Predicting the Number of Students, grades K-5, in NYC at the Census Tract Level

Ben Jakubowski and Michael Higgins

***Abstract*—In 2016, the NYC Department of Education released a Call for Innovations requesting submission of predictive models that gave MSE-minimizing predictions of the number of students in each grade K-5, for each census tract in School District 20, for each year from 2011-12 to 2015-2016. We report on development of two types of probabilistic graphical models developed for this predictive task. Both models are premised on the hypothesis that priors which induce spatial smoothness could help reduce overfitting and yield better predictive models. Unfortunately, this core hypothesis was not supported by our experiments. Regardless, we report our results, provide potential explanations, and suggest alternative approaches.**

***Keywords*—Areal spatiotemporal data, Gaussian Markov Random Fields, Stan**

I. Introduction

In the fall of 2016, the New York City (NYC) Department of Education (DoE) released a Call for Innovations, or public call for submissions addressing a pressing civic need. The DoE specifically was soliciting proposals for "Enhancing School Zoning Efforts by Predicting Population Change." Since the DoE serves over 1.1 million students, they were interested in more accurately projecting population changes in order to inform school zoning and resource allocation decisions.

Their Call for Innovations was divided into multiple stages; we developed a submission for the initial modeling stage as our project for the NYU class Inference and Representation. Specifically, in this initial modelling stage of the challenge, teams were asked to develop predictive models for the following predictive task:

- For each census tract k in NYC School District 20 (142 tracts):
  - For each grade g in K-5:
    - Using data from the school years 2001-2002 to 2010-2011, for each year t from 2010-2011 to 2014-2015:
      - **Target**: Predict the total number of students in grade g in tract k enrolled in public or charter schools in year t.

An accurate predictive model for this task would help the DoE better plan for and serve the evolving needs of NYC families. Unfortunately, this problem is very difficult because it requires making a large number of predictions given a small amount of training data. For each tract, only 10 years of counts are available for each grade, and these 10 years of data are being used to support 5 years of projections.

Given this problem structure (making a large number of predictions using a very small data set), we decided to test models that allow for sharing of information across tracts. Specifically, we hypothesized that gains could be made by exploiting the spatial structure of the data: if counts for adjacent census tracts are similar, then models that capture neighborhood structure would be more robust than independently modeling each tract.

In addition to exploring models that capture this hypothesized spatial structure, we also chose to constrain our models by not consuming side data. This choice reflected the following deployment constraints and hypotheses:

- **Constraint**: Since the model is intended to project counts for five years, we would need to restrict the allowed feature set to features available prior to the prediction window.
- **Hypothesis** Additionally, in the second modeling stage for this contest, predictions are supposed to be made at the city block level. We assumed the following: given the numbers of students in grades K-5 for 10 past years, no readily available city-block level features (available prior to the five year prediction window) would provide additional information regarding the targets.

Based on these hypotheses and constraints, we tested the performance of two types of models:

1) Spatially correlated linear models.
2) Spatially correlated markov chain models.

Results and analysis for each of these models is presented in the subsequent sections.

II. Related Work

Our models are all inspired by the areal spatiotemporal models presented by Lee, Rushworth, and Napier in the vignette for their R package `CARBayesST` [1]. In this package[1], Lee *et al* developed Gibbs samplers for a number of spatiotemporal mixed effects models, including a temporally linear, spatially correlated model. Letting t be time, temporal linearity is apparent from the first component of the model specification (which, using the semantics of the applied statistics community, is simply a hierarchical, or random effects, model for a single

[1] As a side note, in the course of this project the authors identified a bug in their implementation of the log likelihood extractor, and were able to contribute to this package by reporting it to Lee.

district):

Target in spatial unit k at time t :

$$Y_{kt} \sim \mathrm{N}(x_{kt}^T \beta + \phi_{kt}, \nu^2)$$

Latent spatiotemporal variable:

$$\phi_{kt} = \underbrace{\beta_1}_{\substack{\text{District} \\ \text{intercept}}} + \overbrace{\phi_k}^{\substack{\text{Tract} \\ \text{intercept}}} + (\underbrace{\alpha}_{\substack{\text{District} \\ \text{slope}}} + \overbrace{\delta_k}^{\substack{\text{Tract} \\ \text{slope}}}) \frac{(t - \bar{t})}{N}$$

While the first component of the model specification implies temporal linearity, it does not impose spatial smoothness on the random effects. This is achieved through the priors placed on ϕ_k and δ_k. Described by Lee *et al* as conditional autoregressive (CAR) priors, they are simply Gaussian Markov Random Fields (GMRFs) (as shown in Fig. 1) where the graph structure maps onto the spatial units' neighborhood structure.

[1]Gaussian Markov Random Field (GMRF) captures spatial correlation between adjacent census tracts (Image: Orchard, University of Edinburg):

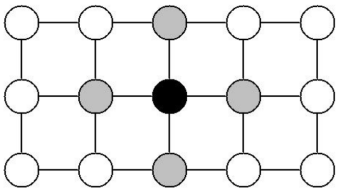

Fig. 1: General structure of a Markov Field on a lattice

The form of the GMRF prior described by Lee *et al* was first proposed by Besag *et al* [2] for use in image restoration (where pixels have a natural neighborhood structure), and then modified by Leroux *et al* [3] to allow for varying strength of the spatial relationship.

Specifically, given the adjacency matrix W for the spatial units, the conditional priors are given by:

$$\phi_k | \phi_{-k}, W \sim \mathrm{N}\left(\frac{\rho_{int} \sum_{j=1}^{K} w_{kj} \phi_j}{\rho_{int} \sum_{j=1}^{K} w_{kj} + 1 - \rho_{int}}, \frac{\tau_{int}^2}{\rho_{int} \sum_{j=1}^{K} w_{kj} + 1 - \rho_{int}} \right)$$

$$\delta_k | \delta_{-k}, W \sim \mathrm{N}\left(\frac{\rho_{slo} \sum_{j=1}^{K} w_{kj} \delta_j}{\rho_{slo} \sum_{j=1}^{K} w_{kj} + 1 - \rho_{slo}}, \frac{\tau_{slo}^2}{\rho_{slo} \sum_{j=1}^{K} w_{kj} + 1 - \rho_{slo}} \right)$$

Note these priors are parametrized by two additional parameters: ρ, which ranges from [0,1] and controls the strength of the spatial relationship, and τ^2, which controls the variance of the latent spatiotemporal effects. Finally, Lee *et al* state they use weakly informative priors on these parameters:

$$\begin{aligned} \tau_{int}^2, \tau_{slo}^2 &\sim \text{Inverse-Gamma}(a, b) \\ \rho_{int}, \rho_{slo} &\sim \text{Uniform}(0, 1) \end{aligned}$$

While this specifies the model, the structure is perhaps obfuscated by notation. Hence we present Lee *et al*'s model (ignoring hyperpriors and other regression coefficients) in figure 2.

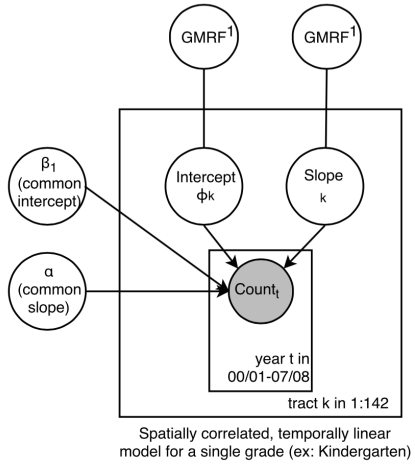

Fig. 2: Graphical model depicting key components of the temporally linear, spatially correlated regression model.

This model (and in particular the priors initially proposed by Leroux [3]) are the basis for our modeling experiments. Importantly, to check our understanding of this model, we first compared the results obtained from (i) using Lee *et al*'s packaged Gibb sampler, and (ii) a from-scratch implemention of the model in Stan, to model counts for a single grade (Kindergarten). These samplers returned similar posterior mean point estimates, providing a sanity check on our from-scratch implemenation (see Fig. 3).

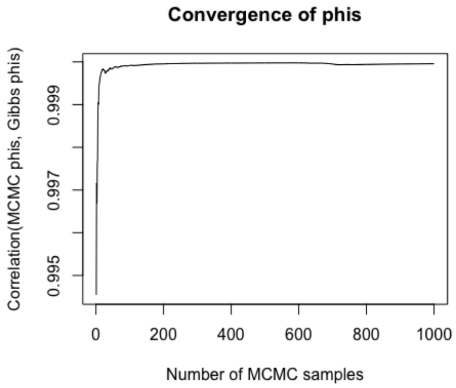

Fig. 3: Correlation between (i) the running posterior mean point estimate of the length-142 ϕ vector obtained from the Stan model, and (ii) the final posterior mean ϕ vector obtained after 60,000 samples

III. Problem Definition and Algorithm

A. Task

Again, our task is to predict the number of students in grade g, year t, and census tract k for:

- g in kindergarden through 5th grade;
- t in academic years 2011-12 to 2015-16
- k in set of 142 census tracts in NYC School District 20

The available data (for training and testing) are counts for these census tracts and grades, for years from 2001-02 to 2010-2011. Finally, the DoE will be evaluating predictions using mean squared error (MSE) as a loss function.

B. Algorithm

We approached this predictive task using the following general framework (note detailed information on the specific model families and models is provided in the methodology section):

1) We tested two different families of models.
2) Within each family, we test multiple model variants.
3) For each specific model, we learned the model (i) with GMRF priors to enforce spatial smoothness, and (ii) without GMRF priors for comparison.
4) Models were written using the Stan probabilistic programming language [4], and the model was fit using Hamiltonian Markov Chain Monte Carlo sampling.
5) Training and test set predictions were made using posterior mean point estimates for necessary parameters following sampling.
6) To evaluate our models we compared the test set MSE performance for (i) the spatially regularized model and (ii) the spatially unregularized model, to determine whether (and to what extent) spatial regularization improved performance.

IV. Experimental Evaluation

A. Data

The NYC DoE provided each challenge team with non-zero counts of students, K-5, in each census tract in District 20 in South Brooklyn. Prior to modeling, we followed the following data cleaning procedure:

1) We first applied a spatial filter to the provided dataset, only passing records for tracts that were in fact in School District 20. Unfortunately, the provided data included a number of tracts that fell outside of District 20; several weeks into the contest this was recognized by the DoE and our choice to drop these tracts was validated by the challenge sponsor.
2) Next, we filled missing values with 0's, since the metadata indicated that missing entries corresponded to an absence of students in that tract for that year.

For greater insight, histograms showing the (smoothed) distributions of counts by grade and year are shown in Fig 4, and maps showing the counts by grade and year are shown in Fig 5. Using this data, we proceeded to experiment with two general approaches to modeling, described in the following two modelling sections.

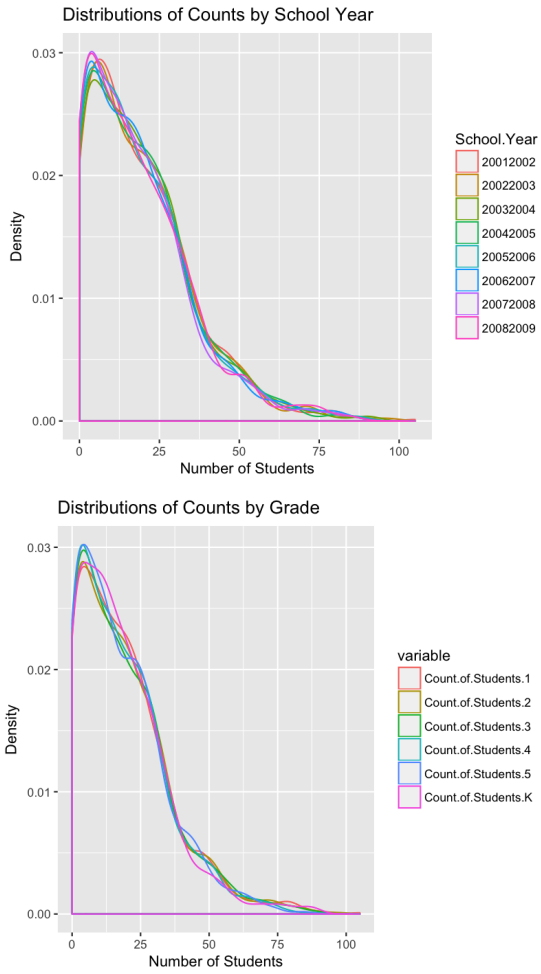

Fig. 4: Distributions of counts by year and grade

B. Model 1: Grade/Time models

Model 1 Methodology

Our first family of models were simply extensions to the spatially correlated linear models presented by Lee *et al*. These models treated grade and year (plus quadratic terms in two models) as covariates in linear models. This is an extension of Lee *et al*'s model, since it introduces additional spatially correlated coeffients (ex: β_g). This model reflects the following hypotheses:

- The number of students in a tract k in grade g during year t (Y_{kgt}) is well modeled by linear models using features constructed from grade and time.
- Moreover, the coefficients in the linear model are drawn from GMRF priors that induce spatial regularization (such that neighboring tracts have similar values for each coefficient).
- However, the priors are not shared across coefficients- there are separate priors- since the scales of the effects are hypothesized to vary (i.e. perhaps there is a stronger

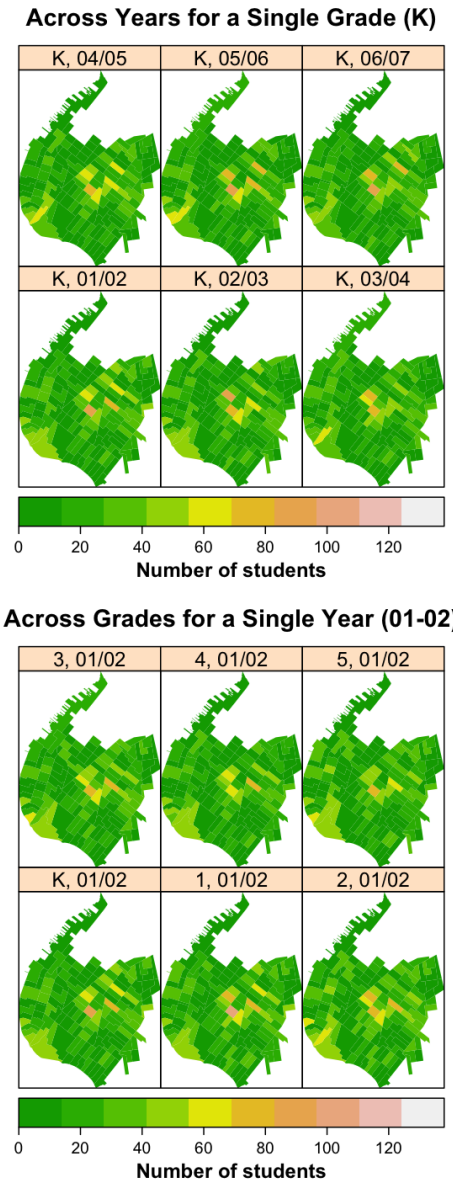

Fig. 5: Maps of counts by year and grade

temporal and a weaker grade effect).

While Fig. 6 illustrates the general structure of these models, we tested three different versions with increasing complexities:

1) Model 1a: Included $\beta_0, \beta_t, \beta_g$.
2) Model 1b (depicted in plate diagram): Included $\beta_0, \beta_t, \beta_g, \beta_{gt}$.
3) Model 1c: Included $\beta_0, \beta_t, \beta_g, \beta_{gt}, \beta_{g^2}, \beta_{t^2}$.

Models were constructed and fit using the Stan probabilistic programming language, using 500 burn-in samples, 1000 samples, and 4 chains. Finally, the posterior mean α's and β's across all chains were extracted and used as point estimates for predictive modeling on the test set.

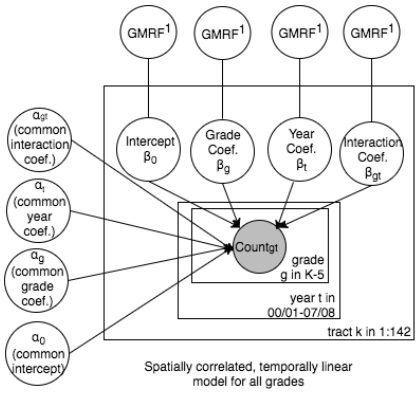

Fig. 6: Example plate diagram for Model 1: spatially correlated linear models. Note we experimented with the three different feature sets.

Model 1 Results

For this model, we used the first 8 years of the provided data as training data, and the last 2 years as test data (as shown in Fig 7).

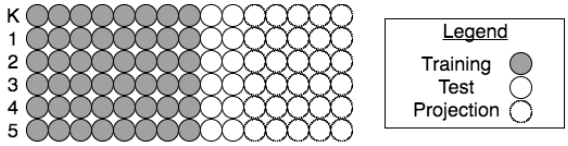

Fig. 7: Training and test sets for Model 1.

For comparison, in addition to learning the model using the CAR/GMRF priors, we also learn equivalent linear models without this spatial regularization. Results from these experiments are provided in Table 1.

Feature set	GMRF prior	Train RMSE	Test RMSE
1a: $\beta_0, \beta_t, \beta_g$	Yes (experimental)	4.25	8.23
1a: $\beta_0, \beta_t, \beta_g$	No (control)	4.22	8.30
1b: $\beta_0, \beta_t, \beta_g, \beta_{gt}$	Yes (experimental)	4.26	8.51
1b: $\beta_0, \beta_t, \beta_g, \beta_{gt}$	No (control)	4.11	8.64
1c: $\beta_0, \beta_t, \beta_g, \beta_{gt}, \beta_{t^2}, \beta_{g^2}$	Yes (experimental)	4.25	8.29
1c: $\beta_0, \beta_t, \beta_g, \beta_{gt}, \beta_{t^2}, \beta_{g^2}$	No (control)	3.98	8.55

Table 1: Experimental results for Model 1

Based on these results, it is apparent that the CAR prior had a slight regularizing effect on the linear model, especially as the potential for overfitting increased (as more terms were added to models in 1b and 1c). However, they did not significantly improve test set performance.

C. Model 2: Cohort models

Model 2 Methodology

Our second family of models are also extensions of the spatially correlated linear models presented by Lee *et al*, in

that they apply the GMRF prior to induce spatial dependence. In Model 1, we treated the grades (rows) and years (columns) as the relevant dimensions of the data matrices, and constructed linear models using (transformations) of these features. In contrast, Model 2 views the diagonals (graduation cohorts) as the relevant observation units. This model structure is based on the observation that there is less variance within cohorts than within years or grades, as shown in Fig. 8. This cohort structure is also clear from heatmaps of single census tract data matrices (an example is provided in Fig 9).

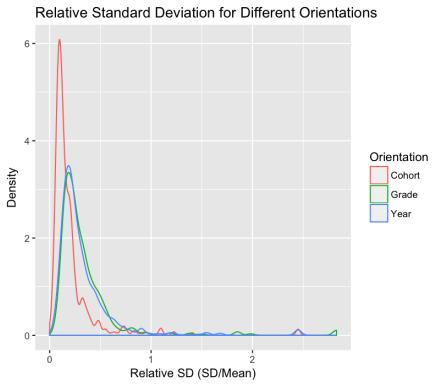

Fig. 8: Comparing the relative within grade, year, and cohort relative standard deviations

In addition to viewing cohorts as relevant unit, the model makes a number of strong assumption and operates under strong constraints:

- First, the model assumes the student counts form a Markov chain, such that given the most recent observation of cohort size, the cohort evoluation is independent of previous observations.
- In addition, the transition operator is parameterized by a single parameter r_k, which governs some simple growth process.
- Moreover, the parameters in this growth function are drawn from GMRF priors that induce spatial regularization (such that neighboring tracts have similar cohort change behavior).
- Finally, this model design suffers from the strong constraint that at least one observation must be made of the cohort size (in grade K-5). As such, it cannot be used to make predictions for the upper triangle in the "projection" section of the matrix depicted in Fig. 10. Hence, it is only useful if it allows for improved prediction in the lower triangle of this matrix and can thus augment a more general model (such as the linear models presented above).

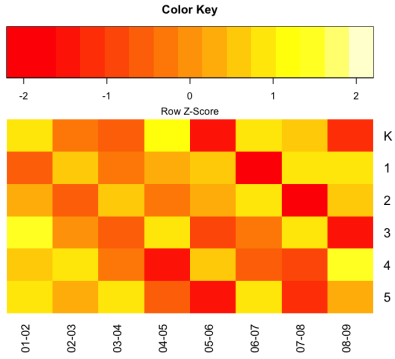

Fig. 9: Prediction surface in grade/time/count space

Fig. 10 illustrates the general structure of these model, with the counts shaded to represent fully observed training data. Given this general model structure, we tested two different versions of the model:

1) Model 2a encoded linear growth model in the tract-level growth parameter r_k, where $Y_{k(g+1)(t+1)}|Y_{kgt} \sim \mathrm{N}(Y_{kgt} + r_k, \nu^2)$.

2) Model 2b encoded exponential growth in the tract-level growth parameter r_k, where $Y_{k(g+1)(t+1)}|Y_{kgt} \sim \mathrm{N}(Y_{kgt}(1 + r_k), \nu^2)$.

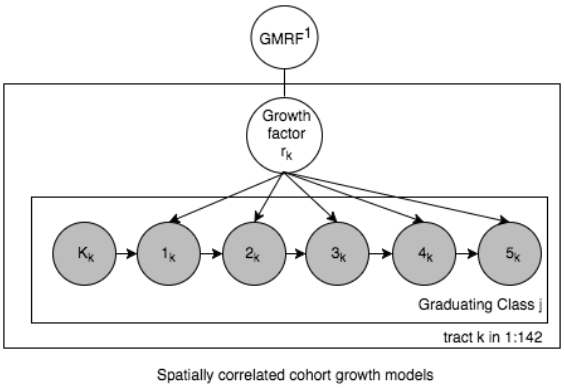

Fig. 10: Example plate diagram for Model 2: Note we experimented with two different transition operators, both parameterized by a single parameter r_k.

Models were constructed and fit using the Stan probabilistic programming language, using 1000 burn-in samples, 5000 samples, and 4 chains. Finally, the posterior mean r's across all chains were extracted and used as point estimates for predictive modeling on the test set.

Model 2 Results

For this model, we used three cohorts with complete observations as training data to learn r. We then used the remaining data for testing. As shown in Fig 11, this training/test split allow us to estimate the prediction error for every possible prediction pair (for example projecting the cohort count in grade 5 given grade 1, versus projecting the cohort count in grade 4 given grade 2).

This is useful since it allows for separate estimation of the error for each potential prediction. As shown in Fig. 11, there

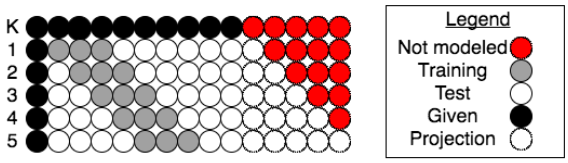

Fig. 11: Training and test sets for Model 2.

are five potential predictions made with one-year time steps, four with two-year time steps, three with three-year time steps, two with four-year time steps, and only one with five-year time steps. As such, the error on one-year time steps should be given more weight in the RMSE over the entire prediction space. Finally, for further illustration of this differentiated loss across different prediction tasks, 12 shows the case where we are predicting grades 3, 4, or 5 using grade 2 (i.e. grade 2 is the last observed grade).

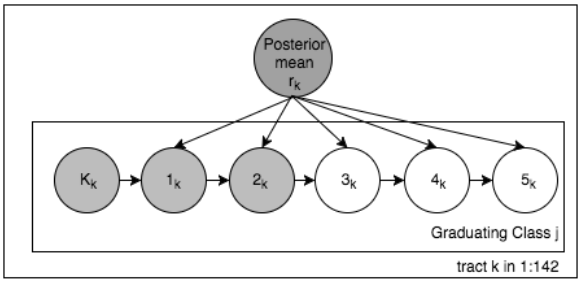

Fig. 12: Example of prediction using Model 2. Note

$$3_k, 4_k, 5_k \perp \mathcal{X} \setminus 2_k, r_k | r_k, 2_k$$

Since we are able to get separate prediction RMSE estimates $\mathrm{E}[Y_{k(g+l)(t+l)} \mid Y_{k(g)(t)}]$ for each g, l pair, we compare the models' performance across all possible pairs. Results are presented as labeled heat-maps in 2.

From these error matrices,we conclude:

- Cohort evolution models perform better than grade/time linear models (given their best test RMSE of $\sim$ 8.2, and the worst task-specific test RMSE for the Markov models of 6.77).
- In both cases, the GMRF prior actually decreases performance slightly. This will be addressed in more detail in the conclusion.
- Overall, the model with no growth actually outperforms any of the growth models.

V. Discussion

At the outset of this paper, we presented our core hypothesis, namely that the census tracts' spatial structure could be exploited to improve the performance of predictive models. To exploit this spatial structure, we use the GMRF priors described by Lee *et al* in [1], and tried applying this prior in two different classes of models: linear models, and markov chain models.

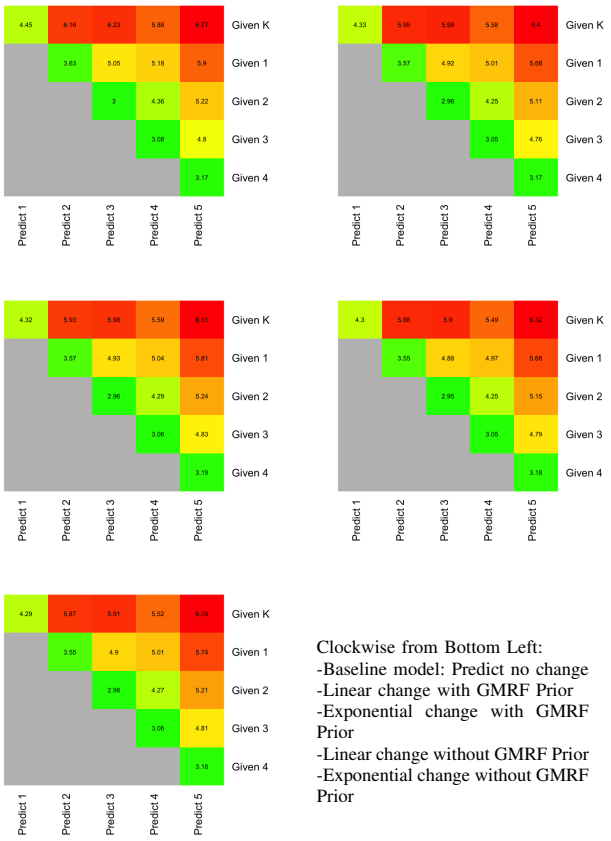

Table 2: RMSE for each prediction task

A. Linear Models

Overall, in our linear models, this prior provided some regularization. As seen in 1, for unregularized models the training set RMSE decreased and the test set RMSE increased as model complexity increased. In contrast, using the GMRF prior, the training error was relatively unaffected by increasing model complexity.

That said, given the near equivalence of the experimental (GMRF prior) and control models for the feature set $\beta_0, \beta_g, \beta_t$, the added time-cost of model fitting (running the hamiltonian monte carlo sampler for several hours) is not justified. Instead, the simple linear model without the GMRF are sufficient-hence these experiments do not support the hypothesis that use of GMRF priors improves the models predictive performance.

B. Markov Chain Models

In model 2 (markov chain growth models), we also got negative results for the hypothesis that spatial struture could be exploited to improve model performance. In these models, the GMRF-regularization actually decreased test set RMSE performance compared to unregularized models. We explain this observed result in terms of the bias-variance trade off. Specifically, this suggests that imposing spatial smoothness on

cohort growth rates adds bias to the tract-level models, and does not return gains in the form of reduced variance. Hence, overall test set performance decreases as a result of this spatial regularization. Finally, we note that our Model experiments demonstrated essentially negligible test set performance gains using growth rate models compared to the simplest possible prediction: simply predicting no change (i.e. $r_k = 0$) for all tracts.

C. Overall

In both cases (linear and Markov chain models), our hypothesis was not supported. Importantly, one explanation for this is the granularity of the data (temporally, spatially, and in terms of the demographic specificity implied in the requirement to make grade-level predictions). While at this level of granularity exploiting spatial structure proved fruitless, we hypothesize this methodology may still prove effective for less granular tasks, since additional aggregation would smooth the data, and thus likely smooth the spatial distribution as well.

VI. Conclusions

Overall, our study was premised on the hypothesis that spatial structure could be exploited to achieve significant gains in predictive performance. Unfortunately, this first premise proved to be incorrect. As such, if we were to continue working on this project, we would likely chose to consume side data and use it to expand the feature space. While we did not identify any features that seemed like reasonable evidence for incorporation in such a fine grained model (i.e. features that would be useful for predicting the number of second graders in a specific census tract, and ultimately a specific city block, for a specific year), perhaps additional research would reveal such a feature.

Additionally, we would also likely explore Hidden Markov Models instead of Markov Chains. Using HMMs, the added hidden layer could potentially usefully smooth the observed counts, and allow for better predictions.

References

[1] D. Lee, A. Rushworth, and G. Napier, "Carbayesst: An r package for spatio-temporal areal unit modelling with conditional autoregressive priors," 2015.

[2] J. Besag, J. York, and A. Mollié, "Bayesian image restoration, with two applications in spatial statistics," *Annals of the institute of statistical mathematics*, vol. 43, no. 1, pp. 1–20, 1991.

[3] B. G. Leroux, X. Lei, and N. Breslow, "Estimation of disease rates in small areas: a new mixed model for spatial dependence," in *Statistical models in epidemiology, the environment, and clinical trials*. Springer, 2000, pp. 179–191.

[4] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *J Stat Softw*, 2016.