Predicting New York City School Enrollment

Jonathan Auerbach, Timothy Jones, and Robin Winstanley

July, 29 2018

Abstract

We propose a Bayesian hierarchical Age-Period-Cohort model to predict elementary school enrollment in New York City. We demonstrate this model using student enrollment data for grades K-5 in each Census Tract of Brooklyn's 20th School District over the 2001-02 to 2010-11 school years. Specifically, our model disaggregates enrollment into grade (age), year (period), and cohort effects so that each can be interpreted and extrapolated over the 2011-12 to 2017-18 school years. We find this approach ideal for incorporating spatial information indicative of the socioeconomic forces that determine school enrollment in New York City. This work is the result of a 2016 "Call for Innovation" initiated by the Department of Education, and it won the grand prize for having a lower prediction error than competing teams on a held out test set. We thank our other two team members: Susanna Makela and Swupnil Sahai, as well as the Department of Education's Office of District Planning, and the Department of City Planning's Population Division.

1. Introduction

School districts predict student enrollment in order to determine whether their schools will be adequately staffed and supplied. For New York City, the largest school district in the United States, billions of dollars in discretionary and capital funds are distributed among more than a million students each year. Officials at the New York City Department of Education depend on enrollment predictions, over small areas as far as a decade into the future, to set administrative boundaries and to decide how to fairly and effectively build infrastructure and allocate funds within those boundaries.

However, accurate enrollment predictions are notoriously 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. In fact, nearly forty percent of New York City residents were born abroad in 2014, and a fifth of all residents moved between 2012 and 2014, according to the Housing Vacancy Survey (Gaumer 2018). These residents follow a complex array of socioeconomic forces, which—despite their complexity—have consistently guided waves of migrant groups from Manhattan, the center of New York City, to the surrounding boroughs. These movements then displace previous migrant groups, resulting in reverberations that systematically remake the demographic profile of every neighborhood in the City.

In this paper, we consider two Bayesian hierarchical Age-Period-Cohort models to predict elementary school enrollment by grade (age) and year (period) for each Census Tract. We use Brooklyn's 20th School District to demonstrate our approach. A key feature of Age-Period-Cohort models are their interpretability, and we devote the remainder of this first section to highlighting the various socioeconomic forces that shape migration. We do this on a Brooklyn-wide scale before considering specific developments within the 20th School District. Both reviews are important because Age-Period-Cohort models do not identify the actual causes that generated the observed data. Instead, they suggest structure, which, not unlike factor models, requires contextualization to interpret.

In Section 2, we review the educational planning literature on enrollment prediction and the demography literature on the Age-Period-Cohort model. We conclude this section by discussing the enrollment data of Brooklyn's 20th School District in light of this review. In Section 3, we fit two models. The first is the traditional log-linear Age-Period-Cohort regression, in which the expected enrollment in each tract is the product of an grade (age), year (period), and cohort effect. Each cohort effect corresponds with the year the students started Kindergarten. The second allows the grade, year and cohort effects to vary by neighborhood, school zone, and land use (zoning district). In both models, we parameterize the 4th and 5th grade effects to be equal in order to eliminate perfect multicollinearity. Finally, we conclude the analysis in Section 4 by interpreting these effects and making predictions.

1.1 Twentieth Century Brooklyn: Industrialization, Segregation, and Revitalization

Prior to the twentieth century, Brooklyn was low density farmland, dotted sparsely with the villages of upper-class residents. In fact, as late as 1880, Brooklyn and Queens were considered the vegetable capital of the United States. (Wallace 2017) Density increased rapidly following the consolidation of New York City (1898), the construction of three bridges: Brooklyn (1883), Williamsburg (1903) and Manhattan (1909), and the extension of the rail system (1920), which made it a viable location for factories and their middle and lower-class workers.[1]

This first wave of workers were European immigrants, many Jewish Americans, who at the time made up the largest ethnic group in New York City. By 1917, immigration law had changed, and economic forces drew African Americans from Southern states, in what is referred to as the first Great Migration. (Wallace 2017) However, industrialization ended after the Second World War. In the three decades which followed, New York City transitioned from the shipping and manufacturing center of the United States to a postindustrial

[1]The institution of education as it currently exists in the United States is closely linked to these events. Consolidation opponents feared the "Manhattanization" of Brooklyn, and, in 1901, school attendance became mandatory in New York City for all children younger than 12. Although the reasons for mandating school attendance are still debated by historians, Wallace (2017) points out that around the time of consolidation, seventy percent of the City's students had been born abroad, and vocal interests expressed the need to "Americanize" immigrant children.

economy. The new economy required a much smaller base of skilled workers instead of the large numbers of factory workers that had traditionally been employed. The transition was characterized by disinvestment, underutilized property, and political turmoil, and it culminated in the near bankruptcy of the City in 1975. (Phillips-Fein 2017)

Concurrent with this economic transition, the racial and ethnic composition of Brooklyn continued to shift. A second Great Migration of younger African Americans and Puerto Rican Americans displaced the earlier wave of European immigrants farther into the suburbs. However, displacement did not occur uniformly over Brooklyn. Policy at all levels of government such as housing codes, urban renewal projects, and low-income housing programs concentrated minority populations. Concomitant practices in the real estate industry, such as red lining—the selective approval of loans in specific locations according to an applicant's race or ethnicity—and blockbusting—the dumping of property to rapidly tip neighborhoods to a particular racial or ethnic composition. These forces markedly increased neighborhood segregation. (Rogers 2006) Segregation was further magnified in public school enrollment because wealthier residents increasingly sent their children to private, religious, or parochial schools. (Gittell 1967)

The turmoil of the 60s and 70s began to reverse in the 80s and 90s, due in part to the reopening of immigration in 1965. Within two decades, immigration increased to the level that had characterized pre-war New York City. This time, however, immigrants arrived from Asia and Central America. Urban revitalization made way to gentrification, and the early twenty-first century was marked by the displacement of African American and Puerto Rican American neighborhoods. Recent changes in student enrollment must be viewed in this context, as the late stages of the revitalization of post-industrialized, underutilized property.

1.2 School District 20: a Brooklyn Microcosm

The 20th School District is one of Brooklyn's most ethnically and racially diverse areas, spanning the neighborhoods of Sunset Park, Bay Ridge, Borough Park, and Bensonhurst. However, its land use is typical of Brooklyn as a whole, making it the ideal case to study enrollment. Figures 1 and 2 show the primary zoning for land use in 2010. The 20th School District was zoned roughly 76 percent residential, 7 percent manufacturing, 15 percent parks, and 2 percent commercial by area. Meanwhile, Brooklyn was zoned roughly 72 percent residential, 16 percent manufacturing, 8 percent parks, and 4 percent commercial by area.

District 20's large manufacturing zones originate from the turn of the twentieth century when the western waterfront was first developed as a manufacturing and garment shipping district, as previously discussed. Most notable was the industrial colony Bush Terminal (1902) and the Brooklyn Army Terminal (1919). The development of the residential area on the eastern edge of the district followed after Long Island Railroad and BRT both completed train lines by 1920, allowing for single family homes populated largely by Norwegian and Finnish Americans.(Wallace 2017)

Despite the end of the manufacturing boom nearly a century ago, this history continues to shape District demographics today, owing to a series of government aid and increased immigration. The area now represents one of the largest concentrations of Asian immigrants, roughly a tenth of all Asian residents in New York City, and the population continues to change. This analysis divides the factors underlying change into three groups: 1. grade-specific factors associated with the typical reasons for matriculation and attrition, 2. year-specific factors associated with short-term fluctuations in migration between other neighborhoods or between charter, private, religious, and parochial schools —in this case largely brought on by the Great Recession— and 3. cohort-specific factors associated with long-term changes to land use and the demographic makeup of the area. We believe estimating these factor groups separately —and even stratifying them by covariates— is important for accurate forecasting as it allows planners to apply heuristics that establish the relative importance of these factor groupings. For example, planners might expect long-term changes to continue at their historic rate throughout the forecast period while period-specific fluctuations might be allowed to change sporadically, and grade-specific factors might remain unchanged.

```
packages <- c("rgeos", "rgdal", "maptools", "plyr", "reshape2", "MASS",
              "ggplot2", "dplyr", "rstan", "StanHeaders", "gridExtra",
              "viridis")
```

```
lapply(packages, require, character.only = TRUE)
rm(packages)
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())

setwd('data/')

# Load NYC Census Tracts
tracts <- readOGR("nyct2010_16c/", "nyct2010")
tracts@data$id <- rownames(tracts@data)
tracts_points <- fortify(tracts, region = "id")
tracts_df <- join(tracts_points, tracts@data, by = "id")
tract_centers <- SpatialPoints(coordinates(tracts),
                               proj4string = tracts@proj4string)

# Load NYC School Districts (2010)
districts <- readOGR("nysd_10c_av/","nysd")
districts@data$id <- rownames(districts@data)
district_points <- fortify(districts, region="id")
district_df <- join(district_points, districts@data, by="id")

# Load NYC School Zones (2010)
zones <- readOGR("2010 - 2011 School Zones/",
                 "geo_export_a2458df3-9b69-4106-9bd4-11c2df6df9b6")
zones@data$id <- rownames(zones@data)
zones <- spTransform(zones, tracts@proj4string)

# Load NYC Zoning Districts January (2010)
land <- readOGR("zoning/nycgiszoningfeatures_201001_shp/",
                "nyzd")
land@data$id <- rownames(land@data)
land@data$ZONEDIST <- toupper(land@data$ZONEDIST)
land_points <- fortify(land, region="id")
land_df <- join(land_points, land@data, by="id")

# Determine which zone and neighborhood each tract is in
districts_tracts <- over(tract_centers, districts)
zones_tracts <- over(tract_centers, zones)
land_tracts <- over(tract_centers, land)

districts_tracts$BoroCT2010 <- tracts@data$BoroCT2010
zones_tracts$BoroCT2010 <- tracts@data$BoroCT2010
land_tracts$BoroCT2010 <- tracts@data$BoroCT2010

# Combine the data for each Census Tract
tracts_data <- data.frame(BoroCT2010 = tracts@data$BoroCT2010,
                          CT2010 = tracts@data$CT2010,
                          NTA = tracts@data$NTACode,
                          Zone = zones_tracts$id,
                          District = districts_tracts$id,
                          Land = land_tracts$ZONEDIST)

# Load Enrollment Data
```

```
doe_data <- read.csv('CSD20_Resident_Data_Phase_1.csv')
enrollment <- expand.grid(unique(doe_data$X2010.Census.Tract),
                          sort(unique(doe_data$School.Year)),
                          levels(doe_data$Grade.Level))
colnames(enrollment) <- c("X2010.Census.Tract", "School.Year", "Grade.Level")
enrollment <- left_join(enrollment, doe_data,
                        by = c("X2010.Census.Tract",
                               "School.Year",
                               "Grade.Level"))
enrollment$Count.of.Students[is.na(enrollment$Count.of.Students)] <- 0

enrollment$CT2010 <- as.character(enrollment$X2010.Census.Tract)
enrollment$CT2010[nchar(enrollment$X2010.Census.Tract) == 4] <-
  paste(substr(enrollment$CT2010,1,2),
        substr(enrollment$CT2010,3,4), sep = ".")[
          nchar(enrollment$X2010.Census.Tract) == 4]
enrollment$CT2010[nchar(enrollment$X2010.Census.Tract) == 5] <-
  paste(substr(enrollment$CT2010,1,3),
        substr(enrollment$CT2010,4,5), sep = ".")[
          nchar(enrollment$X2010.Census.Tract) == 5]
enrollment$CT2010 <- 100 * as.numeric(enrollment$CT2010)
enrollment$CT2010 <- ifelse(nchar(enrollment$CT2010) == 4,
                            paste0("00", enrollment$CT2010),
                            paste0("0", enrollment$CT2010))
enrollment$CT2010 <- factor(enrollment$CT2010,
                            levels = levels(tracts_df$CT2010))
enrollment <- enrollment[!is.na(enrollment$CT2010),]

enrollment$BoroCT2010 <- factor(paste0("3", enrollment$CT2010),
                                levels = levels(tracts_df$BoroCT2010))

tracts_df <- left_join(tracts_df, tracts_data,
                       by = c("CT2010", "BoroCT2010"))
tracts_df <- left_join(tracts_df, enrollment,
                       by = c("CT2010", "BoroCT2010"))
tracts_df$Grade.Level <- relevel(tracts_df$Grade.Level, "K")
```

```
#Percent of Borough/District by Primary Zoning
districts_land <- over(SpatialPoints(coordinates(land),
                                     proj4string = land@proj4string),
                       districts)
boroughs_land <- over(SpatialPoints(coordinates(land),
                                    proj4string = land@proj4string),
                      unionSpatialPolygons(tracts, tracts@data$BoroCode))
land@data$District <- districts_land$id
land@data$BoroCode <- boroughs_land

area_bk <- aggregate(SHAPE_area ~ substr(ZONEDIST,1,1),
                     land@data[land@data$BoroCode == 3,],
                     sum)
area_d20 <- aggregate(SHAPE_area ~ substr(ZONEDIST,1,1),
                      land@data[which(land@data$BoroCode == 3 &
                                      land@data$District == 4),],
                      sum)
```

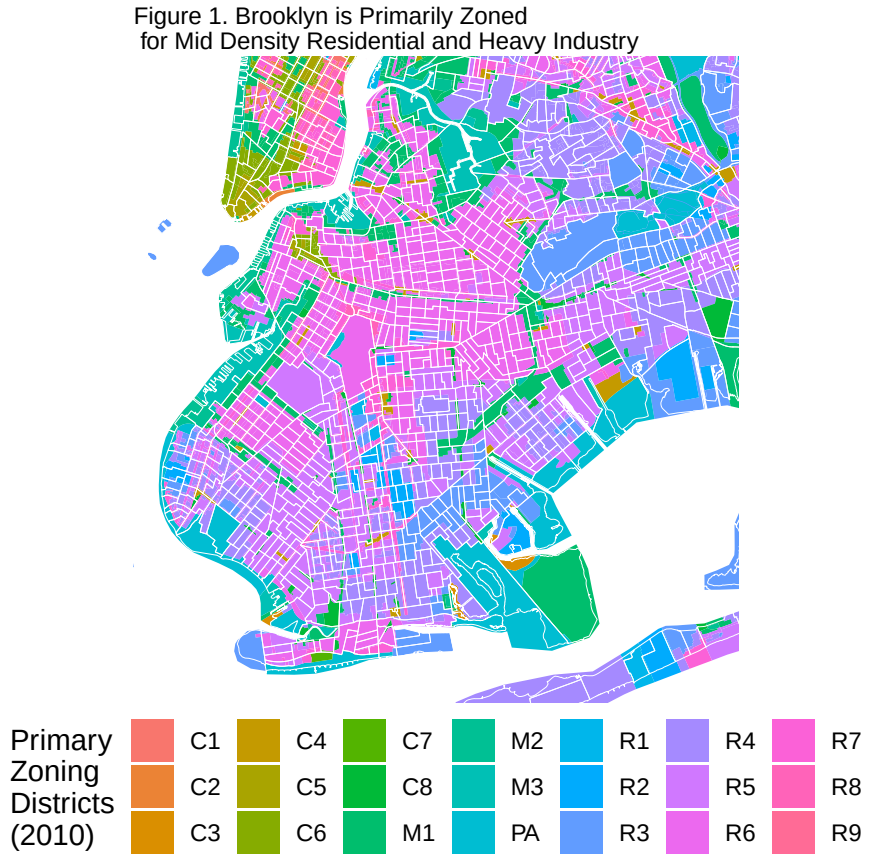

Figure 1: Primary zoning for land use in Brooklyn in 2010. Brooklyn is primarily zoned for Mid Density Residential and Heavy Industry. The letters in the legend correspond to the following classifications, R: residential, M: manufacturing, C: commercial, P: park. The numbers correspond to density, from lowest density (1) to highest density (9).

```
ggplot() +
  theme_void() +
  geom_polygon(aes(long, lat, group = group,
                   fill = substr(ZONEDIST,1,2)),
               data = land_df) +
  geom_polygon(aes(long, lat, group = group), fill = NA, color = "white",
               size = .1,
               data = tracts_df) +
    coord_fixed(xlim = range(tracts_df$long[tracts_df$BoroCode == 3],
                             na.rm = TRUE),
              ylim = range(tracts_df$lat[tracts_df$BoroCode == 3],
                             na.rm = TRUE)) +
  theme(legend.position = "bottom",
        plot.title = element_text(hjust = 0,
                                  size = 9)) +
  guides(fill=guide_legend(ncol=7)) +
  labs(fill = "Primary\nZoning\nDistricts\n(2010)",
       title = "Figure 1. Brooklyn is Primarily Zoned\n for Mid Density Residential and Heavy Industry")
```

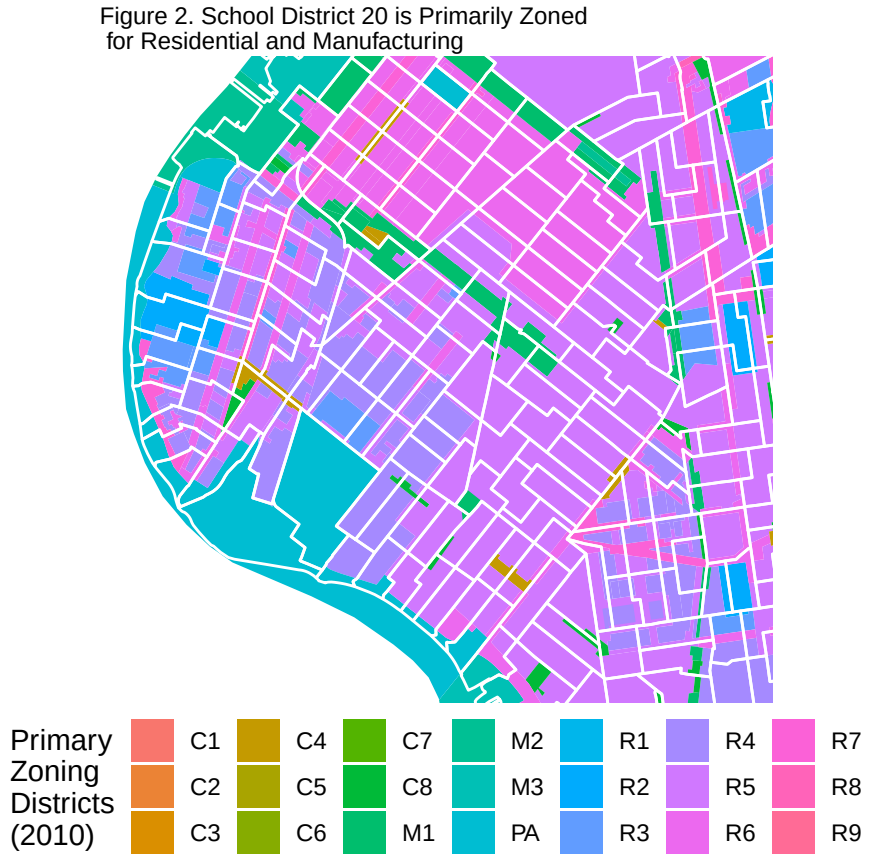

Figure 2: Primary zoning for land use in School District 20 in 2010. The District is primarily zoned for Residential and Manufacturing. The letters in the legend correspond to the following classifications, R: residential, M: manufacturing, C: commercial, P: park. The numbers correspond to density, from lowest density (1) to highest density (9).

```
ggplot() +
  theme_void() +
  geom_polygon(aes(long, lat, group = group,
                   fill = substr(ZONEDIST,1,2)),
               data = land_df) +
  geom_polygon(aes(long, lat, group = group), fill = NA, color = "white",
               data = tracts_df) +
    coord_fixed(xlim = range(tracts_df$long[tracts_df$District == 4],
                             na.rm = TRUE),
                ylim = range(tracts_df$lat[tracts_df$District == 4],
                             na.rm = TRUE)) +
  theme(legend.position = "bottom",
        plot.title = element_text(hjust = 0,
                                  size = 9)) +
  guides(fill=guide_legend(ncol=7)) +
  labs(fill = "Primary\nZoning\nDistricts\n(2010)",
       title = "Figure 2. School District 20 is Primarily Zoned\n for Residential and Manufacturing")
```

2. Data

The use of enrollment projections for educational planning dates back at least half a century. In the two decades following the Second World War, the academic literature was concerned with addressing the inequality that resulted from the United States' abrupt shift to a postindustrial economy. Researchers quickly recognized the importance of measuring the factors that determine population change, such as family stability, housing policy, economic status, and social class. (Gittell 1967) — even before Ryder popularized cohort analysis at the 1970 meeting of the American Sociological Association. (Ryder 1985) But it took much longer to appreciate the methodological issues that complicate the identification of the effects of these factors from the data. Issues that preclude many of the perfunctory analyses often used by researchers.

These methodological issues were summarized for educational planners by Vinovskis, who criticized the widespread practice of researchers "data gathering and analysis without adequately trying to conceptualize the issue they want to investigate". (Goodenow and Ravitch 1983) Vinovskis cited multiple examples where a careful reframing of a hypothesis reversed the conclusions of high-profile results. However, the problems he cites are not confined to the Twentieth Century. Ravitch (2010) provides a recent, high-profile example of how sociodemographic shifts continue to be conflated with policy outcomes.

The majority of these criticisms are recognized in the social sciences as problems of identification: the researcher cannot meaningfully characterize the stated quantity of interest from the data alone, and any analysis is sensitive to the framing or context of the investigation.[2] The lack of identification encountered in the current analysis —where the goal is to estimate latent factors that determine enrollment— is referred to in sociology as the "Age-Period-Cohort" problem, which we now briefly review.

2.1 The Age-Period-Cohort Problem

The Age-Period-Cohort model is a general framework for analyzing longitudinal data, where groups of subjects are repeatedly measured at regular intervals. The random variable of interest, Y_{ij}, a group-level outcome measured during period j when the group is age i, is thought to be the sum of period, age and cohort effects. i.e.:

$$Y_{ij} = \mu + P_i + G_j + C_{i-j} + \epsilon_{ij}$$

where P, G, and C are fixed effects for period, age, and cohort, and ϵ is random measurement error, all satisfying the usual constraints: $\sum P_i = \sum G_j = \sum C_{c=i-j} = \sum \sum \epsilon_{ij} = 0$.[3]

Sociologists quickly realized that the model parameters, although constrained, are still unidentified due to the linear relationship between cohort, age, and period (Kupper et al. 1985), (K. O. Mason et al. 1973). The design matrix is one less than full rank, thus adding any additional constraint — such as setting two period effects equal — will identify the model. (Fienberg and Mason 1979) If the constraint is true, the least squares estimate will be unbiased. However, the modeler may not know a priori if any such constraint holds and even if the modeler correctly specifies a true constraint, measurement and sampling error can lead to highly inaccurate estimates. (Rodgers 1982)

Yang (2006) propose a hierarchical Age-Period-Cohort model that addresses unidentifiability by incorporating informative priors. However, Bell and Jones (2018) demonstrate this approach cannot disentangle the age, period, and cohort effects in general. They argue hierarchical priors imply constraints that are no more useful than setting two parameters equal and provide simulations where the model does not recover the parameters used to generate the data.

[2] We use the word "identification" more broadly than the traditional definition that two likelihoods are equal only if the parameters are equal. We can find no agreed upon definition among Bayesians, owing perhaps to the various philosophical stances that occupy the field. See http://andrewgelman.com/2014/02/12/think-identifiability-bayesian-inference/.

[3] Interactions between the effects are identified and could be included in the model. However, interactions complicate the interpretation of the parameters and have been excluded from our analysis.

For this analysis, we treat the Age-Period-Cohort model like a factor model. We view age, period, and cohort effects as proxies for a variety of underlying, unobserved factors. (Rodgers 1982) That is, our reliance on age, period, and cohort indices is purely for convenience. But in using these indices, we see no practical way to establish constraints or informative priors that correspond with actual knowledge of the unobserved factors. To interpret the model parameters, we rely on a hierarchical model that incorporates spatial variation within and between shapefiles demarking school zones, neighborhoods and land use.

2.2 Graphical Inspection of Enrollment Data

A portion of the enrollment data is displayed with maps in Figure 3. The maps are arranged in a contingency table, and the color of each Census Tract within each map represents the number of students enrolled in each grade (columns) and each school year (rows). Lighter colors mean more students are enrolled. In this case, the Age-Period-Cohort model in Section 2.1 corresponds to a log linear model with column (grade) effects, row (year) effects, and diagonal (cohort) effects. Note that for each tract, we have six grade parameters, ten year parameters, and fifteen (6 + 10 - 1) cohort parameters. Without any additional structure, there are thirty-one parameters per Census Tract.

Enrollment changes systematically about all three indices, although this is difficult to see from Figure 3. In Figure 4, enrollment has been aggregated across all District 20 Census Tracts for each year and cohort. The color of each line corresponds to a different cohort. The twenty-five percent increase in enrollment after the Great Recession of 2008 is now visible, as is the slower, steady enrollment decrease between 2001 and 2007. However, these changes appear to result from different factors.

The 2001-07 decrease is driven by cohort changes: each successive cohort begins with fewer enrolled than the previous one. There appears to be little, if any, systematic change among cohorts within each period. However, the increase in 2008-10 corresponds to a large period effect: all cohorts increase their enrollment from the previous year by similar amounts. The approximate cause of these changes can be deduced by plotting them against neighborhood, school zone, and land use shapefiles.

Figures 5-8 suggest that these changes are exclusive to specific neighborhoods, school zones, and land uses. In Figures 5 and 6, Census Tracts have been colored green for increases in enrollment and red for decreases in enrollment from the 2001-02 to 2010-11 school years. Brighter colors indicate larger changes. The Tracts are then stratified by land use (in the labels, R represents residential, M represents manufacturing, C represents commercial and P represents Park. Larger numbers after the first letter refer to allowing greater density). 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.

Figures 7 and 8 recreate Figure 4 for select land use types. Figure 7 stratifies by neighborhood, and Figure 8 stratifies by school zone. The log-linear model of ages, periods, and cohorts, specified in Section 2.1 appears plausible within covariate strata. In the final version of this paper, we will end this section with a more detailed discussion of these plots.

```
tracts_df$year_labels <- factor(tracts_df$School.Year,
                                labels = paste0(2001:2010,"-",
                                                c(rep(0,8),rep("",2)),
                                                2:11))

ggplot(district_df, aes(x = long, y = lat, group = group)) +
  theme_void() +
  geom_polygon(aes(x = long, y = lat, group = group,
                   fill = log(Count.of.Students + 1)),
               tracts_df[which(tracts_df$School.Year > 20060000),]) +
  geom_polygon(color = "white", fill = NA) +
  facet_grid(year_labels ~ Grade.Level) +
  coord_fixed(xlim = range(tracts_df$long[tracts_df$District == 4],
```

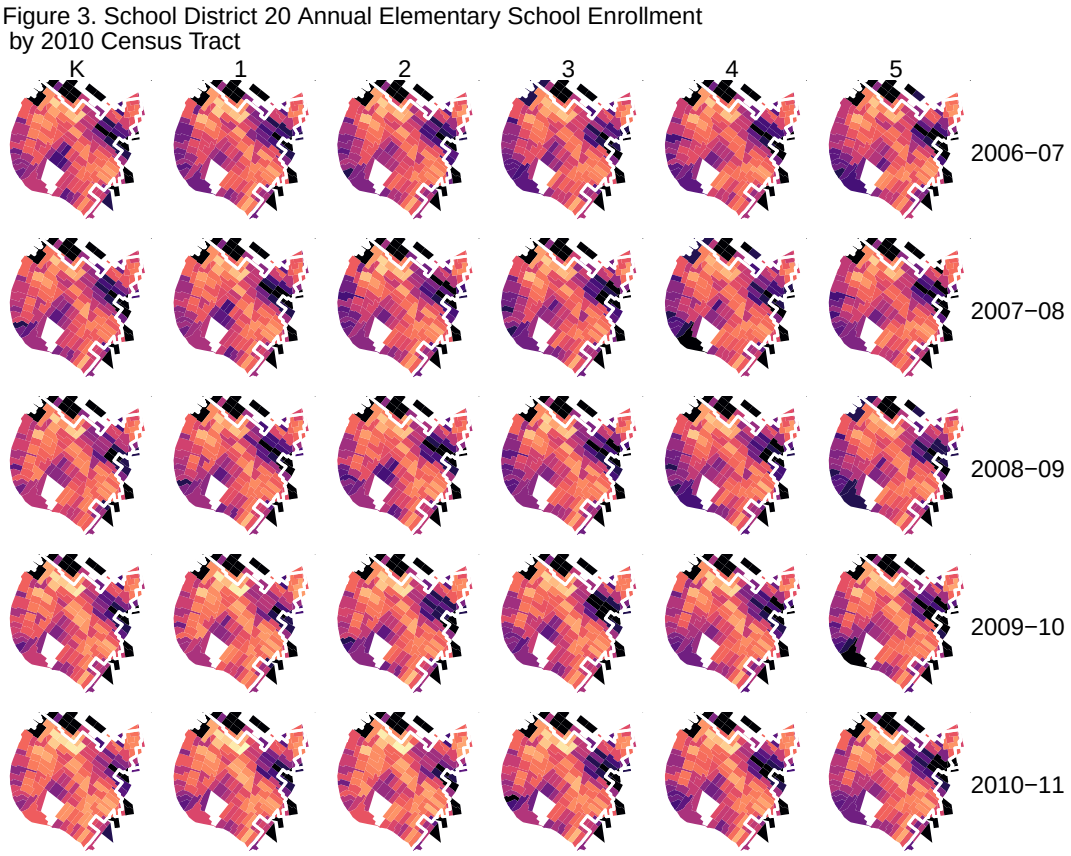

Figure 3: Aggregate student enrollment in School District 20 by 2010 Census Tract. The columns represent grades, ranging from K through 5. The rows represent years, ranging from 2006-2007 to 2010-2011. The diagonals correspond to cohorts.

```
                            na.rm = TRUE),
              ylim = range(tracts_df$lat[tracts_df$District == 4],
                            na.rm = TRUE)) +
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0,
                                  size = 9)) +
  labs(title = "Figure 3. School District 20 Annual Elementary School Enrollment\n by 2010 Census Tract
  scale_fill_viridis(option = "magma")
```

```
enroll_plot <- enrollment
enroll_plot$Year <- as.numeric(factor(enrollment$School.Year))
enroll_plot$Grade <- as.numeric(relevel(enrollment$Grade.Level, "K"))
enroll_plot$Cohort <- enroll_plot$Year - enroll_plot$Grade

ggplot(aggregate(Count.of.Students ~ Year + Cohort,
                 data = enroll_plot,
                 FUN = sum)) +
  theme_bw() +
  aes(2000 + Year, Count.of.Students, color = factor(Cohort)) +
  geom_line() +
  labs(y = "number enrolled", x = "year",
```

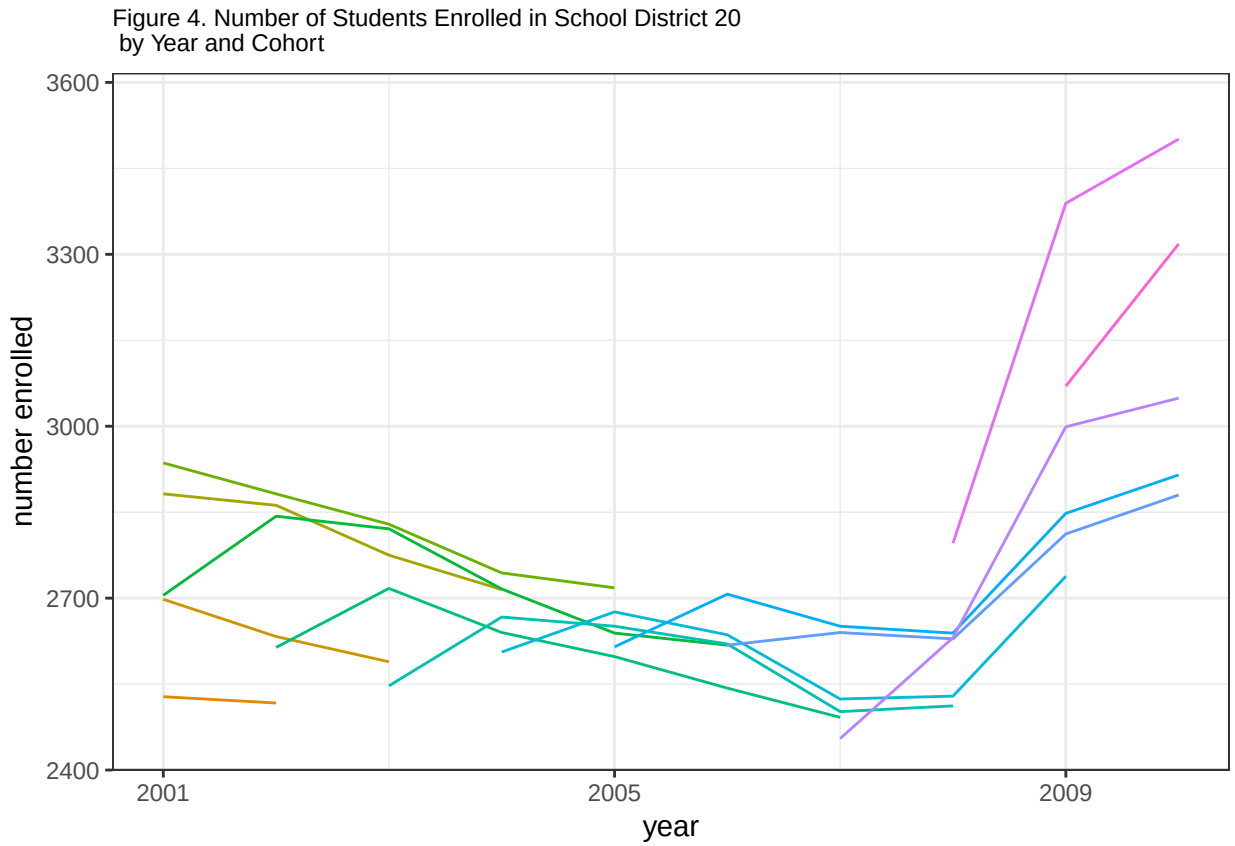

Figure 4: Aggregate student enrollment in School District 20 by year. Each colored line indicates a separate cohort of students.

```
      title = "Figure 4. Number of Students Enrolled in School District 20\n by Year and Cohort") +
  scale_x_continuous(breaks = c(2001, 2005, 2009)) +
  theme(plot.title = element_text(hjust = 0, size = 9),
        legend.position = "none")
```

```
change_df <- inner_join(tracts_df[which(tracts_df$School.Year == 20012002), ],
                        tracts_df[which(tracts_df$School.Year == 20102011), ],
                        by = colnames(tracts_df)[c(1:23,25)])

change_df <- aggregate(cbind(Count.of.Students.x, Count.of.Students.y) ~ long +
                     lat + order + hole + piece + id + group + CTLabel +
                     BoroCode + BoroName + CT2010 + BoroCT2010 + CDEligibil +
                     NTACode + NTAName + PUMA + Shape_Leng + Shape_Area + NTA +
                     Zone + District + Land + X2010.Census.Tract +
                     School.Year.x + School.Year.y,
                   data = change_df,
                   FUN = sum)

ggplot() +
  theme_void() +
  geom_polygon(aes(x = long, y = lat, group = group,
                    fill = log(Count.of.Students.y) - log(Count.of.Students.x)),
               data = change_df[which(change_df$Land %in%
```

Figure 5. School District 20 Enrollment Change from 2001−2002 to 2010−2011
by Neighborhood and Select Primary Zoning District

Figure 5: Changes in aggregate student enrollment in School District 20 from 2001-2002 to 2010-2011, by primary zoning district. Red indicates a decrease in enrollment while green indicates an increase. In Figure 5, neighborhood boundaries are shown in grey.

```
                        names(which(table(change_df$Land,
                                          change_df$District == 4)[,2] > 125))),
                ]) +
  facet_wrap(~ Land, ncol = 4) +
     geom_polygon(aes( x = long, y = lat, group = group),
                fill = NA, color = "grey",
                data = fortify(unionSpatialPolygons(tracts,
                                                    tracts@data$NTAName))) +
  coord_fixed(xlim = range(tracts_df$long[tracts_df$District == 4],
                           na.rm = TRUE),
              ylim = range(tracts_df$lat[tracts_df$District == 4],
                           na.rm = TRUE)) +
  theme(legend.position = "bottom",
        plot.title = element_text(hjust = 0,
                                  size = 9)) +
  labs(color = "Primary Zoning",
    title = "Figure 5. School District 20 Enrollment Change from 2001-2002 to 2010-2011\n by Neighborho
  scale_fill_gradient2(low = "red", high = "green", midpoint = 0,
                       na.value = "white", guide = FALSE)
```

```
ggplot() +
  theme_void() +
  geom_polygon(aes(x = long, y = lat, group = group,
                   fill = log(Count.of.Students.y) - log(Count.of.Students.x)),
               data = change_df[which(change_df$Land %in%
                          names(which(table(change_df$Land,
```

Figure 6. School District 20 Enrollment Change from 2001−2002 to 2010−2011
by School Zone and Select Primary Zoning District

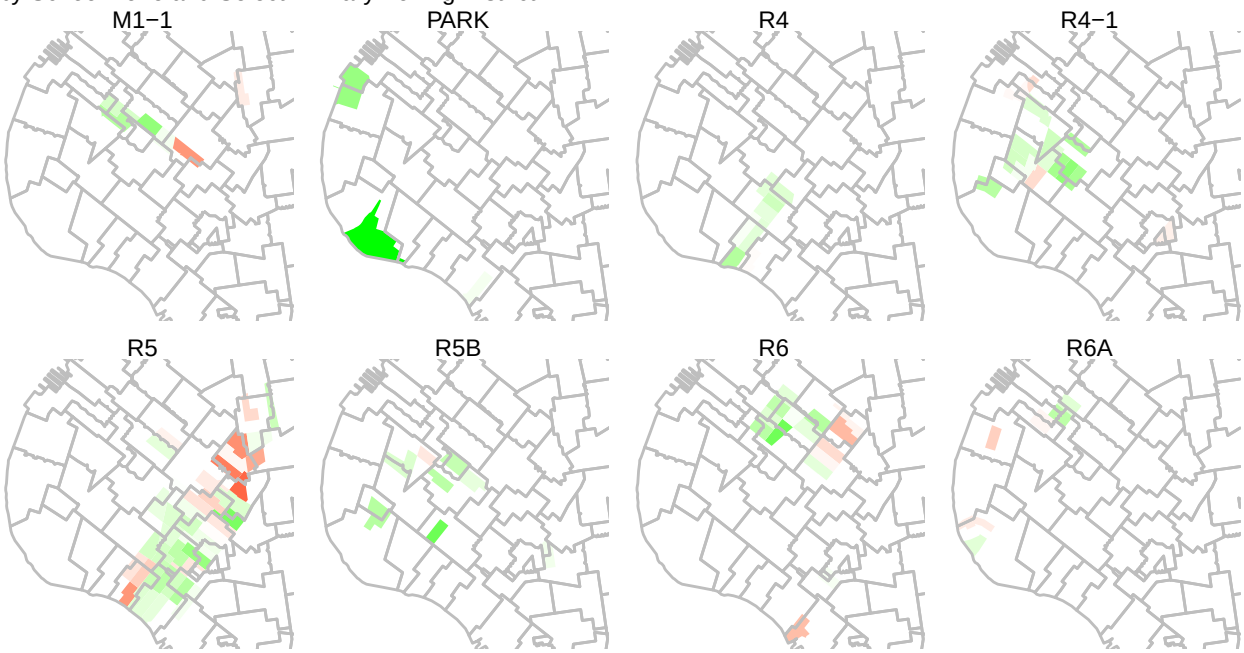

Figure 6: Changes in aggregate student enrollment in School District 20 from 2001-2002 to 2010-2011, by primary zoning district. Red indicates a decrease in enrollment while green indicates an increase. School Zone boundaries are shown in grey.

```
                                        change_df$District == 4)[,2] > 125))),
                ]) +
  facet_wrap(~ Land, ncol = 4) +
     geom_polygon(aes( x = long, y = lat, group = group),
                  fill = NA, color = "grey",
                  data = fortify(zones)) +
  coord_fixed(xlim = range(tracts_df$long[tracts_df$District == 4],
                           na.rm = TRUE),
              ylim = range(tracts_df$lat[tracts_df$District == 4],
                           na.rm = TRUE)) +
  theme(legend.position = "bottom",
        plot.title = element_text(hjust = 0,
                                  size = 9)) +
  labs(color = "Primary Zoning",
    title = "Figure 6. School District 20 Enrollment Change from 2001-2002 to 2010-2011\n by School Zone
  scale_fill_gradient2(low = "red", high = "green", midpoint = 0,
                       na.value = "white", guide = FALSE)
```

```
land_tracts$primary <- substr(land_tracts$ZONEDIST, 1, 2)
enroll_sum1 <- aggregate(Count.of.Students ~ Year + Cohort + NTACode + primary,
                  data = join_all(list(enroll_plot,
                                       tracts@data,
                                       land_tracts),
                                  by = "BoroCT2010"),
                  FUN = sum)
```

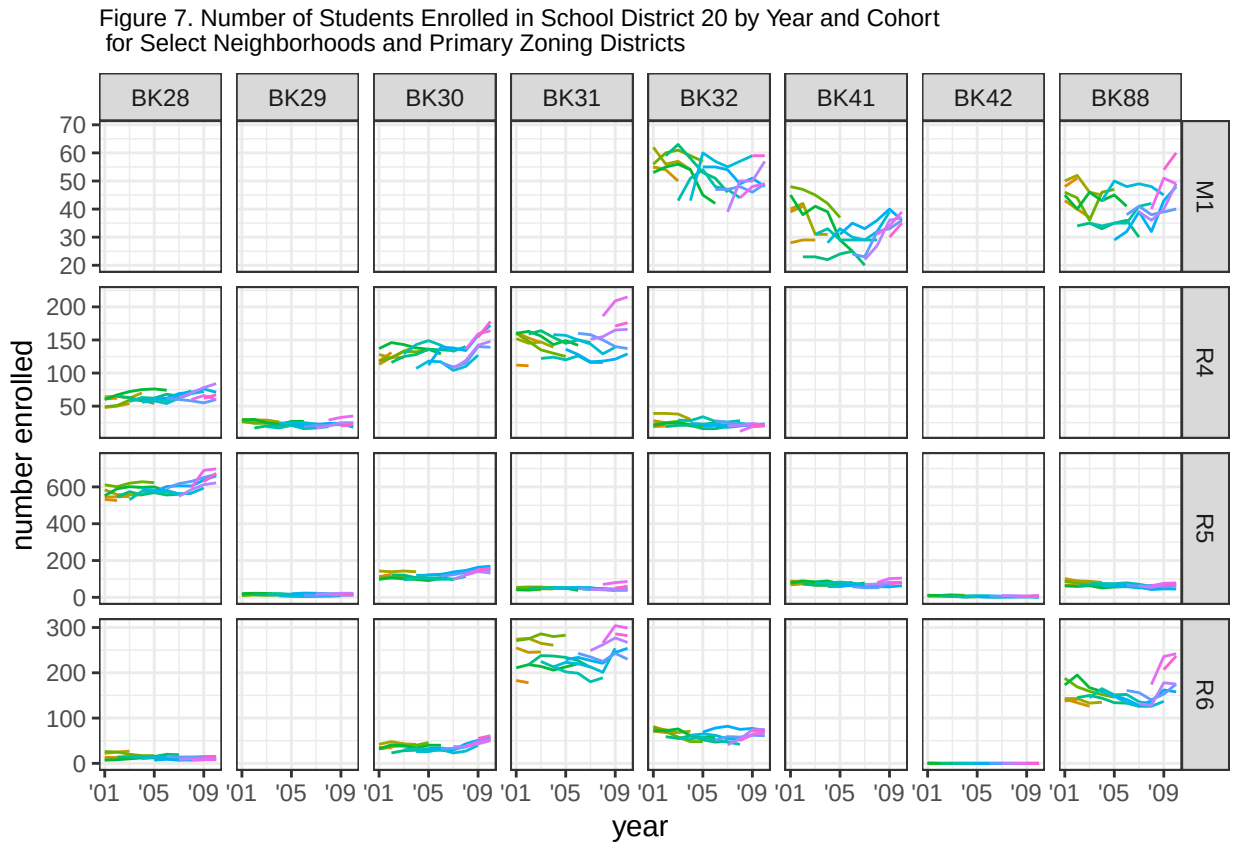

Figure 7: Aggregate student enrollment in School District 20 by year for select spatial boundaries and four Primary Zoning Districts. The colored lines indicate separate cohorts of students. The rows represent Primary Zoning Districts. The columns represent select Neighborhoods in School District 20.

```
ggplot(enroll_sum1[which(enroll_sum1$NTACode %in% c("BK28","BK29","BK30","BK31",
                                                    "BK32","BK41","BK42","BK88") &
                         enroll_sum1$primary %in% c("M1", "R4", "R5", "R6")),]) +
  theme_bw() +
  aes(2000 + Year, Count.of.Students, color = factor(Cohort)) +
  geom_line() +
  labs(y = "number enrolled", x = "year",
       title = "Figure 7. Number of Students Enrolled in School District 20 by Year and Cohort\n for Se
  scale_x_continuous(breaks = c(2001, 2005, 2009),
                     labels = c("'01", "'05", "'09")) +
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0,
                                  size = 9)) +
  facet_grid(primary ~ NTACode, scales = "free")
```

```
enroll_sum2 <- aggregate(Count.of.Students ~ Year + Cohort + id + primary,
                  data = join_all(list(enroll_plot,
                                       zones_tracts,
                                       land_tracts),
                                  by = "BoroCT2010"),
                  FUN = sum)
```

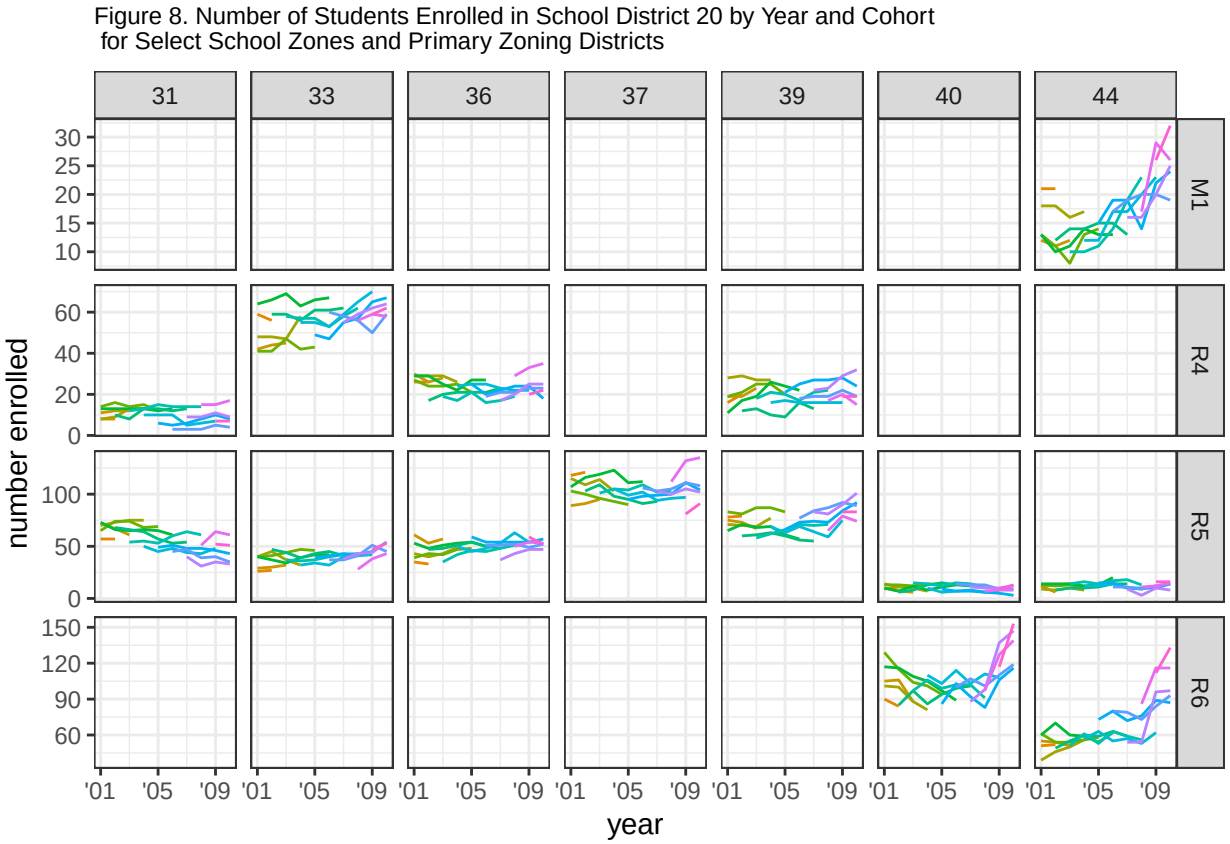

Figure 8: Aggregate student enrollment in School District 20 by year for select spatial boundaries and four Primary Zoning Districts. The colored lines within each plot indicate separate cohorts of students. The rows represent Primary Zoning Districts. The columns represent select School Zones in School District 20.

```
ggplot(enroll_sum2[which(enroll_sum2$id %in% c(31,33,36,37,39,40,44) &
                         enroll_sum2$primary %in% c("M1", "R4", "R5", "R6")),]) +
  theme_bw() +
  aes(2000 + Year, Count.of.Students, color = factor(Cohort)) +
  geom_line() +
  labs(y = "number enrolled", x = "year",
       title = "Figure 8. Number of Students Enrolled in School District 20 by Year and Cohort\n for Se
  scale_x_continuous(breaks = c(2001, 2005, 2009),
                     labels = c("'01", "'05", "'09")) +
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0,
                                  size = 9)) +
  facet_grid(primary ~ id, scales = "free")
```

3. Models

Let Y_{ij}^{n} denote the number of students enrolled in Census Tract $t \in \{1 \dots T\}$ for grade $i \in \{K \dots 5\}$ and school year $j \in \{2001-02 \dots 2010-11\}$. Enrollment is not considered inherently random. Instead, fluctuations are thought to be the result of unobserved covariates that vary both between grades and over time, and cause the outcome. The goal of our model is to partition the effect of these covariates into three groups. Time specific

period effects, P_i, age specific grade effects, G_j, and cohort specific effects C_{i-j}. We constrain $G_4 = G_5$.

We use the Hamiltonian Monte Carlo algorithm to sample from the posterior distribution induced by the following generative, log-linear model using Stan Development Team (2016). (Gelman et al. 2014) We run four chains for one thousand iterations each. The first half of the chains are discarded as warm-up:

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

The key assumption of this first model is that enrollment varies systematically about these effects. In our second model, we further subdivide these grade, period, and cohort effects that more plausibly resemble enrollment behavior, reflecting our discussion in Sections 1 and 2. These effects are pooled across similar school zones Z_i, neighborhoods, H_i, and land uses, L_i. For example, the period effect in two tracts of the same neighborhood are closer together on average than the period effect in the tracts of two separate neighborhoods.

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

Weakly informative gamma priors are put on the σ's as suggested by the
Stan Prior Choice Recommendation Wiki, although the posterior samples do not appear sensitive to this constraint. We note that low BFMI was reported for model 2 when sampling with the default maximum treedepth. We reran Stan with a higher maximum tree depth as suggested in the Stan Warning Guide. Posterior samples from both models were retained for the following Results section.

```
data_all <-
join_all(list(enrollment,
              tracts@data[,c("BoroCT2010","NTACode")],
              zones_tracts[,c("BoroCT2010", "id")],
              land_tracts[,c("BoroCT2010", "ZONEDIST")]),
         by = "BoroCT2010")
colnames(data_all)[7:9] <- c("NTA","Zone","Land")

data_all2 <- data.frame(
  count = data_all$Count.of.Students,
  CT2010 = data_all$CT2010,
  BoroCT2010 = data_all$BoroCT2010,
```

```
  year = sapply(data_all$School.Year, function(x) as.numeric(substr(x, 1, 4))),
  grade = as.numeric(data_all$Grade.Level),
  nta = data_all$NTA,
  zone = data_all$Zone,
  land = data_all$Land
)

data_all2$grade[data_all2$grade == 6] <- 0
data_all2$cohort <- data_all2$year - data_all2$grade
data_all2 <- data_all2[which(data_all2$BoroCT2010 %in%
                           as.character(unique(
                             na.omit(
                               tracts_df$BoroCT2010[
                                 tracts_df$District == 4])))),]

data_all2$grade[data_all2$grade == 5] <- 4
stan_data <- list(N = nrow(data_all2),
                  T = length(unique(data_all2$CT2010)),
                  G = length(unique(data_all2$grade)),
                  Y = length(unique(data_all2$year)),
                  C = length(unique(data_all2$cohort)),
                  num_students = data_all2$count,
                  tract = as.numeric(as.factor(as.numeric(data_all2$CT2010))),
                  grade = as.numeric(as.factor(data_all2$grade)),
                  year = as.numeric(as.factor(data_all2$year)),
                  cohort = as.numeric(as.factor(data_all2$cohort)))

stan_data$zone <- unique(cbind(as.numeric(as.factor(as.numeric(data_all2$zone))),
                               stan_data$tract))[order(unique(stan_data$tract)),1]
stan_data$Z <- length(unique(stan_data$zone))
stan_data$nbhd <- unique(cbind(as.numeric(as.factor(as.numeric(data_all2$nta))),
                               stan_data$tract))[order(unique(stan_data$tract)),1]
stan_data$H <- length(unique(stan_data$nbhd))
stan_data$land <- unique(cbind(as.numeric(as.factor(as.numeric(data_all2$land))),
                               stan_data$tract))[order(unique(stan_data$tract)),1]
stan_data$L <- length(unique(stan_data$land))
```

```
fit1 <- stan(file = "model1.stan",
             data = stan_data,
             iter = 1000,
             chains = 4)

fit2 <- stan(file = "model2.stan",
             data = stan_data,
             iter = 1000,
             chains = 4,
             control = list(max_treedepth = 15))

fit1_output <- extract(fit1)
fit2_output <- extract(fit2)
```

4. Results

In our enrollment projections, we wish to separate long-term demographic changes that are predictive of future trends from short term fluctuations that are unlikely to continue far into the future. After interpretation of the model parameters, we believe planners can use their expert knowledge to make these determinations. We offer a demonstration of how this might be done in the remainder of this section.

In School District 20, planners might reasonably consider fixed grade effects that do not fluctuate over time, short-term trends in the year effects, and long-term trends in the cohort effect. This determination was made using the following figures where it was found that, in general, year effects reflect economic bubbles, recessions, and transient government policies. These factors likely impact alternatives to public school such as charter, private, religious, or parochial enrollment. Cohort effects seem to reflect the types of students who reside in a Census Tract, including school specific factors.

For example, Figure 9 shows the posterior mean of the age, period, and cohort effects aggregated across Census Tracts. The estimates correspond with the observations made in Section 2.2: the decrease in enrollment from 2001 to 2007 is explained by decreasing cohort effects, while the increase in enrollment from 2008-09 to 2010-11 is explained by two atypical years (2009-10, 2010-11) and one atypical cohort effect (2010-11).

Figure 10 shows boxplots of the distribution of posterior means (within) and posterior variances (between) of each of the model parameters across tracts. The within portion of the Figure shows that cohort effects explain more variation than period or grade effects, while the between portion shows large uncertainty within some school zones and land uses.

Figures 11 and 12 recreate Figures 3 and 4 in order to compare the model fit with actual enrollment numbers. The fit is the exponentiated posterior mean, after summing across grade, year, and cohort effects. Figure 11 displays the difference between actual and fited values. No fit is more than 15 students from the actual enrollment, with the greatest error occuring in the largest Census Tracts. Figure 12 aggregates the fitted values over all Census Tracts by year and cohort. The mild regularization of the model parameters is observable both within and across cohorts.

Figures 13 and 14 map the posterior mean of the period effects spatially for the final school years: 2008-09, 2009-10, and 2010-11. Figure 13 maps the total effect, while Figure 14 stratifies by school zone, neighborhood and land use. These two Figures demonstrate that year effects are determined by factors correlated with neighborhood. Meanwhile, Figures 15 and 16 map the posterior mean of the cohort effects spatially for the 2008-2010 school years. These two Figures demonstrate that cohort effects are determine by factors correlated with school zones.

Figures 17 and 19 display example predictions for the first nine tracts of the dataset. We made these predictions as follows: we fit a simple linear regression to the cohort effects, an MA(1) (or GP in Figure 19) to the period effects and kept the grade effects constant. We note that fitting an MA(1) to the period effects is the same in expectation as holding the last period constant. The predictions are aggregated over Census Tracts by year and cohort in Figures 18 and 20.

```r
apc <- data.frame(stan_data[
  which(names(stan_data) %in% c('tract', 'grade', 'year', 'cohort', 'zone', 'nbhd', 'land'))])

mu <- colMeans(fit2_output$mu) * mean(fit2_output$sigma)
beta_grade_zone <- colMeans(fit2_output$tau_grade_zone) * colMeans(fit2_output$beta_grade_zone)
beta_grade_nbhd <- colMeans(fit2_output$tau_grade_nbhd) * colMeans(fit2_output$beta_grade_nbhd)
beta_grade_land <- colMeans(fit2_output$tau_grade_land) * colMeans(fit2_output$beta_grade_land)
beta_year_zone <-  colMeans(fit2_output$tau_year_zone) * colMeans(fit2_output$beta_year_zone)
beta_year_nbhd <-  colMeans(fit2_output$tau_year_nbhd) * colMeans(fit2_output$beta_year_nbhd)
beta_year_land <-  colMeans(fit2_output$tau_year_land) * colMeans(fit2_output$beta_year_land)
beta_cohort_zone <- colMeans(fit2_output$tau_cohort_zone) * colMeans(fit2_output$beta_cohort_zone)
beta_cohort_nbhd <- colMeans(fit2_output$tau_cohort_nbhd) * colMeans(fit2_output$beta_cohort_nbhd)
beta_cohort_land <- colMeans(fit2_output$tau_cohort_land) * colMeans(fit2_output$beta_cohort_land)

mu_df <- data.frame(stan_data[which(names(stan_data) %in% c('tract','nbhd', 'land', 'zone'))])
mu_df <- mu_df[!duplicated(mu_df), ]
mu_df$mu <- mu_df$ mu

apc$beta_grade_zone <- beta_grade_zone[cbind(apc$grade, apc$zone)]
apc$beta_grade_nbhd <- beta_grade_nbhd[cbind(apc$grade, apc$nbhd)]
apc$beta_grade_land <- beta_grade_land[cbind(apc$grade, apc$land)]
apc$beta_year_zone <-  beta_year_zone[cbind(apc$year, apc$zone)]
apc$beta_year_nbhd <-  beta_year_nbhd[cbind(apc$year, apc$nbhd)]
apc$beta_year_land <-  beta_year_land[cbind(apc$year, apc$land)]
apc$beta_cohort_zone <- beta_cohort_zone[cbind(apc$cohort, apc$zone)]
apc$beta_cohort_nbhd <- beta_cohort_nbhd[cbind(apc$cohort, apc$nbhd)]
apc$beta_cohort_land <- beta_cohort_land[cbind(apc$cohort, apc$land)]

apc_plot <- data.frame(variable = c(apc$grade, apc$year, apc$cohort),
                       label = c(rep("grade", length(apc$grade)),
                                 rep("year", length(apc$year)),
                                 rep("cohort", length(apc$cohort))),
                       value = c(apc$beta_grade_land +
                                   apc$beta_grade_nbhd +
                                   apc$beta_grade_zone,
                                 apc$beta_year_land +
                                   apc$beta_year_nbhd +
                                   apc$beta_year_zone,
                                apc$beta_cohort_land +
                                   apc$beta_cohort_nbhd +
                                   apc$beta_cohort_zone))

apc_plot$variable[apc_plot$label == "year"] <-
  2000 + apc_plot$variable[apc_plot$label == "year"]

ggplot() +
   theme_bw() +
   geom_line(aes(variable, exp(value)), color = "black",
             data = aggregate(value ~ variable + label, apc_plot, mean)) +
   facet_wrap(~label, drop = TRUE, scales = "free_x") +
   labs(x = "", y = expression(e^beta[.]),
        title = "Figure 9. Posterior Mean of Age, Period, and Cohort Effects") +
  theme(plot.title = element_text(hjust = 0, size = 9)) +
```

Figure 9. Posterior Mean of Age, Period, and Cohort Effects

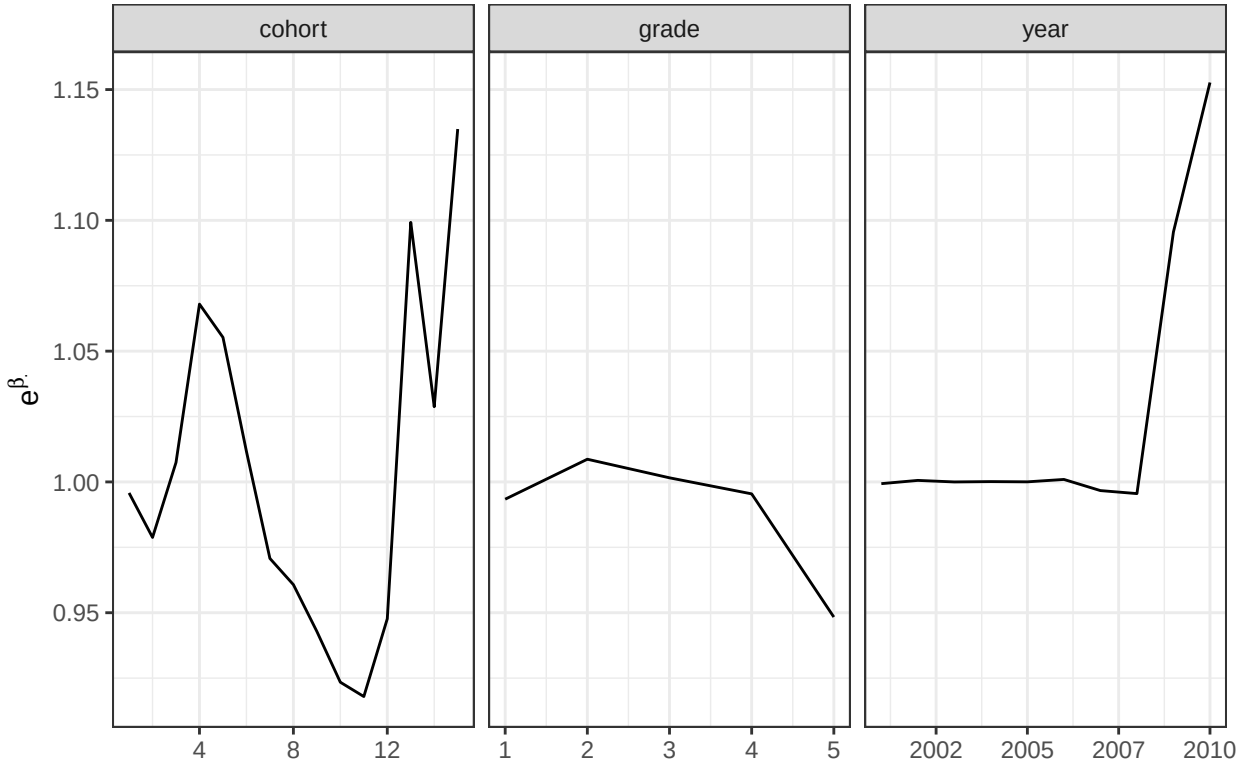

Figure 9: Posterior mean of the age (grade), period (year), and cohort effects, aggregated across Census Tracts. There are 15 cohort variables, 5 grade variables, and 10 year variables, with posterior means plotted on an exponential scale.

```
  scale_x_continuous(labels = function(x) floor(x))
```

```
beta_grade_zone <- colMeans(fit2_output$tau_grade_zone) *
  apply(fit2_output$beta_grade_zone, c(2,3), sd)
beta_grade_nbhd <- colMeans(fit2_output$tau_grade_nbhd) *
  apply(fit2_output$beta_grade_nbhd, c(2,3), sd)
beta_grade_land <- colMeans(fit2_output$tau_grade_land) *
  apply(fit2_output$beta_grade_land, c(2,3), sd)
beta_year_zone <- colMeans(fit2_output$tau_year_zone) *
  apply(fit2_output$beta_year_zone, c(2,3), sd)
beta_year_nbhd <- colMeans(fit2_output$tau_year_nbhd) *
  apply(fit2_output$beta_year_nbhd, c(2,3), sd)
beta_year_land <- colMeans(fit2_output$tau_year_land) *
  apply(fit2_output$beta_year_land, c(2,3), sd)
beta_cohort_zone <- colMeans(fit2_output$tau_cohort_zone) *
  apply(fit2_output$beta_cohort_zone, c(2,3), sd)
beta_cohort_nbhd <- colMeans(fit2_output$tau_cohort_nbhd) *
  apply(fit2_output$beta_cohort_nbhd, c(2,3), sd)
beta_cohort_land <- colMeans(fit2_output$tau_cohort_land) *
  apply(fit2_output$beta_cohort_land, c(2,3), sd)

apc$sd_grade_zone <- beta_grade_zone[cbind(apc$grade, apc$zone)]
apc$sd_grade_nbhd <- beta_grade_nbhd[cbind(apc$grade, apc$nbhd)]
```

```r
apc$sd_grade_land <- beta_grade_land[cbind(apc$grade, apc$land)]
apc$sd_year_zone <-  beta_year_zone[cbind(apc$year, apc$zone)]
apc$sd_year_nbhd <-  beta_year_nbhd[cbind(apc$year, apc$nbhd)]
apc$sd_year_land <-  beta_year_land[cbind(apc$year, apc$land)]
apc$sd_cohort_zone <- beta_cohort_zone[cbind(apc$cohort, apc$zone)]
apc$sd_cohort_nbhd <- beta_cohort_nbhd[cbind(apc$cohort, apc$nbhd)]
apc$sd_cohort_land <- beta_cohort_land[cbind(apc$cohort, apc$land)]

apc_long <- reshape2::melt(apc, id = colnames(apc)[1:7])
apc_long$label <- factor(ifelse(substr(apc_long$variable,1,2)=="sd","between","within"),
                         levels = c("within", "between"))
ggplot(apc_long[apc_long$variable!="mu",]) +
  theme_bw() +
  aes(variable, value) +
  geom_boxplot(outlier.size = 0) +
  facet_wrap(~label, drop = TRUE, nrow = 2, scales = "free") +
  coord_flip() +
  scale_y_continuous(name = "") +
  scale_x_discrete(name = "", labels = c("beta_grade_zone" = expression(bar(beta)[g*","*z]),
                                         "beta_grade_nbhd" = expression(bar(beta)[g*","*h]),
                                         "beta_grade_land" = expression(bar(beta)[g*","*l]),
                                         "beta_year_zone" = expression(bar(beta)[y*","*z]),
                                         "beta_year_nbhd" = expression(bar(beta)[y*","*h]),
                                         "beta_year_land" = expression(bar(beta)[y*","*l]),
                                         "beta_cohort_zone" = expression(bar(beta)[c*","*z]),
                                         "beta_cohort_nbhd" = expression(bar(beta)[c*","*h]),
                                         "beta_cohort_land" = expression(bar(beta)[c*","*l]),
                                         "sd_grade_zone" = expression(sigma(beta)[g*","*z]),
                                         "sd_grade_nbhd" = expression(sigma(beta)[g*","*h]),
                                         "sd_grade_land" = expression(sigma(beta)[g*","*l]),
                                         "sd_year_zone" = expression(sigma(beta)[y*","*z]),
                                         "sd_year_nbhd" = expression(sigma(beta)[y*","*h]),
                                         "sd_year_land" = expression(sigma(beta)[y*","*l]),
                                         "sd_cohort_zone" = expression(sigma(beta)[c*","*z]),
                                         "sd_cohort_nbhd" = expression(sigma(beta)[c*","*h]),
                                         "sd_cohort_land" = expression(sigma(beta)[c*","*l]))) +
  labs(title = "Figure 10. Boxplot Depicting Within and Between Variation of Age, Period, and Cohort Ef
  theme(plot.title = element_text(hjust = 0, size = 9))
```

Figure 10. Boxplot Depicting Within and Between Variation of Age, Period, and Cohort Effect Posteriors

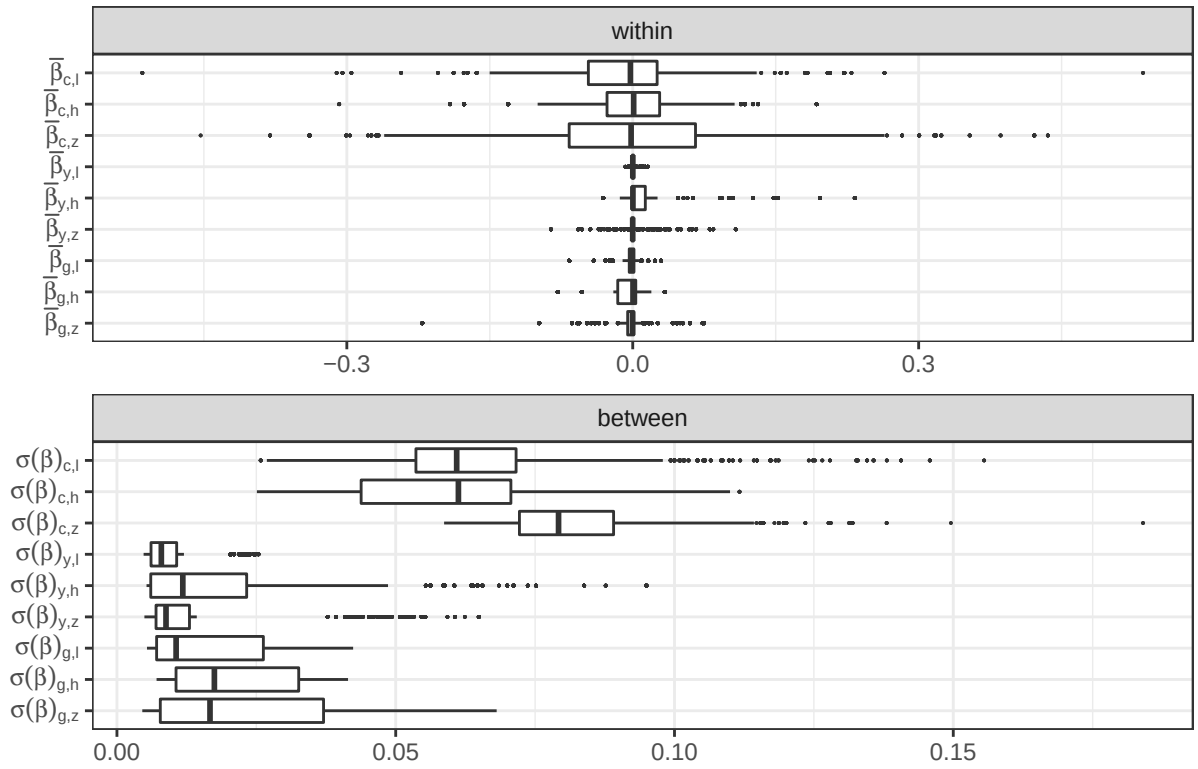

Figure 10: Posterior mean of the age (grade), period (year), and cohort effects, aggregated across Census Tracts. Boxplots show Within and Between variation of Age, Period, and Cohort posterior effects for different spatial partitions. The letters indexing the effects correspond to the following: c: cohort, y: year (period), g: grade (age), l: land use zoning classification, h: neighborhood, z: school zone.

```r
df_pred <- with(stan_data, data.frame(num_students,
                                      tract,
                                      grade,
                                      cohort,
                                      year))

df_pred$mu <- colMeans(fit1_output$mu)[df_pred$tract]
df_pred$alpha_grade <-
  colMeans(fit1_output$alpha_grade)[
    matrix(c(df_pred$grade, df_pred$tract),ncol=2)]
df_pred$alpha_year <-
  colMeans(fit1_output$alpha_year)[
    matrix(c(df_pred$year, df_pred$tract),ncol=2)]
df_pred$alpha_cohort <-
  colMeans(fit1_output$alpha_cohort)[
    matrix(c(df_pred$cohort, df_pred$tract),ncol=2)]

df_pred$sigma <-  mean(fit1_output$sigma)
df_pred$tau_grade <- colMeans(fit1_output$tau_grade)[df_pred$grade]
df_pred$tau_year <-colMeans(fit1_output$tau_year)[df_pred$year]
df_pred$tau_cohort <- colMeans(fit1_output$tau_cohort)[df_pred$cohort]

tract_pred <- function(i, df_pred){
  x <- df_pred[i,]
  x['sigma'] * x['mu'] +
    x['tau_year'] * x['alpha_year'] +
    x['tau_grade'] * x['alpha_grade'] +
    x['tau_cohort'] * x['alpha_cohort']
}

df_pred$pred <-  unlist(sapply(1:nrow(df_pred), FUN = tract_pred, df_pred))

data_for_joining <- data_all2
data_for_joining$Year <- data_for_joining$year - 2000
data_for_joining$Grade <- data_for_joining$grade + 1
data_for_joining$Cohort <- data_for_joining$Year - data_for_joining$Grade + 6
data_for_joining$Tract <- as.numeric(as.factor(as.numeric(data_for_joining$CT2010)))
colnames(df_pred)[2:5] <- c("Tract", "Grade", "Cohort", "Year")

enrollment_apc <- left_join(data_for_joining, df_pred, by = c("Year", "Grade", "Cohort", "Tract"))
enrollment_apc_plot <- left_join(tracts_df, enrollment_apc, by = c("CT2010", "BoroCT2010"))
enrollment_apc_plot1 <- enrollment_apc_plot
enrollment_apc_plot1$label <- ifelse(is.na(enrollment_apc_plot1$year), NA,
                                     paste(enrollment_apc_plot1$year,
                                     substr(enrollment_apc_plot1$year + 1,3,4),
                                     sep = "-"))
enrollment_apc_plot1$label_grade <- relevel(factor(
  ifelse(enrollment_apc_plot1$grade == 0, "K",
         enrollment_apc_plot1$grade)), "K")

ggplot(enrollment_apc_plot1[enrollment_apc_plot1$year %in% 2006:2011, ]) +
  theme_void() +
  geom_polygon(aes(x = long, y = lat, group = group,
```

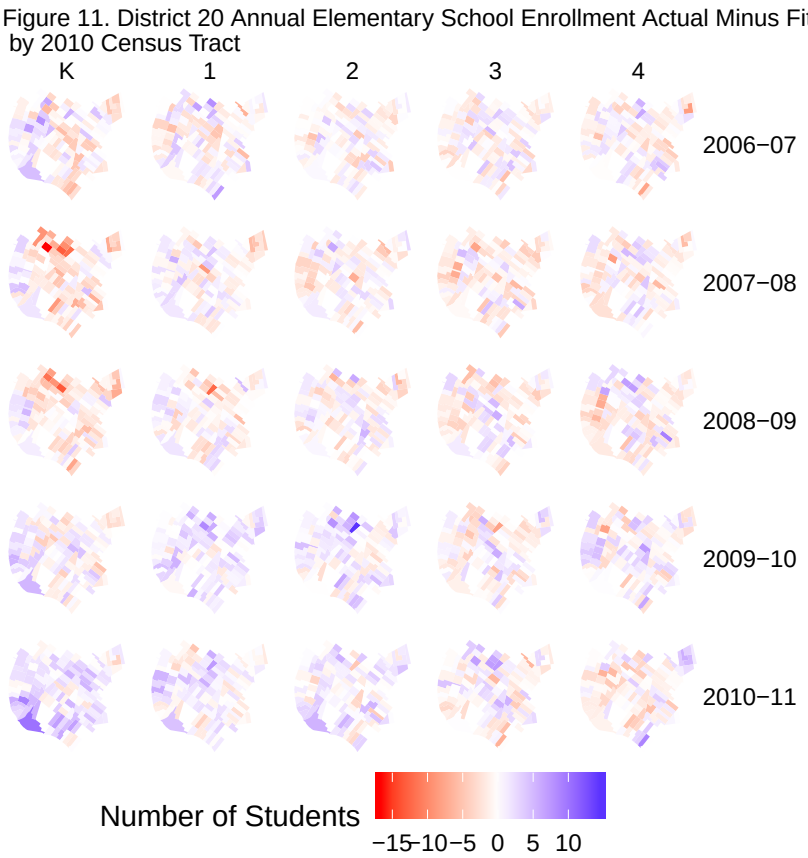

Figure 11: Difference in enrollment counts between predictions from our model and true enrollment counts in School District 20 by 2010 Census Tract. The columns represent grades, ranging from K through 4. The rows represent years, ranging from 2006-2007 to 2010-2011. The diagonals correspond to cohorts.

```
                  fill = num_students - exp(pred))) +
  facet_grid(label ~ label_grade) +
  coord_equal() +
  labs(title = "Figure 11. District 20 Annual Elementary School Enrollment Actual Minus Fitted Values \n
      scale_fill_gradient2("Number of Students",
                           low = "red", high = "blue", midpoint = 0,
                       na.value = "white") +
  theme(plot.title = element_text(hjust = 0, size = 9),
        legend.position = "bottom")
```

```
ggplot() +
  theme_bw() +
  geom_line(aes(Year + 2000, num_students, color = factor(Cohort)),
    data = aggregate(num_students ~ Year + Cohort, df_pred, sum )) +
  geom_line(aes(Year + 2000, `exp(pred)`, color = factor(Cohort)),
            linetype = 2,
            data = aggregate(exp(pred) ~ Year + Cohort, df_pred, sum )) +
    labs(y = "number enrolled", x = "year",
       title = "Figure 12. Number of Students Enrolled in School District 20\n by Year and Cohort with
  scale_x_continuous(breaks = c(2001, 2005, 2009)) +
  theme(plot.title = element_text(hjust = 0, size = 9),
```

Figure 12. Number of Students Enrolled in School District 20
by Year and Cohort with Actual (solid line) and Fitted Values from Posterior Mean (broken line)

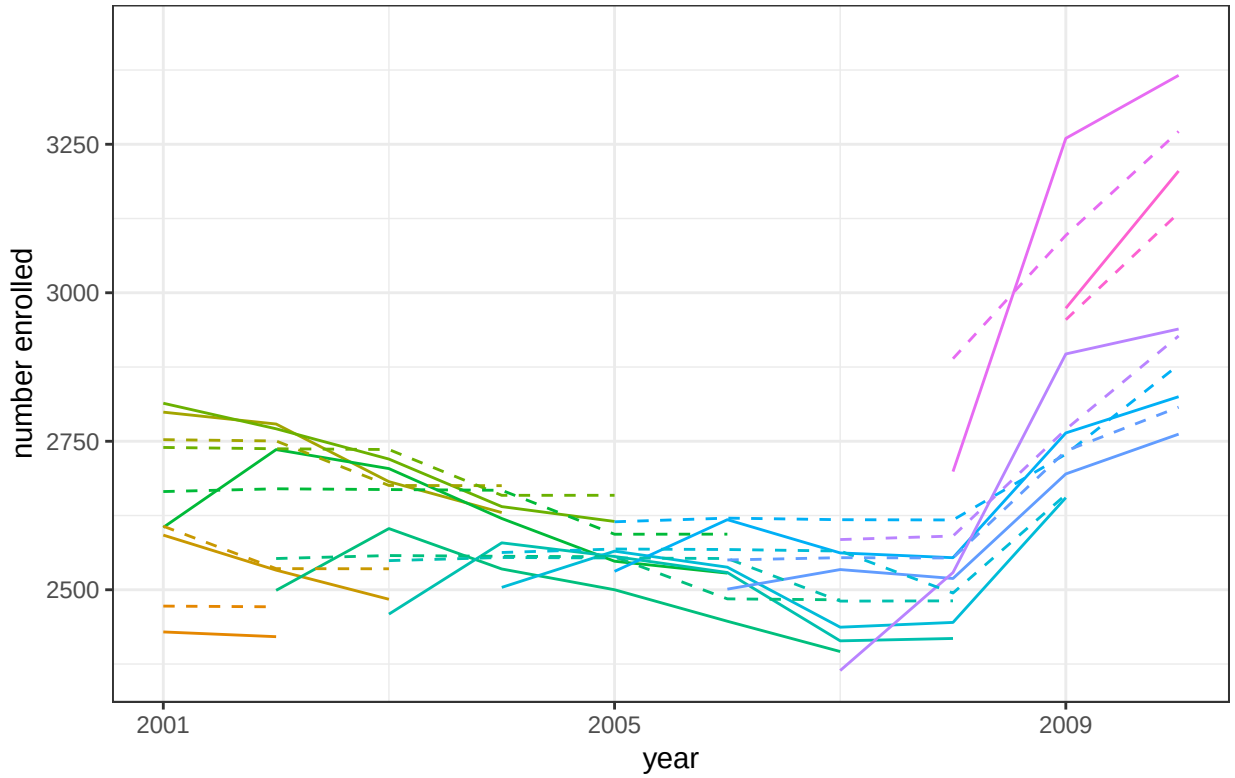

Figure 12: Aggregate student enrollment in School District 20 by year. Each color indicates a separate cohort of students. The solid lines correspond to true enrollment counts. The dashed lines correspond to enrollment estimates using the posterior mean from our model.

```
        legend.position = "none")
```

```
ggplot(enrollment_apc_plot1[enrollment_apc_plot1$year %in% 2008:2010, ]) +
  theme_void() +
  geom_polygon(aes(x = long, y = lat, group = group,
                   fill = alpha_year)) +
  facet_wrap(~ label, ncol = 3) +
  coord_equal() +
  scale_fill_gradient2(low = "red", high = "blue", midpoint = 0,
                       na.value = "white", guide = FALSE) +
  labs(title = "Figure 13. Select Period Effects by Census Tract") +
  theme(plot.title = element_text(hjust = 0, size = 9))
```

```
enrollment_apc <- left_join(data_for_joining, apc,
                            by = c("Year" = "year", "Grade" = "grade",
                                   "Cohort" = "cohort", "Tract" = "tract"))
enrollment_apc_plot <- left_join(tracts_df, enrollment_apc, by = c("CT2010", "BoroCT2010"))

enrollment_apc_plot2 <-
melt(enrollment_apc_plot[enrollment_apc_plot$year %in% 2008:2010,
                     c("long", "lat", "group", "year", "beta_year_zone",
                       "beta_year_nbhd", "beta_year_land")],
```

Figure 13. Select Period Effects by Census Tract

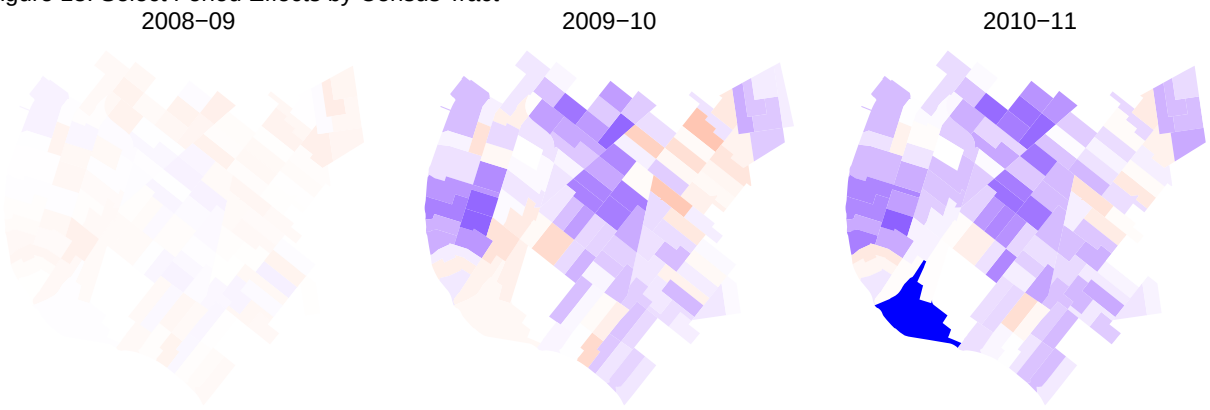

Figure 13: Posterior mean of year (period) effects for the school years 2008-2009 to 2010-2011 by Census Tract. Red values indicate negative effects while purples indicate positive effects.

```
    id.vars = c("long", "lat", "group", "year"))
enrollment_apc_plot2$variable <- factor(enrollment_apc_plot2$variable,
       labels = c('beta[y~","~z]', 'beta[y~","~h]', 'beta[y~","~l]'))
enrollment_apc_plot2$label <- paste0("'", ifelse(is.na(enrollment_apc_plot2$year), NA,
                                     paste(enrollment_apc_plot2$year,
                                     substr(enrollment_apc_plot2$year + 1,3,4),
                                     sep = "-")), "'")

ggplot(enrollment_apc_plot2) +
  theme_void() +
  geom_polygon(aes(x = long, y = lat, group = group,
                   fill = value)) +
  facet_grid(variable ~ label, labeller = label_parsed) +
  coord_equal() +
  scale_fill_gradient2(low = "red", high = "blue", midpoint = 0,
                       na.value = "white", guide = FALSE) +
  labs(title = "Figure 14. Select Period Effects by\n School Zone, Neighborhood, Land Use, and Census T
  theme(plot.title = element_text(hjust = 0, size = 9))
```

```
ggplot(enrollment_apc_plot1[enrollment_apc_plot1$cohort %in% 2008:2010, ]) +
  theme_void() +
  geom_polygon(aes(x = long, y = lat, group = group,
                   fill = alpha_cohort)) +
  facet_wrap(~ label, ncol = 3) +
  coord_equal() +
  scale_fill_gradient2(low = "red", high = "blue", midpoint = 0,
                       na.value = "white", guide = FALSE) +
  labs(title = "Figure 15. Select Cohort Effects by Census Tract") +
  theme(plot.title = element_text(hjust = 0, size = 9))
```

```
enrollment_apc_plot3 <-
  melt(enrollment_apc_plot[enrollment_apc_plot$cohort %in% 2008:2010,
                    c("long", "lat", "group", "year", "beta_cohort_zone",
                      "beta_cohort_nbhd", "beta_cohort_land")],
    id.vars = c("long", "lat", "group", "year"))
enrollment_apc_plot3$variable <- factor(enrollment_apc_plot3$variable,
```

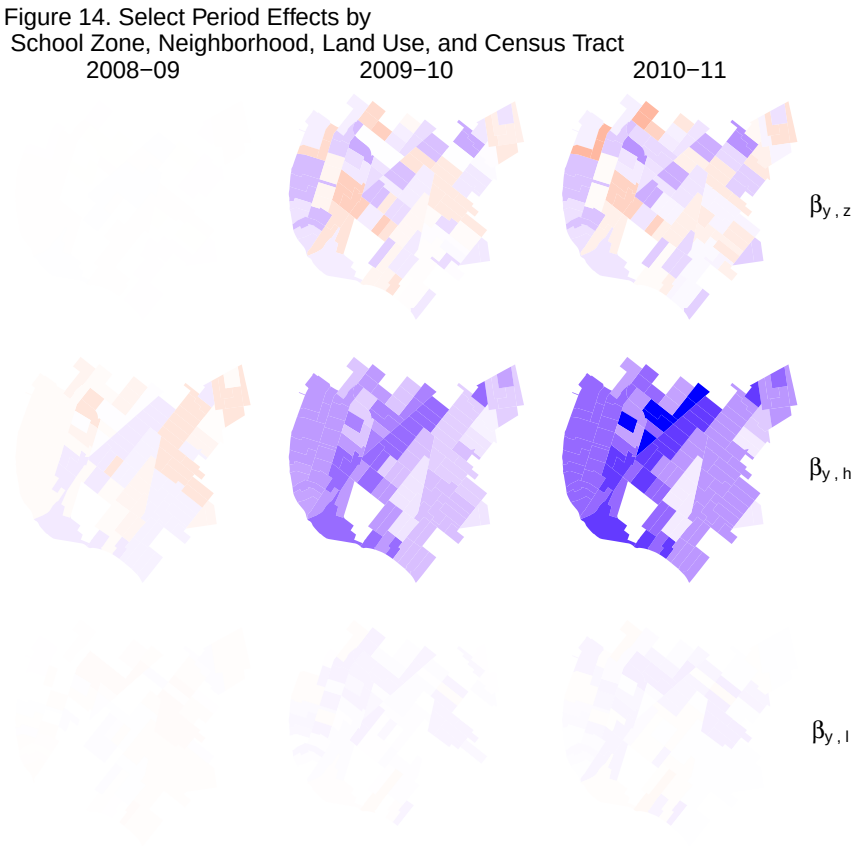

Figure 14: Posterior mean of year (period) effects by spatial partition for the school years 2008-2009 to 2010-2011. Each row corresponds to a separate spatial boundary: School Zone (top), Neighborhood (middle), Land Use Zoning (bottom). Red values indicate negative effects while purples indicate positive effects.

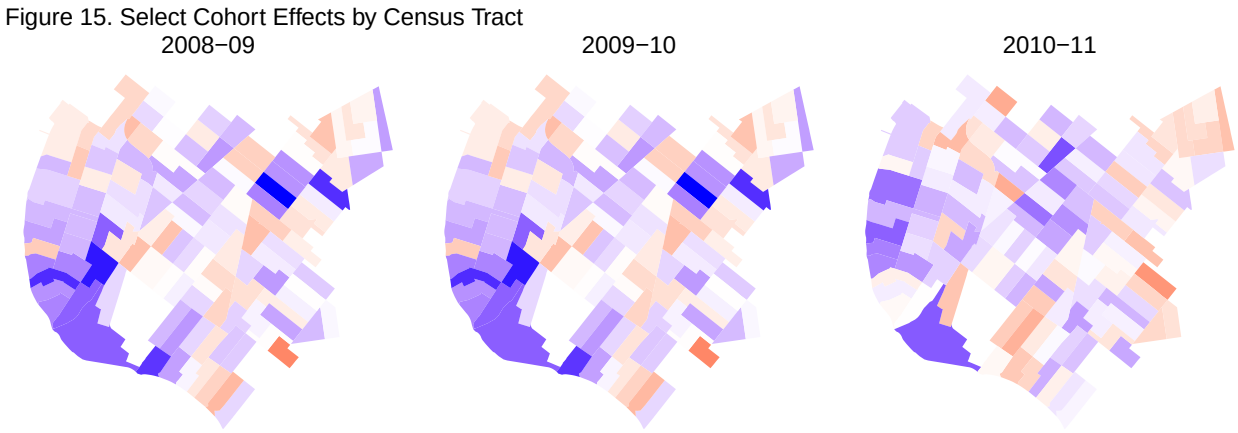

Figure 15: Posterior mean of cohort effects for the years 2008-2009 to 2010-2011 by Census Tract. Red values indicate negative effects while purples indicate positive effects.

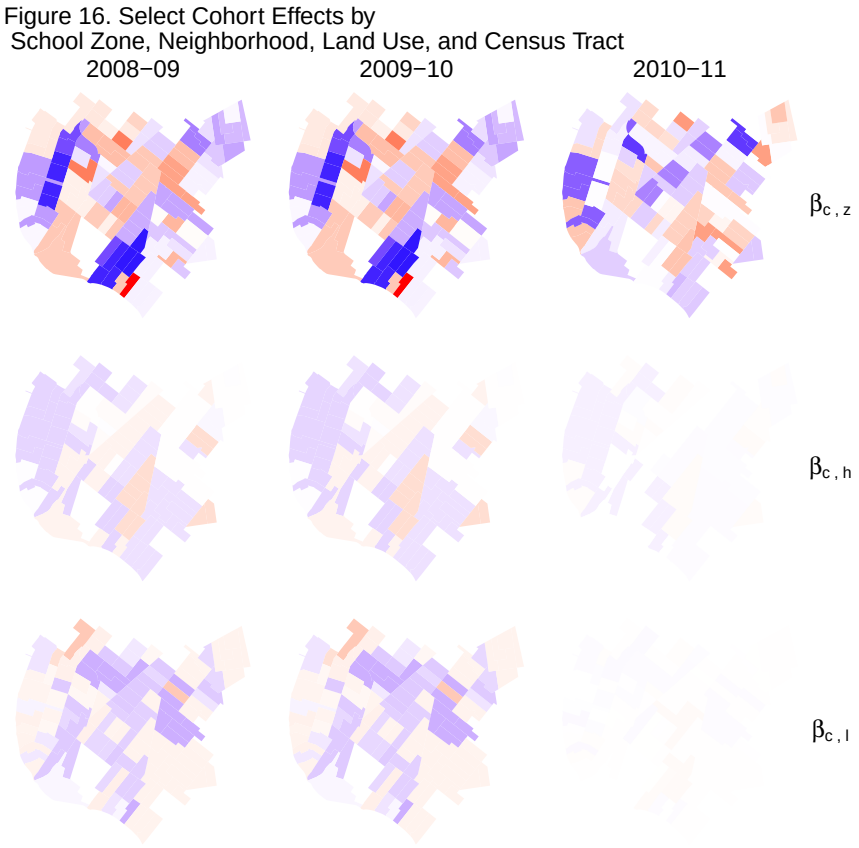

Figure 16: Posterior mean of cohort effects by spatial partition for the school years 2008-2009 to 2010-2011. Each row corresponds to a separate spatial boundary: School Zone (top), Neighborhood (middle), Land Use Zoning (bottom). Red values indicate negative effects while purples indicate positive effects.

```
      labels = c('beta[c~","~z]', 'beta[c~","~h]', 'beta[c~","~l]'))
enrollment_apc_plot3$label <- paste0("'",
                                     ifelse(is.na(enrollment_apc_plot3$year), NA,
                                     paste(enrollment_apc_plot3$year,
                                    substr(enrollment_apc_plot3$year + 1,3,4),
                                    sep = "-")),"'")
ggplot(enrollment_apc_plot3) +
  theme_void() +
  geom_polygon(aes(x = long, y = lat, group = group,
                   fill = value)) +
  facet_grid(variable ~ label, labeller = label_parsed) +
  coord_equal() +
  scale_fill_gradient2(low = "red", high = "blue", midpoint = 0,
                       na.value = "white", guide = FALSE) +
  labs(title = "Figure 16. Select Cohort Effects by\n School Zone, Neighborhood, Land Use, and Census T
  theme(plot.title = element_text(hjust = 0, size = 9))
```

```
colnames(df_pred)[2:5] <- c("tract", "grade", "cohort", "year")
tract_pred <- function(t){
  x <- df_pred[df_pred$tract == t,]
  df_new <- expand.grid(grade = c(1:5,5), year = 11:20)
```

```
  df_new$cohort <- df_new$year - df_new$grade + 5
  df_new <-
  merge(df_new, data.frame(cohort = 11:24,
                           alpha_cohort =
                             predict(lm(tau_cohort * alpha_cohort ~ cohort,
                                        data = x),
                                     newdata = data.frame(cohort = 11:24))))
  df_new$alpha_year <- x$tau_year[x$year == 10][1] * x$alpha_year[x$year == 10][1]
  df_new <- merge(df_new,
                  data.frame(grade = x$grade[!duplicated(x$grade)],
                             alpha_grade = (x$tau_grade * x$alpha_grade)[!duplicated(x$grade)]))
  df_new$mu <- x$sigma[1] * x$mu[1]
  df_new$pred <- exp(df_new$mu +
                       df_new$alpha_cohort +
                       df_new$alpha_year +
                       df_new$alpha_grade)

  df_old <- aggregate(cbind(num_students, exp(tau_grade * alpha_grade)) ~ cohort,
                      x[x$year == 10, ], sum)
  colnames(df_old)[2] <- "last_students"
  df_old <- merge(aggregate(pred ~ cohort + year, df_new, sum), df_old,
                  all.x = TRUE)
  df_old$pred_adj <- ifelse(df_old$pred < df_old$last_students &
                            !is.na(df_old$last_students),
                            df_old$V2 * df_old$last_students,
                            df_old$pred )

    c(aggregate(num_students ~ year, x, sum)[,2],
    aggregate(pred_adj ~ year, df_old, sum)[, 2])
}

pred <- data.frame(sapply(1:9, tract_pred))
colnames(pred) <- paste("Census Tract", 1:9)
pred <- melt(pred)
pred$year <- rep(2001:2020, 9)

ggplot(pred) +
  theme_bw() +
  aes(year, value) +
  geom_line() +
  facet_wrap(~ variable, scales = "free") +
  geom_vline(xintercept = 2010, linetype = 2) +
  labs(title = "Figure 17. Prediction of Student Enrollment for First Nine Census Tracts \n using Linea
       y = "number enrolled") +
  theme(plot.title = element_text(hjust = 0, size = 9))
```

```
tract_pred <- function(t){
  x <- df_pred[df_pred$tract == t,]
  df_new <- expand.grid(grade = c(1:6), year = 11:20)
  df_new$cohort <- df_new$year - df_new$grade + 5
  df_new$grade[df_new$grade == 6] <- 5
  df_new <-
    merge(df_new, data.frame(cohort = 11:24,
                             alpha_cohort =
```

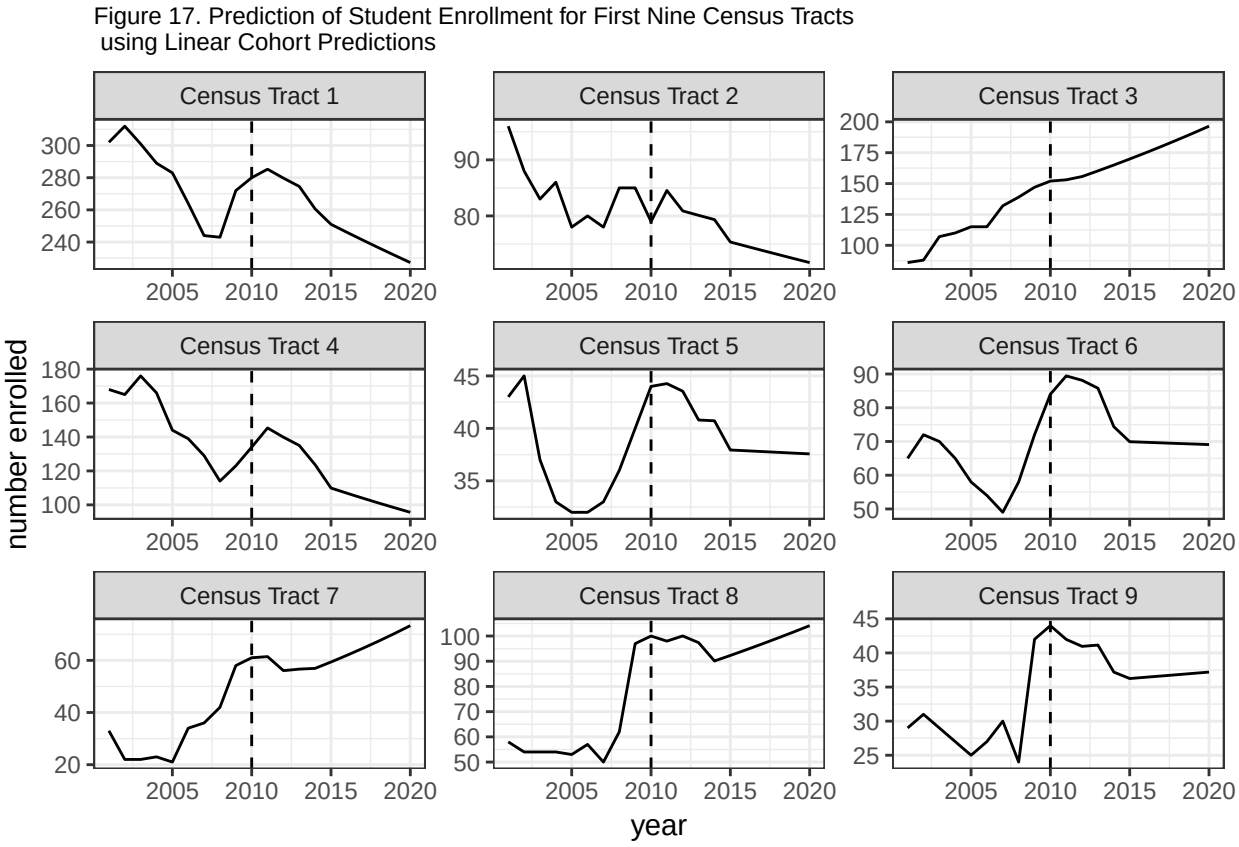

Figure 17: Student enrollment predictions for the first nine census tracts using linear cohort predictions. True enrollment counts are shown to the left of the vertical dashed line and model predictions are projected to the right of the line.

```
                                predict(lm(tau_cohort * alpha_cohort ~ cohort,
                                           data = x),
                                        newdata = data.frame(cohort = 11:24))))
  df_new$alpha_year <- x$tau_year[x$year == 10][1] * x$alpha_year[x$year == 10][1]
  df_new <- merge(df_new,
                  data.frame(grade = x$grade[!duplicated(x$grade)],
                             alpha_grade = (x$tau_grade * x$alpha_grade)[!duplicated(x$grade)]))
  df_new$mu <- x$sigma[1] * x$mu[1]
  df_new$pred <- exp(df_new$mu +
                       df_new$alpha_cohort +
                       df_new$alpha_year +
                       df_new$alpha_grade)

  df_old <- aggregate(cbind(num_students, exp(tau_grade * alpha_grade)) ~ cohort,
                              x[x$year == 10, ], sum)
  colnames(df_old)[2] <- "last_students"
  df_old <- merge(aggregate(pred ~ cohort + year, df_new, sum), df_old,
                  all.x = TRUE)
  df_old$num_students <- ifelse(df_old$pred < df_old$last_students &
                                  !is.na(df_old$last_students),
                                df_old$V2 * df_old$last_students,
                                df_old$pred )
  df_old <-
  rbind(aggregate(num_students ~ cohort + year, x, sum),
        aggregate(num_students ~ cohort + year, df_old, sum))
  df_old$tract <- t
  df_old
}

pred <- lapply(1:stan_data$T, tract_pred)
pred <- Reduce(rbind, pred)

ggplot(aggregate(num_students ~ cohort + year, pred, sum)) +
  theme_bw() +
  aes(2000 + year, num_students, color = factor(cohort)) +
  geom_vline(xintercept = 2010, linetype = 2) +
  geom_line() +
  labs(y = "number enrolled", x = "year",
       title = "Figure 18. Number of Students Enrolled in School District 20\n by Year and Cohort using
  scale_x_continuous(limits = c(2001, 2017),
                     breaks = c(2001, 2005, 2009, 2013, 2017, 2021)) +
  theme(plot.title = element_text(hjust = 0, size = 9),
        legend.position = "none")
```

```
library("mgcv")
tract_pred <- function(t){
  x <- df_pred[df_pred$tract == t,]
  df_new <- expand.grid(grade = c(1:5,5), year = 11:20)
  df_new$cohort <- df_new$year - df_new$grade + 5
  df_new <-
  merge(df_new, data.frame(cohort = sort(unique(df_new$cohort)),
                           alpha_cohort =
                             predict(lm(tau_cohort * alpha_cohort ~ cohort,
                                        data = x),
```

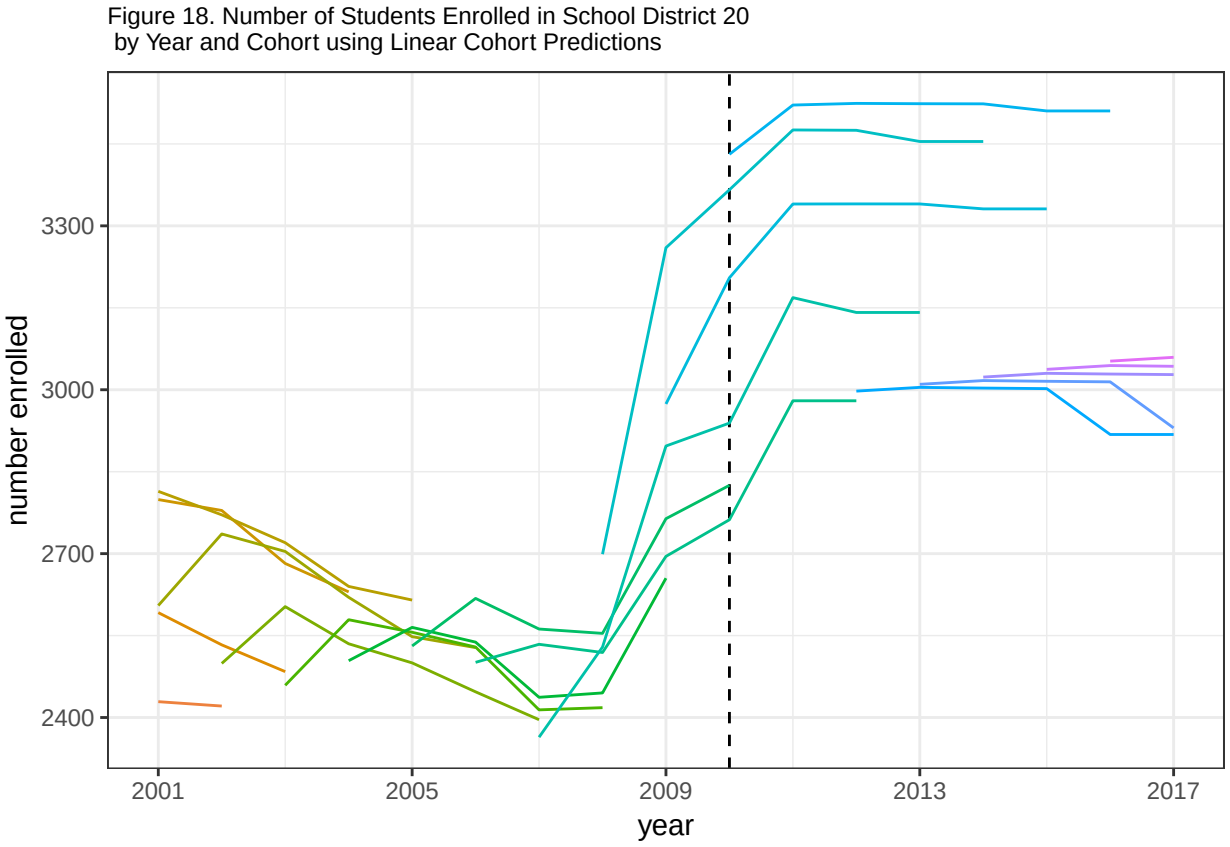

Figure 18: Aggregate student enrollment predictions in School District 20 by year using linear cohort predictions. Each colored line indicates a separate cohort of students. True aggregate counts are shown to the left of the vertical dashed line and model predictions are projected to the right of the line.

```
                                    newdata = data.frame(cohort =
                                                           sort(unique(df_new$cohort))))))
  df_new <-
  merge(df_new, data.frame(year = sort(unique(df_new$year)),
                           alpha_year =
                             predict(gam(tau_year * alpha_year ~ s(year, bs = "gp"),
                                         data = x),
                                     newdata = data.frame(year =
                                                           sort(unique(df_new$year))))))
  df_new <- merge(df_new,
                  data.frame(grade = x$grade[!duplicated(x$grade)],
                             alpha_grade = (x$tau_grade * x$alpha_grade)[!duplicated(x$grade)]))
  df_new$mu <- x$sigma[1] * x$mu[1]
  df_new$pred <- exp(df_new$mu + df_new$alpha_cohort + df_new$alpha_year + df_new$alpha_grade)

  df_old <- aggregate(cbind(num_students, exp(tau_grade * alpha_grade)) ~ cohort,
                      x[x$year == 10, ], sum)
  colnames(df_old)[2] <- "last_students"
  df_old <- merge(aggregate(pred ~ cohort + year, df_new, sum), df_old,
                  all.x = TRUE)
  df_old$pred_adj <- ifelse(df_old$pred < df_old$last_students &
                            !is.na(df_old$last_students),
                            df_old$V2 * df_old$last_students,
                            df_old$pred )

    c(aggregate(num_students ~ year, x, sum)[,2],
    aggregate(pred_adj ~ year, df_old, sum)[, 2])
}

pred <- data.frame(sapply(1:9, tract_pred))
colnames(pred) <- paste("Census Tract", 1:9)
pred <- reshape2::melt(pred)
pred$year <- rep(2001:2020, 9)

ggplot(pred) +
  theme_bw() +
  aes(year, value) +
  geom_line() +
  facet_wrap(~ variable, scales = "free") +
  geom_vline(xintercept = 2010, linetype = 2) +
  labs(title = "Figure 19. Prediction of Student Enrollment for First Nine Census Tracts \n using Linea
       y = "number enrolled") +
  theme(plot.title = element_text(hjust = 0, size = 9))
```

```
tract_pred <- function(t){
  x <- df_pred[df_pred$tract == t,]
  df_new <- expand.grid(grade = c(1:6), year = 11:20)
  df_new$cohort <- df_new$year - df_new$grade + 5
  df_new$grade[df_new$grade == 6] <- 5
  df_new <-
    merge(df_new, data.frame(cohort = sort(unique(df_new$cohort)),
                             alpha_cohort =
                               predict(lm(tau_cohort * alpha_cohort ~ cohort,
                                          data = x),
```

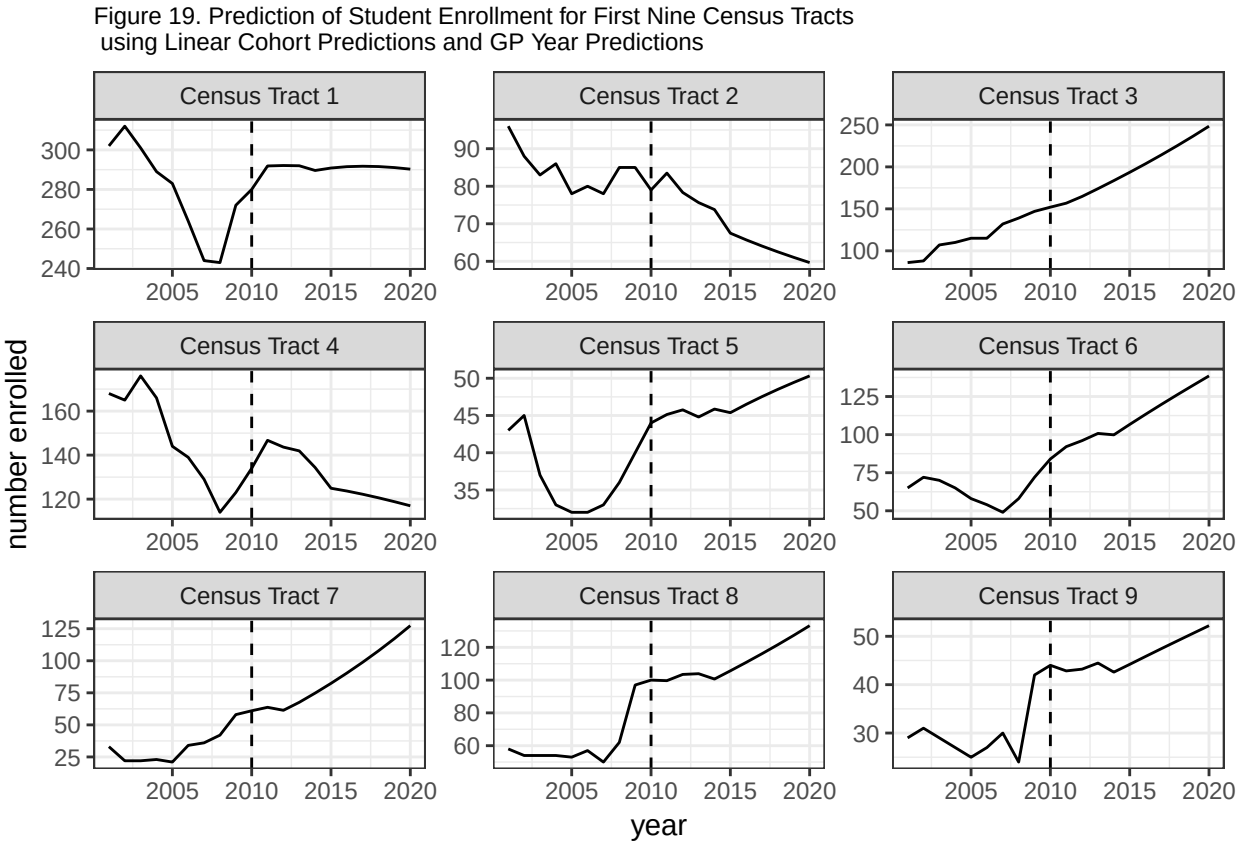

Figure 19: Student enrollment predictions for the first nine census tracts using linear cohort predictions and GP year predictions. True enrollment counts are shown to the left of the vertical dashed line and model predictions are projected to the right of the line.

```
                                        newdata = data.frame(cohort =
                                                               sort(unique(df_new$cohort))))))
  df_new <-
    merge(df_new, data.frame(year = sort(unique(df_new$year)),
                             alpha_year =
                               predict(gam(tau_year * alpha_year ~ s(year, bs = "gp"),
                                           data = x),
                                       newdata = data.frame(year =
                                                              sort(unique(df_new$year))))))

  df_new <- merge(df_new,
                  data.frame(grade = x$grade[!duplicated(x$grade)],
                             alpha_grade = (x$tau_grade * x$alpha_grade)[!duplicated(x$grade)]))
  df_new$mu <- x$sigma[1] * x$mu[1]
  df_new$pred <- exp(df_new$mu +
                       df_new$alpha_cohort +
                       df_new$alpha_year +
                       df_new$alpha_grade)

  df_old <- aggregate(cbind(num_students, exp(tau_grade * alpha_grade)) ~ cohort,
                      x[x$year == 10, ], sum)
  colnames(df_old)[2] <- "last_students"
  df_old <- merge(aggregate(pred ~ cohort + year, df_new, sum), df_old,
                  all.x = TRUE)
  df_old$num_students <- ifelse(df_old$pred < df_old$last_students &
                                  !is.na(df_old$last_students),
                                df_old$V2 * df_old$last_students,
                                df_old$pred)
  df_old <-
    rbind(aggregate(num_students ~ cohort + year, x, sum),
          aggregate(num_students ~ cohort + year, df_old, sum))
  df_old$tract <- t
  df_old
}

pred <- lapply(1:stan_data$T, tract_pred)
pred <- Reduce(rbind, pred)

ggplot(aggregate(num_students ~ cohort + year, pred, sum)) +
  theme_bw() +
  aes(2000 + year, num_students, color = factor(cohort)) +
  geom_vline(xintercept = 2010, linetype = 2) +
  geom_line() +
  labs(y = "number enrolled", x = "year",
       title = "Figure 20. Number of Students Enrolled in School District 20\n by Year and Cohort using
  scale_x_continuous(limits = c(2001, 2017),
                     breaks = c(2001, 2005, 2009, 2013, 2017)) +
  theme(plot.title = element_text(hjust = 0, size = 9),
        legend.position = "none") +
  ylim(2500, 4500)
```

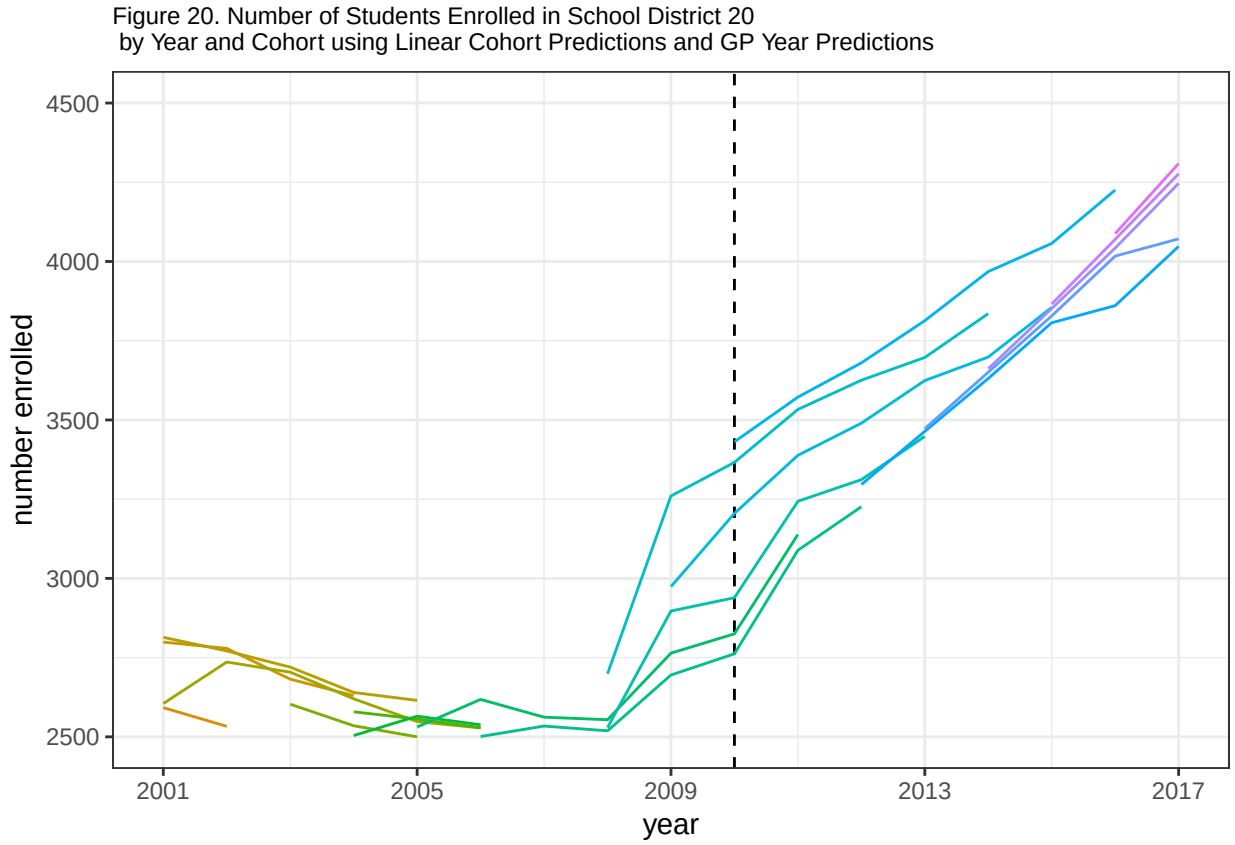

Figure 20: Aggregate student enrollment predictions in School District 20 by year using linear cohort predictions and GP year predictions. Each colored line indicates a separate cohort of students. True aggregate counts are shown to the left of the vertical dashed line and model predictions are projected to the right of the line.

5. References

Bell, Andrew, and Kelvyn Jones. 2018. "The Hierarchical Age–period–cohort Model: Why Does It Find the Results That It Finds?" *Quality & Quantity* 52 (2): 783–99. doi:10.1007/s11135-017-0488-5.

Fienberg, Stephen E., and William M. Mason. 1979. "Identification and Estimation of Age-Period-Cohort Models in the Analysis of Discrete Archival Data." *Sociological Methodology* 10: 1. doi:10.2307/270764.

Gaumer, Elyzabeth. 2018. "New York City Housing and Vacancy Survey Summary." "http://www1.nyc.gov/assets/hpd/downloa HVS-initial-Findings.pdf".

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2014. *Bayesian Data Analysis*. Vol. 2. CRC press Boca Raton, FL.

Gittell, Marilyn. 1967. *Educating an Urban Population*. Sage Publications, Inc.

Goodenow, Ronald K, and Diane Ravitch. 1983. *Schools in Cities: Consensus and Conflict in American Educational History*. New York: Holmes & Meier.

Kupper, Lawrence L., Joseph M. Janis, Azza Karmous, and Bernard G. Greenberg. 1985. "Statistical Age-Period-Cohort Analysis: A Review and Critique." *Journal of Chronic Diseases* 38 (10): 811–30. doi:10.1016/0021-9681(85)90105-5.

Mason, Karen Oppenheim, William M. Mason, H. H. Winsborough, and W. Kenneth Poole. 1973. "Some Methodological Issues in Cohort Analysis of Archival Data." *American Sociological Review* 38 (2): 242. doi:10.2307/2094398.

Phillips-Fein, Kim. 2017. *Fear City: New York's Fiscal Crisis and the Rise of Austerity Politics*. Metropolitan Books.

Ravitch, Diane. 2010. "The Death and Life Ofthe Great American School System." *New York: Basic Books* 101.

Rodgers, Willard L. 1982. "Estimable Functions of Age, Period, and Cohort Effects." *American Sociological Review* 47 (6): 774. doi:10.2307/2095213.

Rogers, David. 2006. "110 Livingston Street: Politics and Bureaucracy in the New York City School System." Percheron Press.

Ryder, Norman B. 1985. "The Cohort as a Concept in the Study of Social Change." In *Cohort Analysis in Social Research*, 9–44. Springer.

Stan Development Team. 2016. *RStan: The R Interface to Stan* (version 2.14.1). http://mc-stan.org.

Wallace, Mike. 2017. *Greater Gotham: A History of New York City from 1898 to 1919*. Oxford University Press.

Yang, Yang. 2006. "2. Bayesian Inference for Hierarchical Age-Period-Cohort Models of Repeated Cross-Section Survey Data." *Sociological Methodology* 36 (1): 39–74. doi:10.1111/j.1467-9531.2006.00174.x.