional patterns learned in this study may help with prevention measures. As

seen in Figure 1, each set of estimates is incomplete, which makes spatial patterns difficult to discern and the sets difficult to compare.

This leads us to our primary inferential questions:

1. Is there a relationship between state-level percent bullying and state-level percent attempted suicide?
2. Are there any spatial patterns present in this dataset?
3. What are the proportions of attempted suicide and bullying at states where YRBSS does not release estimates?

In this article, we use spatial statistics to answer each of these important inferential questions. In particular, we adopt a hierarchical modeling approach that is commonly used for spatial data[5].

In Section 2, we present an exploratory analysis where we will investigate the distributional properties of the percent attempted suicide and percent bullying, statistics measuring the linear and spatial relationships, and summaries identifying cross-correlations between bullying and suicide. We answer Item 1 above by investigating linear relationships between state-level bullying and attempted suicide rates. Then, we answer Item 2 using adjacency matrices and the Geary *C* statistic. In Section 3, we present our Bayesian hierarchical model, which we will use to predict the proportions of bullying and attempted suicide at states where YRBSS does not release estimates. Finally, in Section 4, we will provide results that answer Items 2 and 3 above, and discuss their implications and limitations.

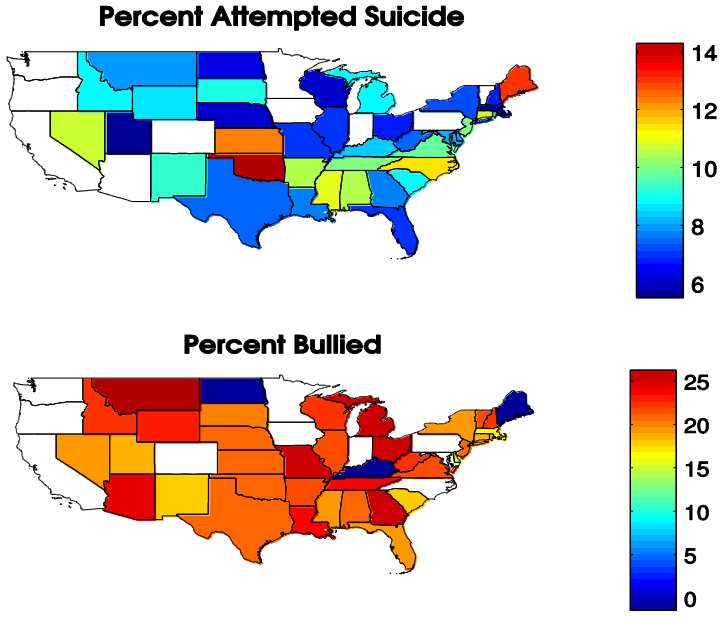

Figure 1: The original percent of attempted suicide (top) and bullying (bottom) among the 48 contiguous states, as provided by YRBSS. Note that white states constitute states with no data released.

2. Exploratory Analysis of YRBSS Estimates of the Percent Reported Bullying and Attempted Suicide

The goal of our study is to apply a Bayesian algorithm to the 2013 YRBSS data (displayed in Figure 1) to address Items 1 through 3 in the list from Section 1. To build this Bayesian model, in a manner that is appropriate for this data, we undertook an extensive exploratory analysis, which we layout in this section. In particular, we set out to determine the distributional properties of the YRBSS data, and the "strength" of the spatial dependencies within, and multivariate dependencies between the percent self-reported bullying and attempted suicide datasets.

To investigate the distributional properties of the proportion of self-reported bullying and attempted suicide, we consider transformations of the data presented in Figure 1. Specifically, let $j = 1$ and $j = 2$ indicate self-reported

bullying and attempted suicide, respectively. Also, let $p_{1,j}, \ldots, p_{n(j),j}$ represent the observed sample proportions for variable j, where $j = 1, 2$, and $n(1) = 41$ and $n(2) = 38$ indicates the number of observed spatial locations for each variable. Then, one might consider the logit transformation:

$$z_{i,j} \equiv \text{logit}(p_{i,j}) \equiv log\left(\frac{p_{i,j}}{1-p_{i,j}}\right); j = 1,2, i = 1, \ldots, n(j). \quad (1)$$

In Figure 2, we provide normal QQ-plots for $\{z_{i,1}\}$ and $\{z_{i,2}\}$, in the left and right panels, respectively. The use of the logit-normal distribution has been successfully used by Mead (1965), and, from Figure 2, appears to be a reasonable assumption for 2013 state-level YRBSS estimates of the percent self-reported bullying and attempted suicide.

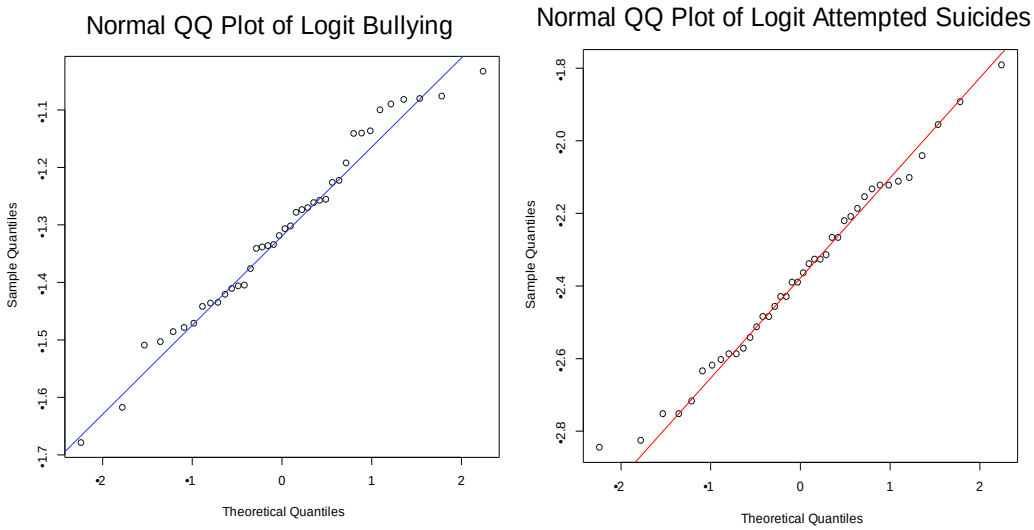

Figure 2: QQ-plots for the logit transformation of percent bullying (left) and percent attempted suicide (right) where it is reasonable to assume that the transformation is normal. Notice that the (logit) transformed data, in each panel, closely follows the 45° line, indicating that the normality assumption appears reasonable.

With the assumption of the logit-normal distribution supported, one can use simple linear regression to explore whether the bullying and attempted suicide datasets have a relationship. The fitted linear regression between logit-suicide (response) and logit-bullying (explanatory variable) is written as,

$$z_{i,2} = 0.09 z_{i,1} - 2.25. \quad (2)$$

Table 1: ANOVA table results for the simple linear regression between attempted suicide and bullying.

	DF	SSE	MSE	F-value	*p*-value
Logit Percent Bullied	1	0.00784	0.00784	0.1225	0.7283
Error	36	2.30348	0.06399	—	
Total	37	2.31132	—		

The ANOVA results associated with this simple linear regression are given in Table 1, and suggest there is not enough evidence to claim a direct linear relationship between bullying and attempted suicide (p-value = 0.73 > 0.05).

Furthermore, the correlation coefficient is equal to 0.03, suggesting a weak correlation. The slope of the regression model is positive (i.e., 0.09 > 0), which reflects the intuition that an increase in bullying implies an increase in attempted suicide; however, the results in Table 1 show that this relationship is not significant.

We end this section by testing whether state-level bullying or state-level attempted suicide have spatial patterns within their datasets. To do this, we create an adjacency matrix, which we denote **W**. For our application, **W** is a 48-dimensional symmetric matrix, where the (*i, j*)-th element, denoted as $w_{i,j}$, is set equal to 1 if state *i* is considered close to state *j*; *i, j* = 1, …, 48. If state *i* is *not considered close* to state *j*, $w_{i,j}$ is set equal to zero. For example, if state *i* is a nearest neighbor of state *j* then $w_{i,j} = 1$ and 0 otherwise (one could also allow for general *k*-nearest neighbors if desired).

This 48 × 48 adjacency matrix **W** is put to use in a myriad of ways throughout spatial statistics[5], but here we will use it in the Geary *C* statistic. The Geary *C* statistic is a measure of spatial association, and it is given by[3],

$$C = \frac{(n-1)\sum_i \sum_j w_{i,j}\left(z_{k,i} - z_{k,j}\right)^2}{2\left(\sum_{i \neq j} w_{i,j}\right)\sum_i \left(z_{k,i} - \bar{z}_k\right)^2}; k = 1,2, \tag{3}$$

where $\bar{z}_k$ is the sample average of $\{z_{k,j} : j = 1, \ldots, n(k)\}$. The Geary *C* statistic is asymptomatically normal with mean 1, under the assumption that the underlying values of $\{z_{k,j} : j = 1, \ldots, n(k)\}$ are i.i.d. (Banerjee et al., 2015). In Table 2, we give the Geary *C* statistic, its asymptomatic variance, and the corresponding *p*-value. Here, we see that there is evidence to claim that there is spatial information in the bullying dataset (*p*-value = 0.0034 < 0.05); however, there is not enough evidence to claim that there is spatial information in the attempted suicide dataset (*p*-value = 0.1223 > 0.05).

Table 2: Geary C statistics for logit percent bullying and logit percent attempted suicide, with asymptomatic variances and p-values.

	Geary C	Variance	*p*-value
Logit Percent Bullied	0.6473	0.0169	0.0034
Logit Attempted Suicide	0.8515	0.0163	0.1223

3. The Bayesian Hierarchical Model

To predict the proportion of bullying and attempted suicide for states that did not provide data, we build a Bayesian hierarchical model. Such a model will allow us to account for all of the uncertainty in the underlying processes (i.e., the true value of bullying and attempted suicide) and the uncertainties of parameters within our model.

The results from Section 2 lead naturally to the following Bayesian hierarchical model:

$$z_{i,1} = \mu_1 + \varepsilon_{i,1}; i = 1, \ldots, 41, \tag{4}$$

$$z_{j,2} = Y_j + \varepsilon_{j,2}; j = 1, \ldots 38, \tag{5}$$

where random variables $\{z_{i,1}\}$ and $\{z_{j,2}\}$ are assumed to be independent, and $\{\varepsilon_{i,1}\}$ and $\{\varepsilon_{j,2}\}$ are independent and identically distributed (i.i.d.) normal with mean zero and variance $\sigma_1^2 > 0$ and $\sigma_2^2 > 0$, respectively. The 48-dimensional process vector $\boldsymbol{Y} \equiv (Y_1, \ldots, Y_{48})'$ follows a Gaussian distribution:

$$\boldsymbol{Y}|\mu_2, \sigma_Y^2 \sim Gaussian(\mu_2 \mathbf{1}, \sigma_Y^2(\boldsymbol{I} - \boldsymbol{W})^{-1}),$$

where μ_2 is a constant, **1** is a 48-dimensional vector of ones, and $\sigma_Y^2 > 0$. This is called an intrinsic conditional autoregressive model (ICAR), and will help incorporate spatial information into our predicted state-level estimates[3, 5]. Notice that Equation (4) follows from the results from Section 2 — that is, the Geary *C* statistic and normal QQ-plots suggest that the logit-suicide data follows an i.i.d. normal distribution. Equation (5) is supported by the normal QQ-plot in Figure 1 and the Geary *C* statistics in Table 2, which suggests that logit-bullying follows a model with spatial dependence (e.g., an ICAR model).

The parameters for our model are given the prior distributions:

$$\sigma_1^2 \sim IG(1,1),$$

$$\sigma_2^2 \sim IG(1,1),$$

$$\sigma_Y^2 \sim IG(1,1),$$

$$\mu_1 \sim Normal(0, 100),$$

$$\mu_2 \sim Normal(0, 100).$$

We chose our prior distributions for our hyperparameters to be relatively flat. This is because we do not have any additional information on what the value of these parameters are; that is, if additional studies suggest values for these parameters, we would center these distributions around those values.

Finding a closed form expression of the posterior distribution is difficult for this statistical model. In order to combat this, we use an algorithm called the Gibbs Sampler[10] to obtain simulated replicates from the following posterior distributions:

$$\mu_1 | Z_{1,1}, .. Z_{41,1}, \tag{6}$$

$$Y_j | Z_{1,2}, .. Z_{38,1}; j = 1, \dots, 48. \tag{7}$$

The closed form expressions of the full-conditionals associated with this Gibbs sampler can be found in Robert and Casella (2004) and Christensen et al. (2010). Notice that $\boldsymbol{Y}$ is 48-dimensional and, thus, leads to an estimate of logit-bullying at each of the states in the contiguous US. Additionally, we assume (via Equation (4)) that each state within the continental US has the same percentage of self-reported attempted suicide (i.e., μ_1).

4. Results and Discussion

As we learned in Section 2, there does not appear to be any direct linear relationship between self-reported bullying and attempted suicide rates. While this may initially surprise some, it is logical in a larger sense. There are many reasons why someone might attempt suicide beyond being bullied — preexisting mental illnesses, destitute economic and financial situations, or distressing emotional events such as parental divorce or a difficult personal breakup could all contribute to the decision to commit suicide. Because bullying is just one factor in a list of many possible confounding factors, its ability to accurately predict suicide rates could be seen as minimal in a broader context. Therefore, to develop a more meaningful relationship between bullying and attempted suicide, one would need to consider additional confounding effects to isolate the effect of bullying on attempted suicide.

We were able to detect spatial dependence among the 41 logit-bullying values. The presence of this dependence allows us to obtain estimates for states that do not have YRBSS estimates. In Figure 3, the predicted bullying estimates are similar for much of the inner contiguous US, and increase as you move toward the eastern and western coasts. The squared prediction errors follow a similar spatial pattern. The way our model was constructed, the estimate of each state is determined by leveraging strength from its neighboring states and incorporating that information into the state's latent process (i.e., $\boldsymbol{Y}$).

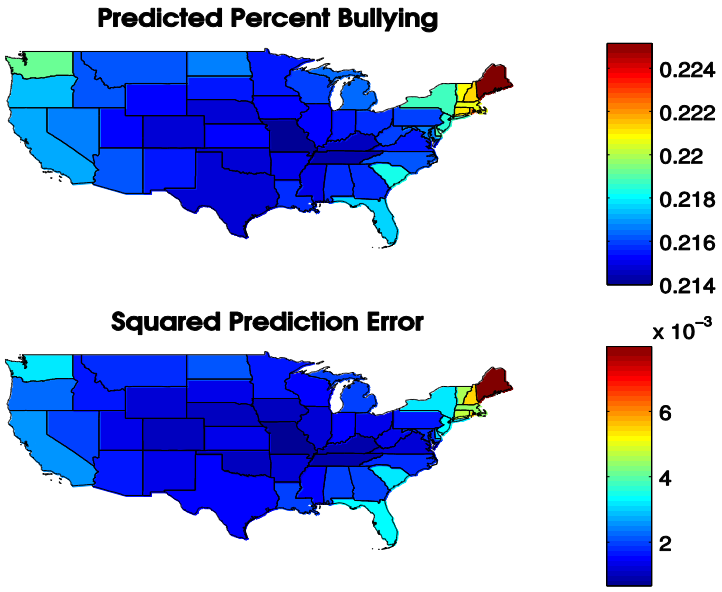

Figure 3: Predicted state-level percentage estimates for bullying (top) and the squared prediction error (bottom) as outputted by our hierarchical model.

For the attempted suicide dataset, we were able to use the model to estimate the unobserved mean assumed to be constant across each state within the contiguous US. In Figure 4, we display a histogram over replicates from the posterior distribution of the mean logit-attempted suicide rate. Overall, our estimate of the mean logit-attempted suicide rate is consistent with the national average of 8.539 percent. Thus, we conclude that our model is relatively accurate because this average is close to the original YRBSS national estimate of 8 percent.

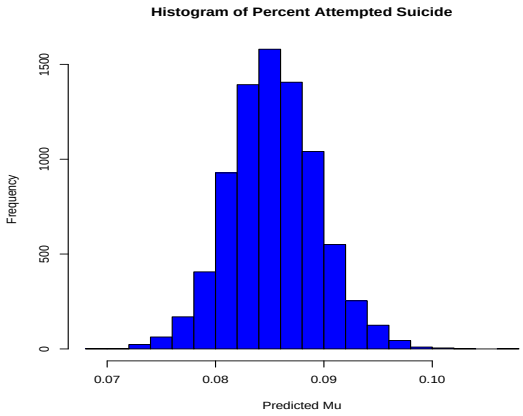

Figure 4: A histogram of the predicted attempted suicide mean percentage centers the national average around 8.539 percent, close to the original YRBSS national estimate of 8 percent.

As can be imagined, there are some limitations that could affect the total accuracy of our model. First, in the state-level bullying model, many of the states that were missing estimates were first-order neighbors. For example, it may be incorrect to conclude that Maine has a higher bullying rate relative to its neighbors; if many of the missing states are located next to each other there is a limited amount of information the model can leverage to estimate the missing state.

Moreover, a phenomenon known as the ecological fallacy could be at work[11]. This means that when looking at estimates on a larger level, such as state-level compared to county-level, some of the intricacies of the spatial patterns are removed upon aggregation of the data. In Figure 3, the larger state averages might not be completely representative of each individual county within the state. For example, Texas is a very large state and what may be happening on the

eastern border may not be the same as what is happening on the western border. To develop a more accurate model for predicting bullying rates, it would be beneficial to analyze the county-level bullying rates and possibly create separate models for different US geographical regions.

5. Acknowledgements

This research was partially supported by the U.S. National Science Foundation (NSF) and the U.S. Census Bureau under NSF grant SES-1132031, funded through the NSF-Census Research Network (NCRN) program.

6. References

1. American Foundation for Suicide Prevention (2015). "State Policy." Retrieved May 18, 2015, from http://www.afsp.org/advocacy-public-policy/state-policy
2. Agresti A. (2007). *Categorical data analysis*. rev edn. New York, NY: John Wiley & Sons.
3. Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*, rev. edn. London, UK: Chapman and Hall.
4. Christensen, R., Johnson, W., Branscum, A., and Hanson T.E. (2010). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. Boca Raton, FL: CRC Press.
5. Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Hoboken, NJ: Wiley.
6. Hinduja, Sameer, and Justin W. Patchin. "Bullying, Cyberbullying, and Suicide." *Archives of Suicide Research* 14, no. 3 (July 28, 2010): 206–21. doi:10.1080/13811118.2010.494133.
7. Mead, R. (1965). "A Generalised Logit-Normal Distribution." *Biometrics*, 21: 721–732.
8. Miniño A.M. (2010). "Mortality among teenagers aged 12-19 years: United States, 1999-2006." *NCHS data brief*: 37. Hyattsville, MD: National Center for Health Statistics.
9. Rigby, K. and Slee, P. (1999), "Suicidal Ideation among Adolescent School Children, Involvement in Bully—Victim Problems, and Perceived Social Support." *Suicide and Life-Threat Behavior*, 29: 119–130.
10. Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer-Verlag
11. Robinson, S. (1950). "Ecological correlations and the behavior of individuals." American Sociological Review, 15, 351–357.
12. Tuckman, Dr. Jacob, Mr. William F. Youngman, and Mr. Garry Kreizman. "Multiple Suicide Attempts." *Community Menal Health Journal* 4, no.2 (April 1, 1968): 164-70. doi: 10.1007/BF01530700
13. Williams, Ruth FG, D. P. Doessel, and Jerneja Sveticic. "Are there regional disparities in suicide rates? Quantifying suicide distributions for Queensland, 1990-2007." *School of Economics La Trobe University Working Paper Series* 2 (2012): 12.