Predicting New York City School Enrollment

Timothy Jones

joint work with
Jonathon Auerbach & Robin Winstanley

StanCon 2018

August 31, 2018

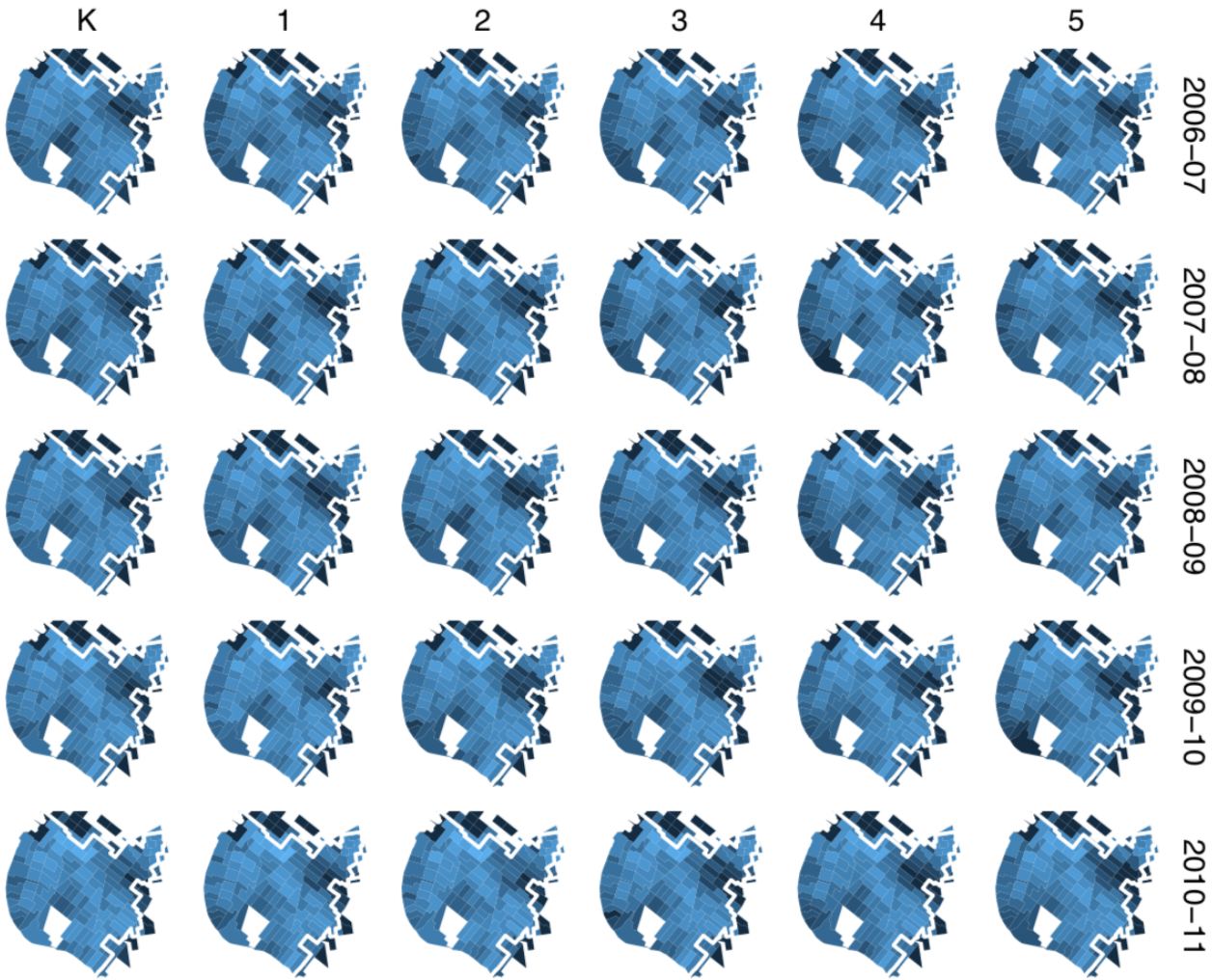
We won a competition predicting school enrollment a year ago...

- we thought a well-reasoned hierarchical model would beat out-of-the-box machine learning methods as spatial-temporal data is difficult to cross validate.
- now we are in on-going project with the NYC Department of Education building out our approach.
- Please see out paper as this presentation is streamlined.

NYC student enrollment is difficult but important to predict

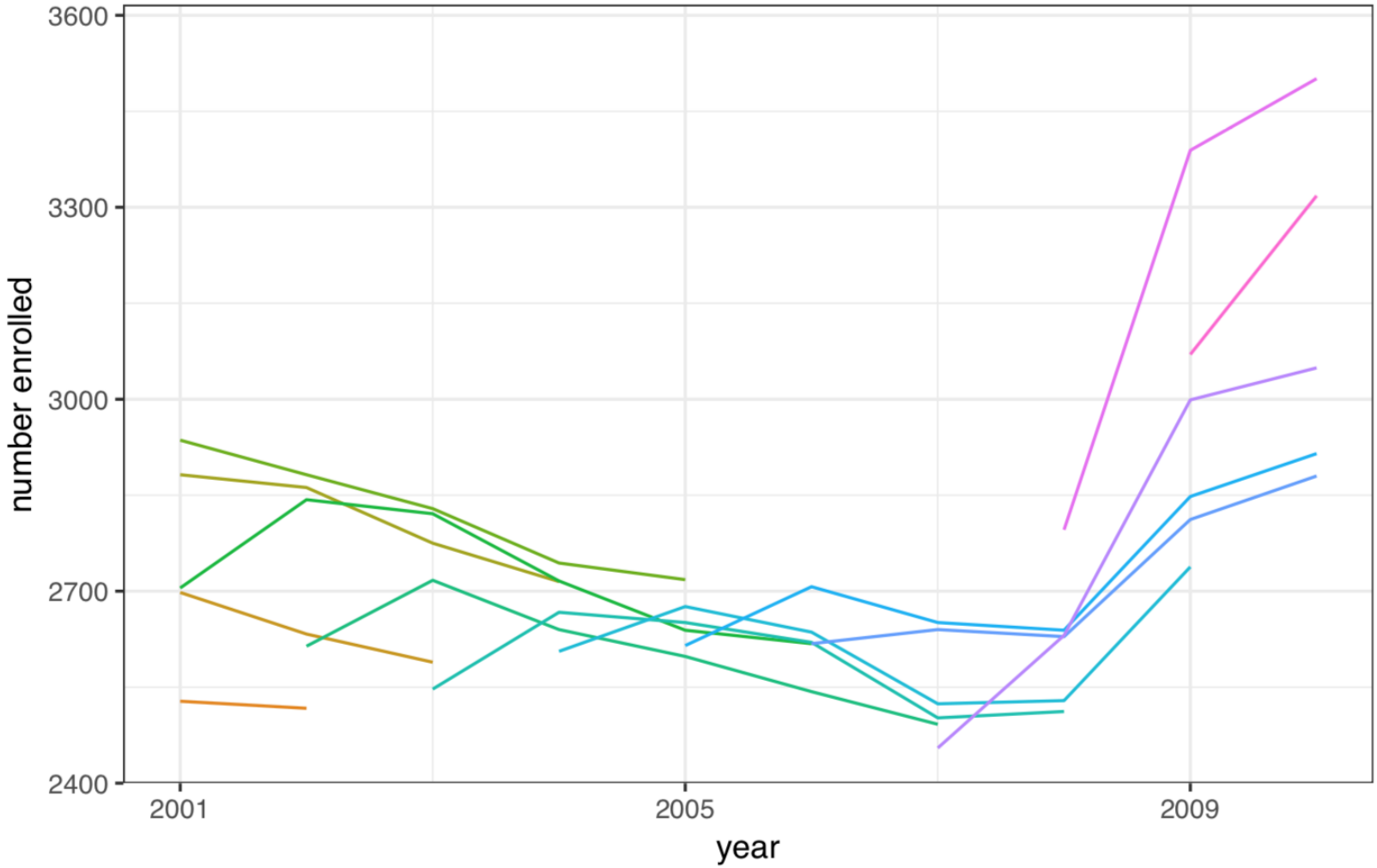
- New York City distributes billions of dollars to schools each year.
- Department of Education depends on student enrollment predictions to fairly and effectively build infrastructure and allocate funds across schools.
- Accurate enrollment predictions are difficult to make in New York City. Even if only for a year into the future. A continuous flow of residents immigrate to the City and then constantly relocate across its neighborhoods.

DOE gave us the School District 20 from 2001-2002 to 2010-2011 school years ...



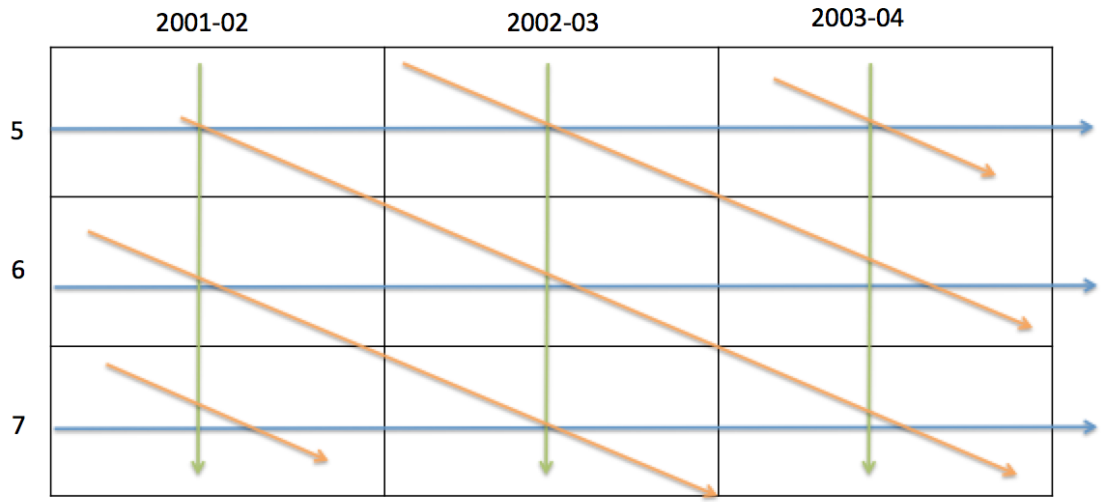
... we aggregated across tracts and plotted cohorts

We see long term decline (gentrification) and a short term spike (recession). Will the spike continue?



Trends can be decomposed into grade (age), period, and cohort effects ...

$$Y_{ij} = \mu + P_i + G_j + C_{i-j} + \epsilon_{ij}$$
$$\epsilon_{ij} \sim N(0, \sigma^2)$$



... but this decomposition cannot be learned from the data alone.

$$Y_{ij} = \mu + P_i + G_j + C_{i-j} + \epsilon_{ij}$$
$$\epsilon_{ij} \sim N(0, \sigma^2)$$

- This model is (Frequentist) unidentifiable - the design matrix is one less than full rank.
- We need another constraint to break unidentifiability.

How do we disentangle long term cohort effects and transient period effects so that we don't overfit and forecast trends that won't persist?

Constraints

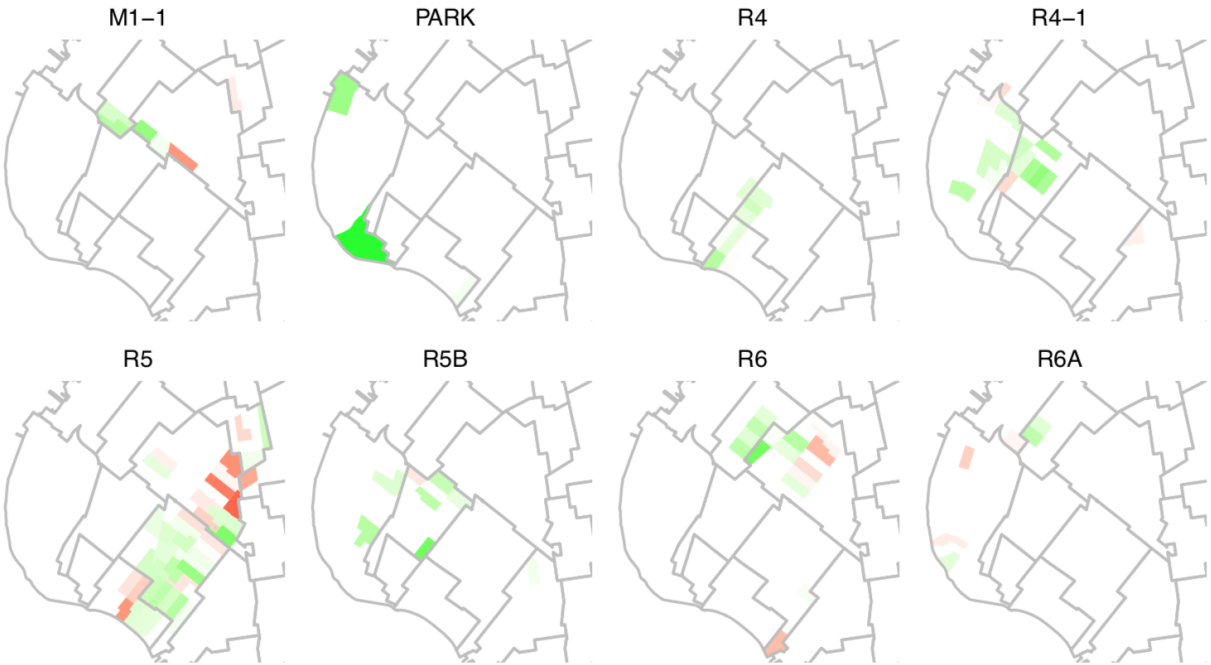
- We can constrain the model parameters by introducing priors.
- How should we do this?

Default priors for the Grade-Period-Cohort Model

$$
\begin{aligned}
Y_{ij}^t &\sim \text{Poisson}(\exp[\mu^t + P_i^t + G_j^t + C_{i-j}^t])) \\
\mu^t &\sim \text{Normal}(0, \sigma_\mu) \\
P_i^t &\sim \text{Normal}(0, \sigma_{P_i}) \\
G_j^t &\sim \text{Normal}(0, \sigma_{G_j}) \\
C_{i-j}^t &\sim \text{Normal}(0, \sigma_{C_{i-j}})
\end{aligned}
$$

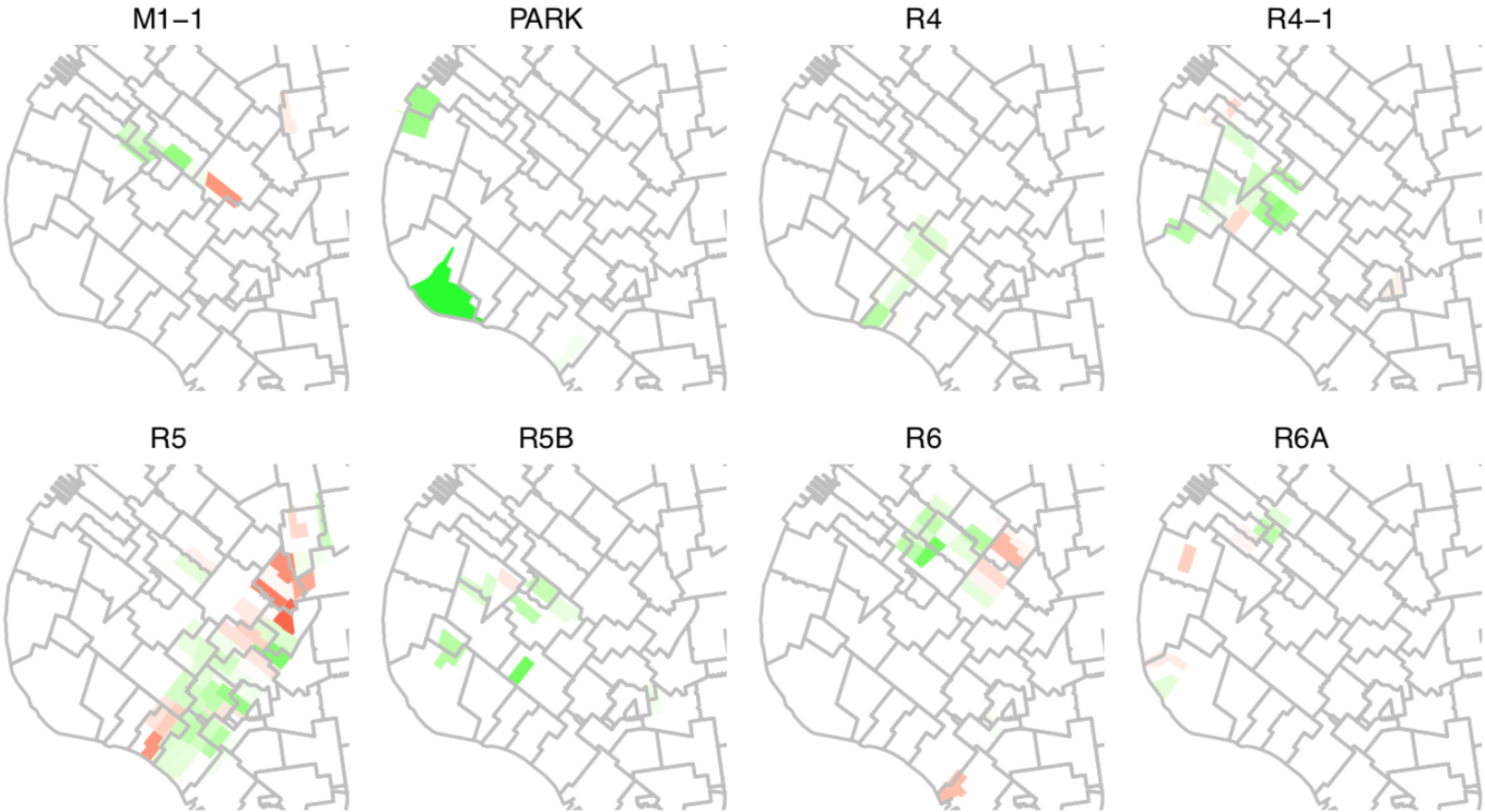
Spatial information can be used to interpret the model

- If the prior is arbitrary, the coefficients are more like factors than effects.
- We can add spatial information to interpret these factors.
- The underlying determinants of these changes can be deduced by plotting them against neighborhood, school zone and land use shapefiles.
- Enrollment change by neighborhood:



Spatial information can be used to interpret model factors

- Enrollment change by School Zone



Takeaway from Spatial Plots

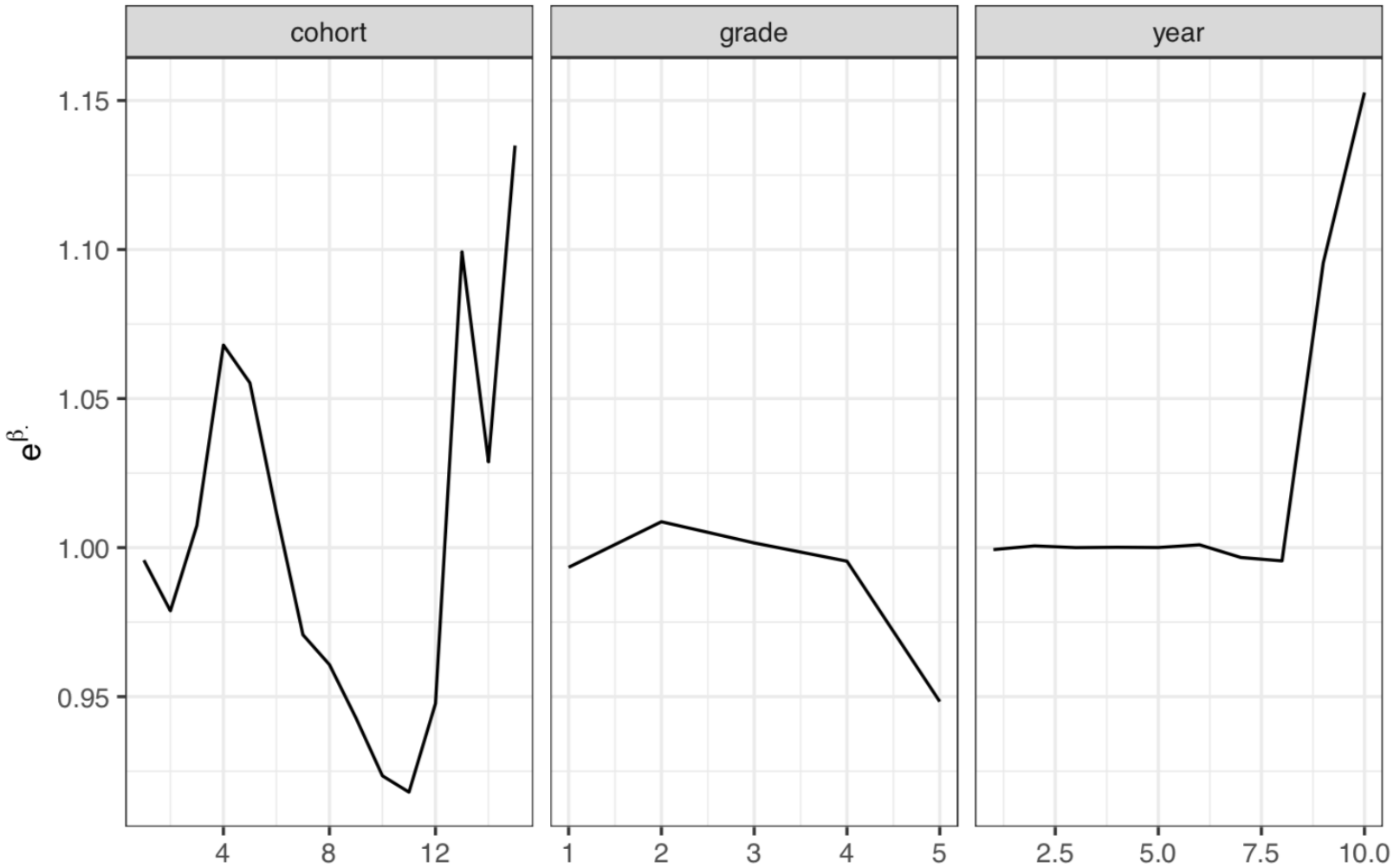
- In general, we find that increases cluster by neighborhood while decreases cluster by school zone.
- This suggests that perhaps the former is the result of short-term economic forces, and the latter is the result of long-term revitalization.

The current model incorporates spatial information

$$
\begin{aligned}
Y_{ij}^t &\sim \text{Poisson}(\exp[\mu^t + P_i^t + G_j^t + C_{i-j}^t])) \\
\mu^t &\sim \text{Normal}(0, \sigma_\mu) \\
P_i^t &\sim \text{Normal}(Z_P^t + H_P^t + L_P^t, \sigma_{P_i}) \\
G_j^t &\sim \text{Normal}(Z_G^t + H_G^t + L_G^t, \sigma_{G_j}) \\
C_{i-j}^t &\sim \text{Normal}(Z_C^t + H_C^t + L_C^t, \sigma_{C_{i-j}}) \\
Z_\cdot^t &\sim \text{Normal}(0, \sigma_\cdot^t) \\
H_\cdot^t &\sim \text{Normal}(0, \sigma_\cdot^t) \\
L_\cdot^t &\sim \text{Normal}(0, \sigma_\cdot^t)
\end{aligned}
$$

Results

- Cohort effects decrease until Great Recession.
- After the recession, the period effect jumps along with the cohort effects.



Now that we interpreted these effects, there are many methods to extrapolate for future years. Please see our paper for more information.

Thank You