A Spatial Model of Home School Enrollment Trends with Missing and Censored Data

Nick Grener

November 14, 2024

1 Motivation

A recent article in *The Washington Post* (Jamison et al. 2023) describes a widespread increase in home schooling over the last decade in America, and furthermore indicates a 54% increase of home school enrollment in the state of Ohio (my home state) over the time frame from 2017–2022. Although this phenomenon was accelerated during the Covid-19 pandemic, the trend predates, and has persisted beyond, 2020. However, *The Washington Post*'s analysis was unable to identify any important drivers of this increase, instead concluding that, "Home schooling's surging popularity crosses every measurable line of politics, geography and demographics." For my creative component project (STAT 6990), I wanted to investigate this phenomenon at the level of local school districts. This project provided me with an opportunity to learn about several important concepts that I had not previously encountered in my coursework, including modeling spatial areal data, missing data mechanisms, Bayesian inference (including MCMC techniques), and spatio-temporal residual autocorrelation, via the analysis of an interesting real-world data set. In this report, I will explain how I developed and evaluated a succession of linear spatial models in an effort to identify which predictors were associated with the recent trends in home school enrollment. In contrast to the newspaper's findings, I discovered that - at least in the case of Ohio- some variables were significantly associated with the increase in homeschooling at the district level.

2 Data Procurement and Associated Challenges

All data used in this analysis is publicly available and was freely obtained; the article in *The Washington Post* linked to a GitHub repository (Hoyer 2023) from which I procured the home school counts, where available, for each school district in Ohio for both of the years 2017 and 2022. The Ohio Department of Education's District Profile Report (commonly referred to as the Cupp Report) (*District Profile Report (Cupp Report)* FY Years 2017 and 2022) was the source for the values of some variables which I hypothesized could potentially be associated with changes in home schooling, including teacher experience levels, per-pupil expenditure, and economic measures for the population in each district. These were supplemented with data regarding attendance rates, graduation rates, and private school enrollments found at the State of Ohio's Education and Workforce Report Portal (*Ohio Department of Education and Workforce Report Portal* n.d.). The remaining information included in the model was provided by the U.S. Census Bureau's American Community Survey (ACS), accessed using the tigris package in R. More details about the relevant predictors can be found in Section 3 below.

Cleaning and merging the data from these disparate sources provided the first challenge of the project. First of all, there were some inconsistencies in the values of key variables used to link data frames. The tigris shapefile of Ohio school districts, which served as the base data frame to which all other data values were attached, consists of 611 rows; however, the supposedly unique identifier known as the information retrieval number (IRN) for 38 of these districts did not match the associated value in the Cupp Reports. Online research and manual adjustments were done to reduce the number of districts without matching data (which were dropped from the analysis) to five, and another three "island districts" located in Lake Erie and consisting of few year-round residents were also excluded from the analysis, following the lead of *The Washington Post*.

Another issue was the change of support between the primary areal unit of interest in this study (school districts) and the areal unit of interest for two salient regressors. Private school counts and political affiliation (as measured by Republican vote share in the 2020 Presidential election) were not available at the level of school districts. Since school districts in Ohio respect county boundaries — that is, no districts straddle two or more counties — the decision was made to assign county-level political affiliation to all districts contained within each county; census tracts would have yielded a finer-grained picture of the political makeup of districts, but unfortunately, there is not an alignment between census tracts and school district boundaries in Ohio. Private school enrollments were proportionally assigned to each district within a county based on the percentage of the overall public school enrollment of the county that each district represented. Geographical coordinates and school-level enrollments possibly could have been used to allocate most of these private school students to a public school district of home residence, but the rule considered for divvying up students in such a way provided no greater confidence than the (much more direct) choice of county-wide proportional allocation, so — at the admitted risk of potentially committing a statistical ecological fallacy — the latter approach was taken.

Of greatest theoretical interest with respect to the challenges encountered during the data cleaning portion of the project was the presence of missing and interval-censored home school enrollment data. Of the 603 verified districts that could be matched from the tigris ACS shapefile to the data from the Ohio Department of Education, nearly 23 percent had home school counts that were either missing or censored for at least one of the years of interest; Figure 1 shows the geographic distribution of these districts. Any district whose home school count for a year was in the single digits was recorded as "< 10". Gelman et al. (2020) advocate for accounting for the process of data collection in a rigorous analysis of such datasets: "If partial information is available (for example, knowing that a measurement exceeds some threshold but not knowing its exact value, or having missing values for variables of interest), then a probability model should be used to relate the partially observed quantity or quantities to the other variables of interest" (pg. 197). Therefore, a primary consideration in the model building process was to incorporate these attributes of the dataset in the inference process.

3 Model Building

First Steps The focus of this analysis was to identify the relative importance of a set of proposal covariates on a response variable in the presence of possible spatial dependence. The ultimate goal was to build a spatial mixed linear model with a conditional autogressive (CAR) formulation, where each estimated element was specified conditionally on the values of its neighbors. As Ver Hoef et al.

(2018) point out, "for Bayesian Markov Chain Monte Carlo methods, CAR models are ready-made for conditional sampling because of their conditional specification" (pg. 45). However, in an effort to first learn Markov Chain Monte Carlo (MCMC) techniques, and to provide a baseline model against which more complex models could be compared, an initial model was considered which made no effort to incorporate missing data mechanisms or spatial dependence; this simple approach is described in this section.

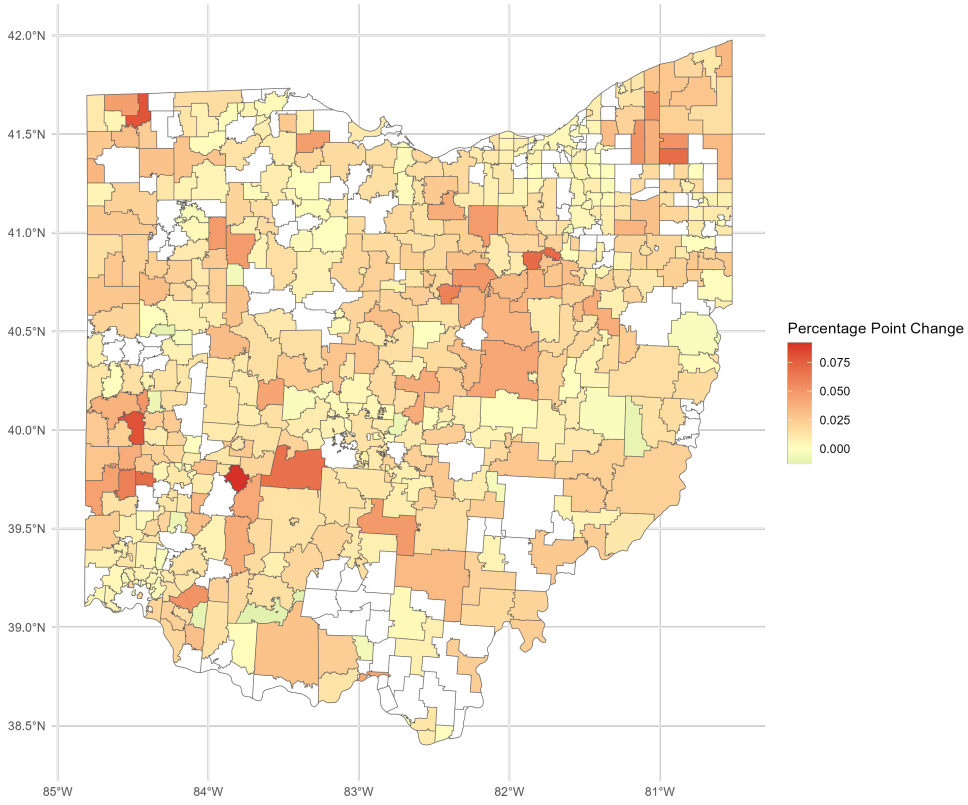

Figure 1: Difference in Annual Proportions of Home Schooled Students, 2017 vs. 2022. Districts with missing or censored data are shown in white

As presented, the response variable consisted of two columns of count data: the number of school-age children in each public school district who were registered with the state of Ohio as home schooled in each of the years 2017 and 2022. Because the interest was in discovering *trends* in the increase of this phenomenon over these years, the response variable was transformed to a single column that estimated the change in the proportion of students who were home schooled in each district. This was calculated using the formula

$$Z_s = \frac{Y_{s,1}}{n_{s,1}} - \frac{Y_{s,2}}{n_{s,2}}, \tag{1}$$

where the Z_s represent each district's change in the proportion of students that are home schooling, $i = 1$ represents the year 2022, $i = 2$ represents the year 2017, the $Y_{s,i}$ values are the counts of registered home school students in district s, and the $n_{s,i}$ values were estimated (using the county proportional allotment described above) via the formula

$$n_{s,i} = \text{public enrollment in year } i + \text{home school enrollment in year } i + \text{est. private enrollment} \tag{2}$$

This choice allowed for a pseudo-Gaussian framework for inference; namely, the initial model I considered takes the familiar form,

$$Z_s \overset{ind}{\sim} \mathbf{N}(\boldsymbol{x}_s^T \boldsymbol{\beta}, \sigma^2 I). \tag{3}$$

Although there are some obvious shortcomings that come with using this model (e.g., Z_s should have support bounded between $(-1,\ 1)$, and its marginal variance is likely dependent on $n_{s,i}$), its simplicity made for a useful starting point given my state of knowledge at the beginning of the project, and allowed me to sequentially incorporate complexity (Bayesian sampling techniques, augmentation of missing and censored data, the inclusion of a spatial random effect) on top of a comfortable base structure. Figure 1 shows a plot of the transformed response variable, with the districts colored white representing those that had missing or censored data in either 2017, 2022, or both years. Summary univariate statistics on the distribution of Z_s justify the Gaussian approach- while there exists a right skew to the distribution, about 98% of the values fall within 3 standard deviations of the mean, and the range of Z_s from -0.01297 to 0.09195 shows that none of the values comes close to the theoretical bounds of -1 and 1.

Six candidate predictors representing potential social drivers of the change in homeschooling were selected for the model: the percent of the enrolled students who identified as White (non-Hispanic), the proportion of classroom teachers in the district with 10 or more years of teaching experience, the median income of the residents of the district as reported by the Ohio Department of Taxation, the 4-year graduation rate for the high schools located within the district, the county-level Republican vote share in the 2020 U.S. Presidential election, and pupil density (the full-time-equivalent count of K-12 students divided by the square mileage of the school district), which is plotted in Figure 2 to give the reader a sense of where Ohio's urban centers are located. (*District Profile Report (Cupp Report)* FY Years 2017 and 2022).

The initial estimation of the parameters of interest for the model — namely, coefficients for each of the predictors as well as the error variance — was conducted by implementing a Metropolis-Hastings MCMC algorithm "from scratch" as outlined in chapters 10 and 11 of Bayesian Data Analysis (Gelman et al. 2020). As alluded to above, districts with any missing or censored values of the response variable were left out of the analysis, and no spatial dependence on observations was assumed. A multivariate Normal prior for the predictor coefficients and a half-Normal prior for the error variance were chosen, and a Markov Chain was set up with each step consisting of separate draws (each conditioned on the current values of the other parameter) from the two proposal distributions:

$$\boldsymbol{\beta} \sim \mathbf{MVN}(\boldsymbol{\beta}^{(k)}, \lambda_\beta \boldsymbol{I}) \tag{4}$$

$$\boldsymbol{\sigma}^2 \sim \mathbf{N}(\boldsymbol{\sigma}^{2,(k)}, \lambda_{\sigma^2}) \tag{5}$$

These proposal values derived from "alternating conditional sampling" (known as the Gibbs sampler) were then independently considered as possible draws from their respective full conditional posterior distributions. Each was accepted or rejected based on an acceptance probability determined by the ratio of the full likelihood of the proposal value given the observed data to the full likelihood of the current value given the observed data. The Markov chain was run until stationarity was suggested by trace plots of the parameters, and the proposal tuning parameters λ_β, λ_{σ^2} were adjusted until acceptance rates of the Gibbs sampler were in the window of 15 to 30 percent. Estimates resulting from this procedure were comparable to those obtained from OLS values obtained by running the

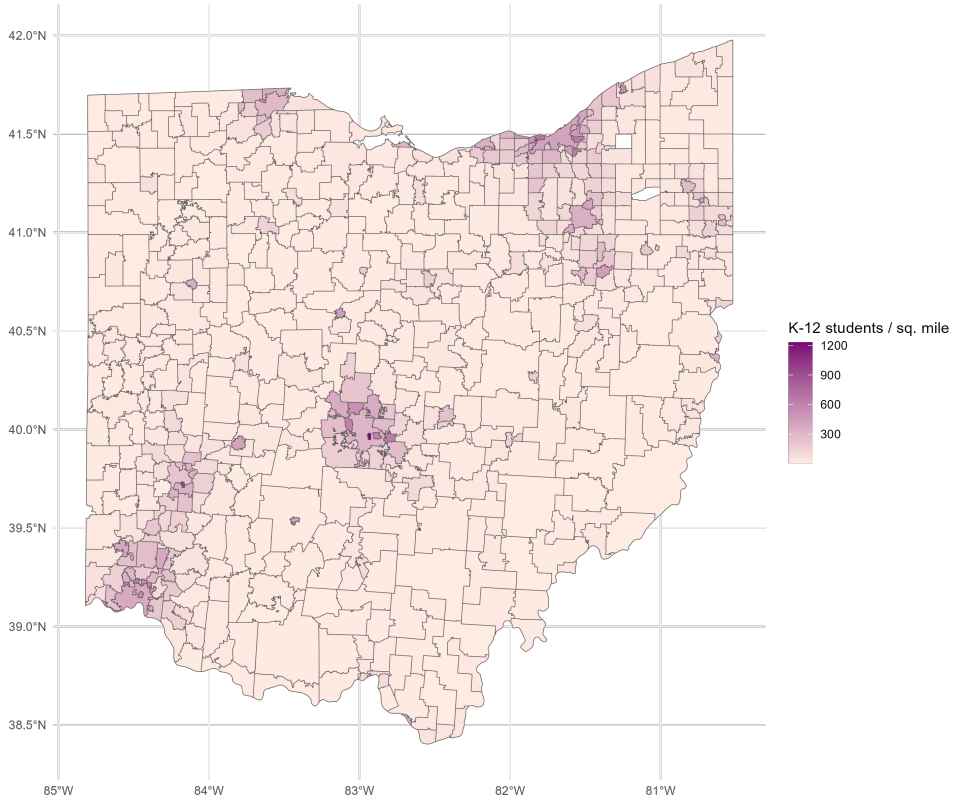

Figure 2: District Pupil Density, 2022

lm() function on the same reduced data frame. While a measure of global spatial dependence in the residuals from such a model could not be reliably calculated due to the missing data, a visual inspection of the plotted residuals suggested strong evidence of spatial dependence. This motivated the need to incorporate the rest of the districts into the linear regression model using a process of stochastic data augmentation as outlined in chapter 8 of the Bayesian Data Analysis text (Gelman et al. 2020), as well as a spatial random effect to account for spatial autocorrelation and allow for improved inference of model parameters (VerHoef et al. 2018).

MCMC Implementation with Data Augmentation and Spatial Random Effect: In addition to the aforementioned Ver Hoef article, the guiding principles for what follows were taken from *The Calculation of Posterior Distributions by Data Augmentation* (Tanner & Wong 1987) and *Gaussian Markov Random Fields: Theory and Applications* (Rue & Held 2005), with the latter being particularly helpful regarding the exploitation of the sparsity of the precision matrix to improve computational efficiency.

Let z_s denote the change in proportion of home schooled students (from 2017–2022) in school district s as described in (1). The complete response vector, $\mathbf{Z} = (Z_1, \ldots, Z_n)$, can be decomposed into three components, $\mathbf{Z} = (\mathbf{Z}_o, \mathbf{Z}_c, \mathbf{Z}_m)$, where

$$\mathbf{Z}_o = \text{ observed data,}$$

$$\mathbf{Z}_c = \text{ censored data,}$$

$$\mathbf{Z}_m = \text{ missing (and uncensored) data.}$$

Furthermore, let $\mathcal{I}^c$ be an indicator vector indicating which locations are censored, and let $\mathcal{I}^m$ be an

indicator vector denoting the locations with missing (and uncensored) data. For our linear mixed model using a CAR structure, we specify the following:

$$\mathbf{Z}|\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2) \sim \mathbf{MVN}\big(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}, \sigma^2 \boldsymbol{I}\big). \tag{6}$$

The $\boldsymbol{\gamma}$ term denotes a zero-mean spatial random effect on the areal units governed by a covariance matrix $\boldsymbol{\Sigma}_\gamma$ parameterized by spatial dependence parameter ρ and random effect marginal variance parameter τ^2, while σ^2 accounts for measurement error. In the context of the present study, this $\boldsymbol{\gamma}$ term can be viewed as an attempt to account for one or more unmeasured spatially varying covariates that are relevant to the change in homeschool rates across Ohio.

For the CAR setting, the covariance matrix $\boldsymbol{\Sigma}_\gamma$ for the random effect will be defined in terms of a neighbor-weights matrix $\mathbf{W}$ via

$$\boldsymbol{\Sigma}_\gamma(\rho, \tau^2) = \tau^2(diag(\mathbf{W}1) - \rho\mathbf{W})^{-1}. \tag{7}$$

To build $\mathbf{W}$, a first-order "queen" definition of neighbors was initially used, in which any school districts that shared at least a boundary point were considered neighbors. Note that the canonical representation of a CAR model — in which the conditional mean of each z_s is written as a weighted average of its neighbors — does not guarantee a full joint distribution; however, we will assume that the proper limitations on $\mathbf{W}$ and ρ, as given in Ver Hoef et al. (p. 42), have been satisfied, and proceed with the jointly Gaussian representation of our response. Furthermore, we also consider a subclass of the CAR model which is known as an intrinsic conditional autoregressive (ICAR) model. In ICAR, the spatial dependence parameter is not estimated, but is rather fixed in advance at $\rho = 1$, and the row-standardization of $\mathbf{W}$ as described above is required, not optional as it is in the case of the more general CAR model. While this leads to an improper distribution as just described, if a sum-to-zero constraint is imposed on the spatial effect parameters, the distribution becomes proper. The appeal of the ICAR approach is identified by Ver Hoef et al. (pg. 43) as such: despite the reduction in the number of parameters to estimate in the model (by fixing ρ), there still exists sufficient flexibility in the surface for fitting the data, and the sum-to-zero constraint keeps the spatial random error estimates under control and anchored near zero (VerHoef et al. 2018).

The conditional structure of our model implies a simple data augmentation algorithm for this setting; we can obtain samples from $f(\boldsymbol{z}^c, \boldsymbol{z}^m|\boldsymbol{z}^o, \boldsymbol{\theta})$ and $f(\boldsymbol{\theta}|\boldsymbol{z})$ using the following steps.

- Start with initial values, $\{\boldsymbol{z}_c^{(0)}, \boldsymbol{z}_m^{(0)}, \boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)}, \rho^{(0)}, \tau^{2,(0)}\sigma^{2,(0)})\}$.
- For k in $1 : N$, obtain samples:
  1. $\boldsymbol{z}_c^{(k)} \sim f\big(\boldsymbol{z}_c|\boldsymbol{z}_o, \mathcal{I}^c, \boldsymbol{z}_m^{(k-1)}, \boldsymbol{\theta}^{(k-1)}\big)$
  2. $\boldsymbol{z}_m^{(k)} \sim f\big(\boldsymbol{z}_m|\boldsymbol{z}_o, \mathcal{I}^m, \boldsymbol{z}_c^{(k)}, \boldsymbol{\theta}^{(k-1)}\big)$
  3. $\boldsymbol{\theta}^{(k)} \sim f\big(\boldsymbol{\theta}|\boldsymbol{z}^{(k)}\big)$, where $\boldsymbol{z}^{(k)} = (\boldsymbol{z}_o^{(k)}, \boldsymbol{z}_c^{(k)}, \boldsymbol{z}_m^{(k)})$.

Together, the first two of these form what Tanner and Wong refer to as the *imputation step*, while the last is referred to as the *posterior step*. We begin with a description of the imputation step:

Imputation Step For notational convenience, let $\boldsymbol{z} = (\boldsymbol{z}_o, \boldsymbol{z}_m, \boldsymbol{z}_c)$, $\boldsymbol{z}_{om} = (\boldsymbol{z}_o, \boldsymbol{z}_m)$, $\boldsymbol{z}_{oc} = (\boldsymbol{z}_o, \boldsymbol{z}_c)$, and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \gamma, \sigma^2)$. Our linear mixed model with CAR specification as given in (6) implies

that we can split out the conditional distribution of $\boldsymbol{z}$ into sub-vectors of the missing data and the rest of the observations, yielding:

$$\mathbf{Z}|\boldsymbol{\theta} = \begin{pmatrix} \mathbf{Z}_m \\ \mathbf{Z}_{oc} \end{pmatrix} |\boldsymbol{\theta} \sim \mathbf{MVN}\big(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}, \sigma^2 \boldsymbol{I}\big). \tag{8}$$

From well-known properties of multivariate Gaussian distributions, this implies that the conditional distribution $\mathbf{Z}_m|\mathbf{Z}_{oc}, \boldsymbol{\theta}$ can be written as a product of its conditionally independent marginal distributions:

$$\mathbf{Z}_m|\mathbf{Z}_{oc}, \boldsymbol{\theta} = \prod_{\{s:\mathcal{I}_s^m=1\}} \mathbf{N}\big(x_s^T\boldsymbol{\beta} + \gamma_s, \sigma^2\big) \tag{9}$$

Therefore, at each step of the Markov Chain, imputing values for missing responses $z_s^{(k)}, s \in \mathcal{I}^m$ was achieved via sampling from a Normal distribution with mean $x_s^T\boldsymbol{\beta}^{(k)} + \gamma_s^{(k)}$ and variance $\sigma^{2,(k)}$. With respect to the censored data, we get a result that is only slightly more complicated. Again using our conditional specification, we write

$$\mathbf{Z}_c|\mathbf{Z}_{om}, \boldsymbol{\theta} = \prod_{\{s:\mathcal{I}_s^c=1\}} \mathbf{N}\big(x_s^T\boldsymbol{\beta} + \gamma_s, \sigma^2\big)\mathbb{1}_{(z_s \in (l_s, u_s))} \tag{10}$$

Thus, sampling values $\boldsymbol{z}_c^{(k)}$ for the censored data at each iteration amounts to sampling from a truncated multivariate normal distribution. In the context of this analysis, since single-digit values of the number of home schoolers were the censored values, the lower bound l_s and upper bound u_s for z_s where $s \in \mathcal{I}^c$ were derived by substituting the value 9 or 0 into the appropriate places for $Y_{s,1}$ or $Y_{s,2}$ in (1).

Posterior Step With tentative imputed values for our censored and missing response values, samples were generated from $f\big(\boldsymbol{\theta}|\boldsymbol{z}\big)$ in the following manner:

- Direct posterior sample of $\boldsymbol{\beta}$
- Probabilistic Gibbs sample of σ^2
- Probabilistic block Metropolis sample of the spatial random effect parameters: τ^2, ρ, and (conditioned on these first two) $\boldsymbol{\gamma}$

The prior distributions that were ultimately settled on for use in the Bayesian likelihood calculations were as follows:

$$\boldsymbol{\beta} \sim \mathbf{N}(0, \lambda_\beta \boldsymbol{I}), \tag{11}$$

$$\rho \sim \mathbf{Beta}(18, 2), \tag{12}$$

while σ^2 and τ^2 were each assigned half-normal priors corresponding to zero-mean normals with respective variances λ_{σ^2} and λ_{τ^2}. To see why the $\boldsymbol{\beta}$ vector of coefficients can be directly sampled from in this formulation, observe that

$$f(\boldsymbol{\beta}|\boldsymbol{Z},\sigma^2,\boldsymbol{\gamma}) \propto f(\boldsymbol{Z}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\gamma}) * \pi(\boldsymbol{\beta}) \tag{13}$$

$$\propto exp\big(\frac{-1}{2\sigma^2}(\boldsymbol{Z}-\boldsymbol{\gamma}-\mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{Z}-\boldsymbol{\gamma}-\mathbf{X}\boldsymbol{\beta})\big) * exp\big(\frac{-\boldsymbol{\beta}^T\boldsymbol{\beta}}{2\lambda_\beta}\big) \tag{14}$$

$$\propto exp\big(\frac{-1}{2}\big[\boldsymbol{\beta}^T\big(\frac{\mathbf{X}^T\mathbf{X}}{\sigma^2}+\frac{1}{\lambda_\beta}\big)\boldsymbol{\beta}-\frac{2(\boldsymbol{Z}-\boldsymbol{\gamma})^T\mathbf{X}\boldsymbol{\beta}}{\sigma^2}\big]\big) \tag{15}$$

This represents a Multivariate Gaussian distribution with mean and variance (each of which are conditioned on the data and the current values of the parameters) given by:

$$\mu_\beta = \big(\mathbf{X}^T\mathbf{X}+\frac{\sigma^2}{\lambda_\beta}\boldsymbol{I}\big)^{-1}\mathbf{X}^T\big(\boldsymbol{Z}-\boldsymbol{\gamma}\big) \tag{16}$$

$$Var_\beta = \sigma^2\big(\mathbf{X}^T\mathbf{X}+\frac{\sigma^2}{\lambda_\beta}\boldsymbol{I}\big)^{-1} \tag{17}$$

While this particular Gaussian would have been easy enough to sample directly from using existing R packages designed for multivariate normal computations, I was interested in learning how to exploit sparseness properties in the covariance matrix to speed up computations. Below, we will see that this effort pays dividends when the covariance matrix is derived from our sparse neighborhood matrix $\mathbf{W}$. Rue and Held (2005) point out that "what makes Gaussian Markov Random Fields (GMRFs) extremely useful in practice is that the things we often need to compute are particularly fast to compute for a GMRF. The key is naturally the sparseness of the precision matrix and the structure of its nonzero terms" (pg. 26). Following the authors' notation of using $\mathbf{Q}$ to denote the precision matrix (which is simply the inverse of the covariance matrix, in this case for $\boldsymbol{\beta}$), we will write

$$\mathbf{Q}_\beta = \frac{1}{\sigma^2}\big(\mathbf{X}^T\mathbf{X}+\frac{\sigma^2}{\lambda_\beta}\boldsymbol{I}\big) \tag{18}$$

While $\mathbf{Q}_\beta$ is not sparse in this example, I was still able to implement an algorithm (pg. 46) that uses the Cholesky decomposition of this precision matrix, along with a sequential series of systems of equations with easy matrix solutions, to achieve the sampling for $\boldsymbol{\beta}$ at each iteration of the MCMC (Rue & Held 2005).

The Gibbs sampler for the parameter σ^2 was borrowed from the simpler model described in an earlier section, using the proposal distribution (5), as σ^2 is independent of the other parameters in the model and this process needed no modification. With respect to the parameters associated with the spatial random effect, though, a Gibbs sampler (cycling through each of the three parameters in turn, drawing a proposal for each one conditional on the value of all the others, then deciding whether to individually accept or reject it based on a likelihood ratio calculation) was discarded due to the fact of the probable high correlation between the values of τ^2, ρ, and γ. In other words, at each step of the MCMC algorithm, if we were to independently accept a proposal value for one of these parameters, the other parameters would tend to "follow it around" in the parameter space. To more efficiently arrive at the target distribution for these three dependent parameters, a block update algorithm was used, in which proposal values $\tau^{2,(*)}, \rho^*$, and γ^* were generated in sequence, and then the set of proposals was accepted or rejected en masse as a representation from the posterior distribution. The outline of the process is as follows:

1. Propose $\tau^{2,*}$ from $\mathbf{N}(\tau^{2,(k)}, \lambda_{\tau^2})$,

2. Propose ρ^* from a truncated Normal with mean $\rho^{(k)}$, variance λ_ρ, and restriction $\rho \in (0, 1)$,

3. Sample γ^* from a multivariate Gaussian (derived using calculations similar to that used for the distribution of $\boldsymbol{\beta}$ above) with parameters

$$\mu = \left(\boldsymbol{\Sigma}_\gamma(\rho^*, \tau^{2,(*)})^{-1} + \frac{1}{\sigma^2}\boldsymbol{I}\right)^{-1} * \frac{\left(\boldsymbol{Z} - \mathbf{X}\boldsymbol{\beta}\right)}{\sigma^2} \tag{19}$$

$$Var = \left(\boldsymbol{\Sigma}_\gamma(\rho^*, \tau^{2,(*)})^{-1} + \frac{1}{\sigma^2}\boldsymbol{I}\right)^{-1} \tag{20}$$

   (Note the dependence of the parameters of this proposal distribution on the just-sampled values of $\tau^{2,(*)}$ and ρ^*), then

4. Accept $(\tau^{2,(*)}, \rho^*, \gamma^*)$ as the posterior draw $(\tau^{2,(k+1)}, \rho^{(k+1)}, \gamma^{(k+1)})$ with a probability ratio of the conditional posterior of τ^2, marginalized over γ, evaluated at $\tau^{2,(*)}$, to the that same posterior evaluated at $\tau^{2,(k)}$. Due to the asymmetric nature of the proposal distribution for ρ, another balancing factor is included in this ratio as well.

With this approach, the CAR specification provides dividends in the computation of (20). While a data set of approximately 600 observations is not terribly huge in the grand scheme of things, the need for multiple large matrix inversions at every step of a chain with 10^5 iterations can be cumbersome (to say the least) if the sparsity inherent in $\boldsymbol{\Sigma}_\gamma(\rho^*, \tau^{2,(*)})$ is not taken advantage of. Since each district in Ohio is the neighbor of (or has adjacency to) only a handful of other districts at most, the precision matrix $\mathbf{Q}_\gamma$ inherits a sparsity property from the neighbor weights matrix $\mathbf{W}$. Computational advantages arise from the fact that our covariance matrix $\boldsymbol{\Sigma}_\gamma$ is defined as the inverse of this sparse precision matrix. So, following the same essential algorithm as used above to sample β, taking posterior draws for the spatial random effect was able to be accomplished in a reasonable time frame.

4 Results/Model Comparison

Three model structures were selected for comparison: a linear (LIN) model with no spatial random effect, a conditional autoregressive (CAR) model, and an intrinsic conditional autoregressive (ICAR) model. Each respective model was run for 50,000 MCMC iterations, and, based on evidence of stationarity seen in trace plots (see example below), the first 10,000 runs in each chain were discarded as burn-in, leaving a posterior sample of size 40,000 for each model. Convergence was confirmed by calculating the Gelman-Rubin test statistic on a subset of the chains produced by each model. The Gelman-Rubin test statistic is a simple (unit-free) ratio that compares variance between chains to that within chains, so that values very close to one are desirable as an indication of a stationary posterior distribution. The chains fed into the test statistic can be chains estimating the same parameter from different starting values, or sub-chains from an arbitrarily "chopped up" longer chain from a single MCMC run. Using the latter option, Gelman-Rubin test statistics sufficiently close to the desired value of one were observed (all were less than 1.001).

Of the many options for model selection criteria available in the literature, Bayesian Information Criterion (BIC) was precluded since the parameter counts in its penalty term are meaningless in the

Bayesian hierarchical settings, Deviance Information Criterion (DIC) could not be used due to the presence of missing data, and the Watanabe-Akaike Information Criterion (WAIC) was inappropriate because its data independence assumption was clearly violated in this setting. Instead, Hobbs and Hooten recommend using the lesser-known posterior predictive loss function for model selection when comparing spatial hierarchical models (Hobbs & Hooten 2015). The idea of the posterior predictive loss function is to quantify the fit of the model by comparing features of the posterior predictive distribution $p(\boldsymbol{Z}^{new}|\boldsymbol{Z})$ to equivalent features of the observed data using a familiar format of a penalty term plus a goodness-of-fit term. Another appealing attribute of this criterion is that it is able to account for an evaluation of the estimates associated with censored data (although missing data had to be ignored). In the current setting, the specific calculation that was carried out was

$$PPL = \sum_{i=1}^{n} Var(\boldsymbol{Z}_i^{new}|\boldsymbol{Z}) + \sum_{i=1}^{n} (\boldsymbol{Z}_i^{new} - \boldsymbol{E}(\boldsymbol{Z}_i^{new}|\boldsymbol{Z}))^2, \tag{21}$$

where, for censored data, the goodness-of-fit term was recorded as zero for predictions that fell within the range of plausible response values, and was calculated as the squared difference from the nearest endpoint of the range of plausible responses otherwise (Gelfand & Ghosh 1998). Table 1 displays the results: the CAR model was found to have the smallest posterior predictive loss, indicating a superior fit compared to the other two models. Effective sample sizes for the components of the CAR MCMC chain were under 1,000 for the variance parameters τ^2 and σ^2, but were over 20,000 for all other parameters of interest.

Model	Posterior Predictive Loss
Linear Model, No Spatial Random Effect	0.11757
CAR Model	**0.11160**
ICAR Model	0.11540

Table 1: Diagnostic Performance Comparison of the Three Models

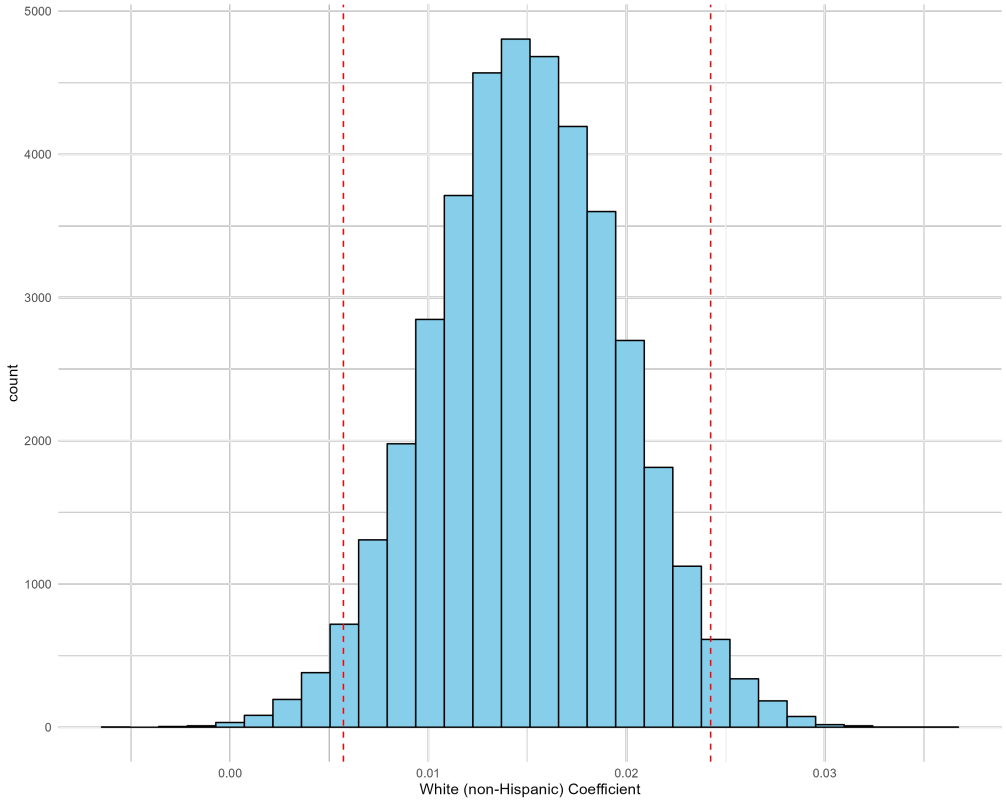

Figure 3: Posterior distribution example. Credible interval endpoints are indicated by the dashed lines.

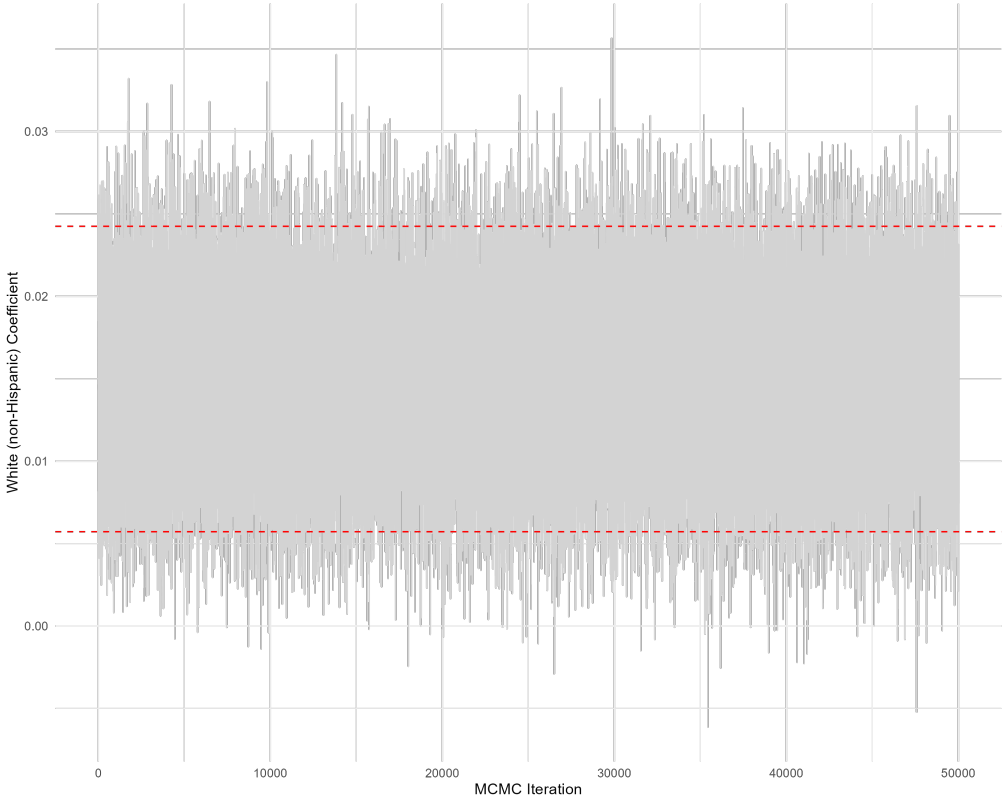

Figure 4: Trace plot example. Credible interval endpoints are indicated by the dashed lines.

While the CAR model appeared to be the preferred model by comparing these diagnostics, it

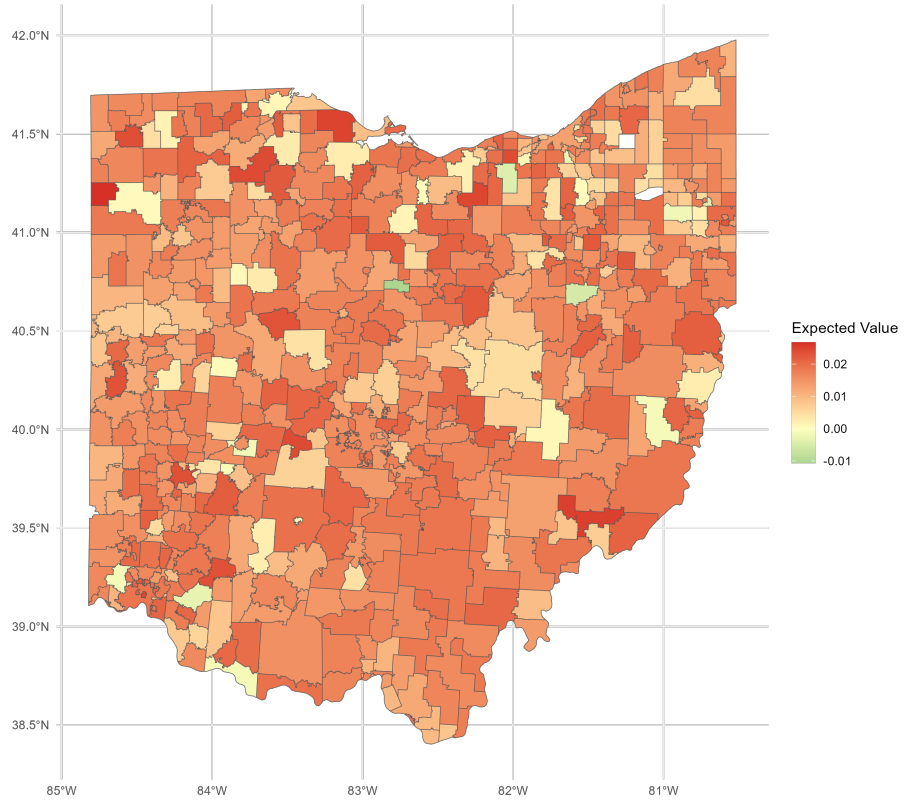

Figure 5: Predicted Difference in Annual Proportions in Home Schooling in All Districts According to CAR Model

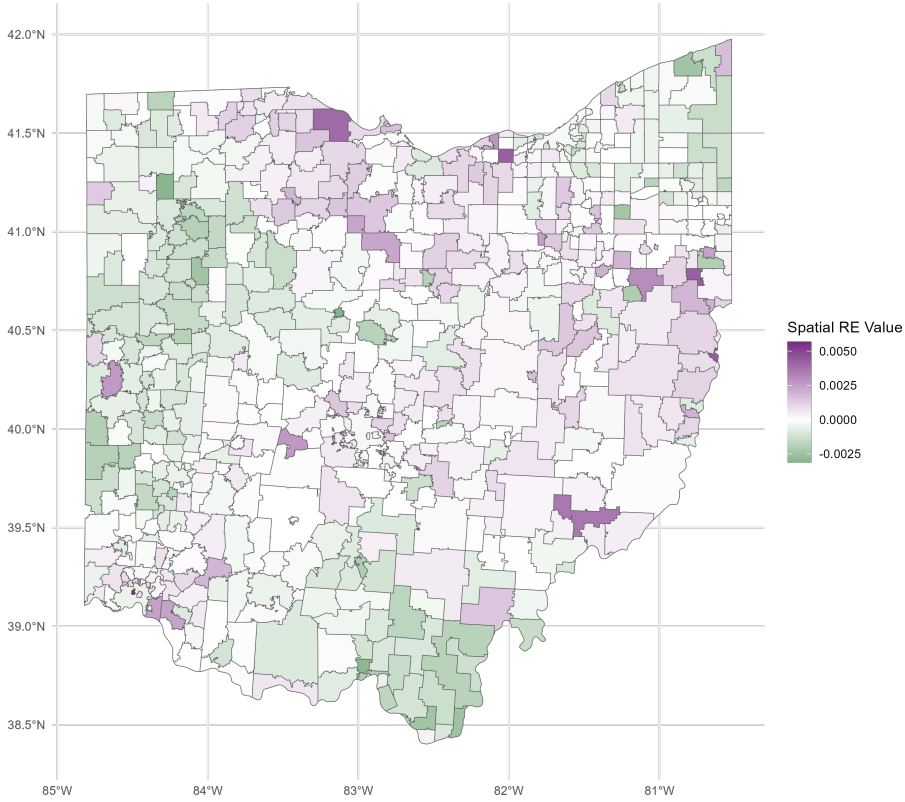

Figure 6: Posterior Mean Spatial Random Effect Surface, CAR Model

did lead to one concerning piece of output: when the distribution of random effect estimates (as

measured by the mean value of random effect estimates in each district over all iterations of the MCMC) was examined, one district appeared as an extreme outlier: St. Bernard-Elmwood Public Schools in southwestern Ohio. After considering possible reasons for this anomaly, it was determined that it most likely was a result of the district being completely surrounded on all sides by just one other district: Cincinnati Public Schools. In fact, it was one of only 11 districts in the state that had a single neighbor according to the queen's contiguity definition initially employed. It was decided that a more expansive definition of neighbors should be employed to see what effect this had on the analysis, so the definition of $\mathbf{W}$ was altered to also give positive weight to second-order neighbors ("neighbors of neighbors") for each district. This change had little to no effect on main parameter estimates and brought the values of the spatial random effect estimates closer to zero, including greatly reducing the magnitude of the St.Bernard-Elmwood district's estimate and bringing it in line with values from other districts. Therefore, although this change came at a computational cost (due to $\mathbf{W}$ no longer being as sparse as it was in the first-order neighbors-only case), the CAR model with second-order neighborhood structure was chosen as the preferred model, and the results that follow correspond to this specification.

Predictor	95% Bayesian credible interval	Posterior mean
(Intercept)	[0.00071, 0.042]	0.021
Pupil Density *	[-0.000028, -0.0000042]	-0.000016
Median Income *	[-0.00000036, -0.00000001]	-0.00000019
Proportion White (non-Hispanic) *	[0.0057, 0.024]	0.015
Proportion of Experienced Teachers *	[-0.026, -0.0035]	-0.015
County Republican Vote Share	[-0.0083, 0.018]	0.0048
Four-year Graduation Rate	[-0.026, 0.018]	-0.0043

Table 2: Summary of Bayesian estimates for all included predictors.
Statistically significant predictors at the .05 level indicated with an asterisk.

Table 2 shows the results from the CAR regression model; four of the six predictors examined were found to have a significant association with an increase in home schooling in Ohio school districts between the years 2017 and 2022. Converting to more interpretable units, we see the following: an increase of 100 students per square mile in the density of a district is associated with a drop in home schooling of about .16 percentage points; an increase of 10,000 dollars in the median income of a school district is associated with a decrease in home schooling of .19 percentage points; an increase of 10 percentage points in the White composition of a district's student body is associated with an increase in home schooling of .15 percentage points, and an increase of 10 percentage points in the composition of a district's teaching staff with ten or more years of experience is associated with an decrease in home schooling of .15 percentage points. Of note, an Ordinary Least Squares analysis of the same predictors on the incomplete data set (that is, ignoring any districts whose home school count for either 2017 or 2022 was missing or censored) reached the same conclusion in terms of which of the six predictors reached statistical significance; however, the magnitude of the coefficients on those predictors were quite different from those found here.

A Moran's I test for spatial dependence was used to confirm the evidence from the graphic above; namely, that the CAR specification accurately captured most of the spatial dependency in the residuals from the original model. The Moran's test statistic for the spatial random effect was equal to 14.31 and was highly significant (p-value approximately zero), while the test statistic for the residuals was equal to -0.76 and had a p-value of 0.78, indicating that there is no evidence of

remaining autocorrelation that is unaccounted for by the CAR model.

5 Discussion

While *The Washington Post*'s claim that, "Home schooling's surging popularity crosses every measurable line of politics, geography and demographics," is supported by Figure 1, this analysis suggests that the surge is greater among some factions of the population than others, at least in the case of Ohio. Of the demographic, economic, political, and educational factors examined here, it was discovered that school districts with more White residents, lower median income, lower population density, and a lower proportion of teachers with 10 or more years of experience were associated with a greater increase in home schooling than other districts, on average. Further research into this topic could try to ascertain if these patterns hold in other states and over more extensive time frames.

Besides its limited scope in time and geography, some other shortcomings in the model warrant skepticism about reading too much into the associations discovered here or extrapolating them to other settings. As a reminder, two of the variables incorporated into the model (private school counts in the response variable transformation and political affiliation as a predictor) present plausible change-of-support issues via the imputation of county-level data to smaller areal units. As with any regression analysis, confounding is also a potential issue. The spatial random effect built into the model serves as something of an insurance policy against the oversight of important covariates, but there still exists the possibility of biased estimators due to incomplete model specification. Follow-up research could map other potential drivers of homeschool increases against the spatial random effect to ascertain if there is another variable that should be included in the model as a plausible covariate. Another potential path for future research is to alter the model so that the original count data of number of homeschooled students is the response variable, as it is possible that some bias was introduced into the model through the estimation process of constructing Z_s. If such a Poisson formulation were to be attempted, it would be important to account for the large variation in district enrollments through an appropriate offset factor. Robustness of the model was not tested against alternative formulations of the spatial weights matrix- a definition based on K-nearest neighbors or intercentroidal distances could alter the inference. Finally, since model selection is a tricky matter when it comes to Bayesian models that allow for missing and censored data, it would be valuable to confirm the relative superiority of the CAR model using a technique other than a Posterior Predictive Loss function.

References

District Profile Report (Cupp Report) (FY Years 2017 and 2022).
URL: *https://data.ohio.gov/wps/portal/gov/data/view/ode-cupp-report*

Gelfand, A. E. & Ghosh, S. K. (1998), 'Model choice: A minimum posterior predictive loss approach', *Biometrika* **85**(1), 1–11.

Gelman, A., Carlin, J., Stern, H., Rubin, D., Dunson, D. & Vehtari, A. (2020), *Bayesian Data Analysis*, 3rd edn, Chapman and Hall/CRC.

Hobbs, N. T. & Hooten, M. B. (2015), *Bayesian Models: A Statistical Primer for Ecologists*, Princeton University Press.

Hoyer, M. (2023), 'The rise of home schooling: Data from the post's analysis of home-schooling enrollment across the u.s.'.
URL: *https://github.com/washingtonpost/data$_h$ome$_s$chooling*

Jamison, P., Meckler, L., Gordy, P., Morse, C. E. & Alcantara, C. (2023), 'Home schooling's rise from fringe to fastest-growing form of education', *The Washington Post* .
URL: *https://www.washingtonpost.com/education/interactive/2023/homeschooling-growth-data-by-district/*

Ohio Department of Education and Workforce Report Portal (n.d.).
URL: *https://reports.education.ohio.gov/overview*

Rue, H. & Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Chapman Hall/CRC.

Tanner, M. A. & Wong, W. H. (1987), 'The calculation of posterior distributions by data augmentation', *Journal of the American Statistical Association* **82**(398), 528–540.

VerHoef, J. M., Petersen, E. E., Hooten, M. B., Hanks, E. M. & Fortin, M.-J. (2018), 'Spatial autoregressive models for statistical inference from ecological data', *Ecological Monographs* **88**(1), 36–59.