A Brief Analysis of BC Grade-to-Grade Transitions

Warren Zhu

20 April 2022

Introduction

Although test scores can be quite informative, they can be misleading–the difficulty of tests can fluctuate dramatically, effectively resulting in student grades also fluctuating. On the other hand, although grade-to-grade transition rates do not reflect the quality of the education too accurately, they are still an excellent indicator of how much support a government is providing to its youth. As such, through the use of open data from British Columbia, we sought to answer the following question:

How has the British Columbian Government's educational support (as indicated by grade-to-grade transition rates) in the past 30 years differed in different kinds of schools, as well as for different types of British Columbian students (i.e., Indigenous? Special needs?)? Are there any other specific attributes that is strongly related to a lower/higher quality of educational support?

Do note that this is the pdf version of the report for this project. If you wish to see the interactive web version of this report that covers things in more detail (for instance, there are interactive figures there), you can simply click here. Various later parts of this PDF will reference the interactive figures of the report on the website. If you are curious about the code that was written for this investigation, you can find the github repository here.

Methods

Data Prep

All the data that was used in this assignment was from British Columbia open data, and all the files that we used came in the form of '.csv' files. The main data set we used was the 1992-2021 data on grade-to-grade transition rates published by British Columbia Education Analytics. Additional data was needed for machine learning, so we also used a dataset on class Sizes from 2006 to 2021, a dataset regarding British Columbian Teachers for each district from 1991 to 2016, and a dataset with information regarding each district's 2018/19 office. As all of our data comes from the British Columbian open data portal and is regarding the British Columbian education system, they all share a similar format–the data is separated into a provincial, district and school level, and it shares common "IDs" for certain variables, such as "School_Numbers". It is worth noting that few of the files were tidy–many of the files had some rows that recorded data on "all students" whilst other rows recorded data on "Indigenous students", so each row should not be treated as its own observation. As such, we had to be careful when working with it. We merged and filtered the data from all of the files into three data tables–one for the provincial level, one for the district level, and one for the school value. Rather than imputing, we removed entries in our table with missing data, as I felt as though imputation would lead to my figures being misleading.

Procedure

Before doing anything meaningful, we should first do some exploratory data analysis. Based on the structure of the data, I felt as though it would be most appropriate to look at the situation from all three levels. We begin by making interactive figures to understand the situation from a provincial and district level. We will also perform basic model fitting on a school level. By mimicking the structure of our data, we can make the most out of it! Our data exploration is primarily for the purpose of answering our first question: **How has the British Columbian Government's educational support (as indicated by grade-to-grade transition rates) in the past 30 years differed in different kinds of schools, as well as for different types of British Columbian students (i.e., Indigenous? Special needs?)?**

After doing some basic Exploratory Data Analysis, we will fit some machine learning models (namely regression trees, bagging, random forests, boosting and extreme gradient boosting) on a school level, as there are not enough features or number of observations on the district and provincial levels to fit these machine-learning models in a meaningful way. The primary focus of this section is to answer our second question: **Are there any other specific attributes that are strongly related to a lower/higher quality of educational support?**

Results

Exploratory Data Analysis

Let us begin with some exploratory data analysis. You can find the interactive figures **on the website** here. We will explore the data on the provincial level, district level, and school level.

Figure 1 **from our website** is an interactive figure that allows you to do Exploratory Data Analysis on a Provincial Level. Overall, as time went by, the percentage of students that successfully transitioned to the next grade increased. This is the most prominent for students in grades 8-11. It is also worth noting that the percentage of students in grades 10/11 that successfully transition to the next grade are significant lower than the percentage of students in the lower grades that successfully transition.

As for specific sub-populations, if we focus on the difference in transition rates between students from public schools and independent schools, we can see that the rates of grade transitions for students of BC public schools are lower than that of BC independent schools, but the increase in the rate of successful grade transitions is significantly higher in public schools. Furthermore, private schools (i.e., independent schools) seem to have a much higher rate of transition, especially for grade levels 8 and above. This does make sense, as private schools are often times more expensive, and so the students attending them naturally live in my affluent families. Likewise, when we focus on the difference in transition rates between Indigenous and Non Indigenous students, we can see that the rate of successful grade transitions is significantly lower in Indigenous students compared to their non-Indigenous counterparts. We can also see that the increase in the rates of grade transitions over the years is higher for Indigenous students compared to non-Indigenous students. Lastly, when we focus on the difference in transition rates between Special Needs and Non Special Needs students, we can see that the rate of successful grade transitions is significantly lower in special needs students compared to students without special needs.

Figures 2 (interactive heat map) and 3 (interactive box plot) **from the website** were made so that one could more easily explore the district level. We can see that overall, the percentage of students who transitioned to the next grade for Grade 11 fluctuates very dramatically. Playing with the interactive box plot, one can see that the variation in the rate of transitions from grade 11 to 12 from 2015 to 2020 has been significantly reduced (when compared with the past years). The median of the rates has also increased over time. As such, these plots suggest that there is more equality (due to the decrease in variation) in terms of education/opportunities across the districts, and the education/opportunities offered is of a higher quality (as indicated by the higher median). Furthermore, have a look at figures 1-4 **on this document**. The colour scale for the transition rates are the same for all four figures. Notice how figures 1 and 2 look relatively the same, while for figures 3 and 4, we can clearly see that the data points have become noticably more red.

This shows us how significantly the transition rates for higher grades (like Grade 11) have increased with respect to the lower grades (like Grade 1).

Again, let us focus on specific sub-populations now, but at a district level. Playing with the settings of Figure 3, we can see that for most grades (particularly the higher grades, like grade 11) the median percentage of Indigenous students per district who transitioned is significantly lower than that of non-Indigenous students in 1993. This difference is much smaller in 2020, but the median percentage is still less. The same can be said about the variation: the interquartile range (IQR) for the Indigenous 11th graders in different districts was much larger than that of the non-Indigenous counterparts in 1993. In 2020, this different has indeed shrunk, but the IQR for the Indigenous children is still roughly twice that of their non-Indigenous counterparts. As such, we can certainly say that there's an improvement in the quality of education/opportunities for both Indigenous children and non-Indigenous children, but more work needs to be done. Likewise, the relationship between students with special needs and those without (again, particularly for the higher grades) is very similar to what was seen between Indigenous and non-Indigenous children–the variation for both of the rates decreased over time, but the variation of the rates for the students with special needs was typically higher than the rate for students with special needs. The difference now is minimal though (at least when compared to the past). The median rates for both increased over time (although there was a period of time in which the special needs rates were decreasing), but the non-special needs students had a higher median rate for a while now. That said, the difference in the median rate in 2020 was very small.

As for the school level, I ended up fitting three linear models that predicted grade-to-grade transition rates (i.e., the grade-to-grade transition rate of a specific school, grade level, and a certain population). The first model further split the groups into Indigenous students and non-Indigenous students, while the second model split the group into ones with special needs and without. The third model simply dealt with all students (i.e., we did not split into subgroups). The predictors I used was the year, the grade level of interest, and whether the school was public or private. I also used the sub-population of interest as a predictor in my first two models.

Note that the point of creating these models was not to create a model for future prediction–we will do that at the machine learning section (which is below this section) instead. The models created here were purely for interpretability. As such, although my first, second, and third models had fairly low R^2 values, as shown in table 1 below, I was not bothered.

All of the predictors had an extremely small p-values (less than 1e-16), which hints at the fact that it is very likely that there is a difference in grade-to-grade transitions between people of different grades, between Indigenous and non Indigenous students, between special needs and non special students, and between students of public and private/independent schools. The coefficients of our models tell us more about these differences. The first model suggests that Indigenous students had a 2.502% lower mean rate of successful transitions, while our second model suggested that students with special needs have a 1.126% lower mean rate of successful transitions compared to their non-Indigenous and non-special needs counterparts, respectively. Looking at the coefficients for the grades for all of our models, we can also see that our model suggests that students in high school (BC high school starts at grade 8) are significantly less likely to transition to their next grade compared to their elementary school counterparts. This is probably due to the riskier behaviors certain high school students have, which may result them getting in all sorts of trouble. Lastly, our model suggests that the transition rate for public schools is smaller than that of private schools, but this different is fairly negligible. This provides a nice answer to our first question at a school level.

Table 1: Various Important Statistics from models

Model	Adjusted R-Squared	Residual Standard Error	Estimated Rate Increase per Year
Model 1: Indigenous/Non-Indigenous	0.2164	5.730	0.151300
Model 2: Special Needs/Non-Special Needs	0.1542	5.482	0.133000
Model 3: All Students	0.1551	4.632	0.089253

Table 2: Estimated Grade Transition Rate (%) Increase Per Year

Grade	Model 1: Indigenous/Non-Indigenous	Model 2: Special Needs/Non-Special Needs	Model 3: All Students
Grade 1	0.15130	0.1330	0.089253
Grade 2	0.30900	0.2897	0.275735
Grade 3	0.05495	-0.2851	0.071903
Grade 4	0.56520	0.6558	0.505705
Grade 5	0.69660	0.8607	0.644274
Grade 6	0.76970	0.8654	0.670936
Grade 7	0.62770	0.7923	0.728806
Grade 8	-1.76870	-0.9410	-1.157668
Grade 9	-3.23370	-2.1500	-2.472243
Grade 10	-5.74970	-4.3720	-5.064621
Grade 11	-8.86170	-7.1360	-7.991081

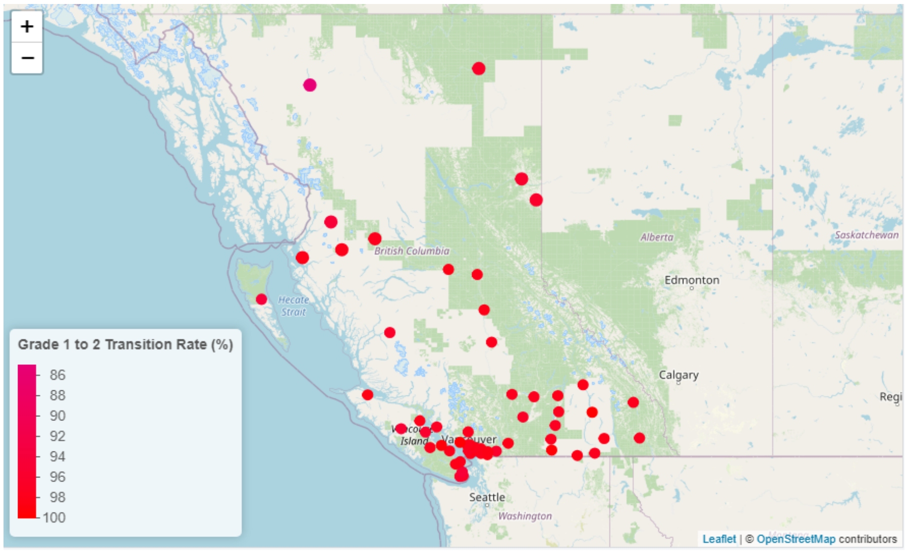

Figure 1: Map displaying the location of the board offices for each district. The color of the points is based on the rate of successful transitions from Grade 1 to 2 in 1993.

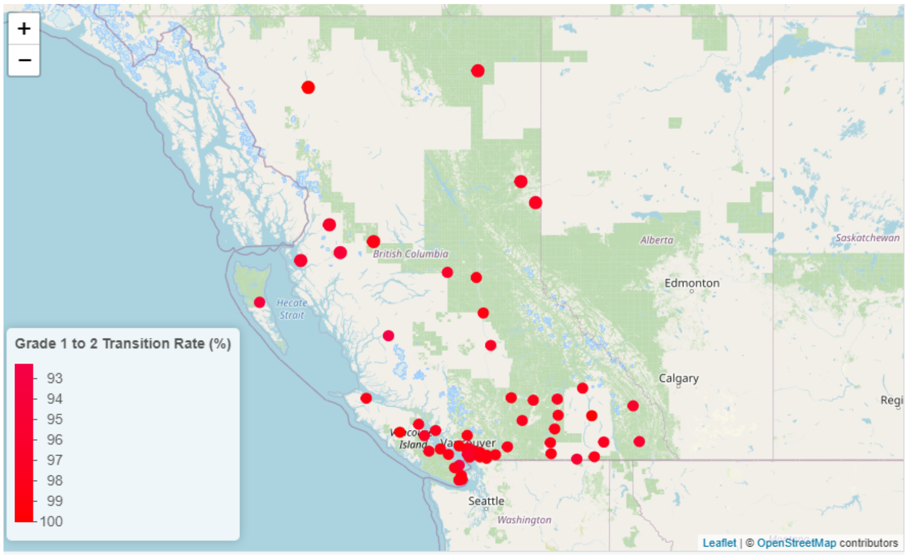

Figure 2: Map displaying the location of the board offices for each district. The color of the points is based on the rate of successful transitions from Grade 1 to 2 in 2020.

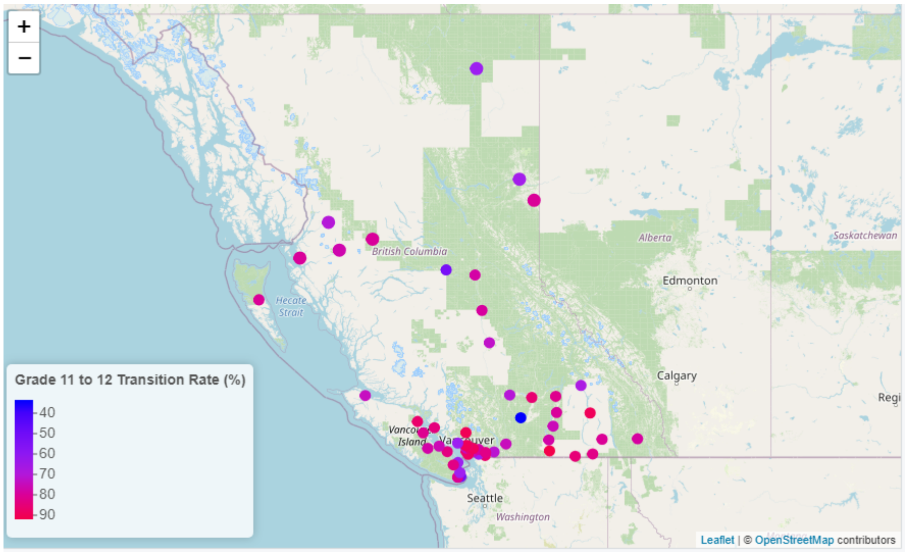

Figure 3: Map displaying the location of the board offices for each district. The color of the points is based on the rate of successful transitions from Grade 11 to 12 in 1993.

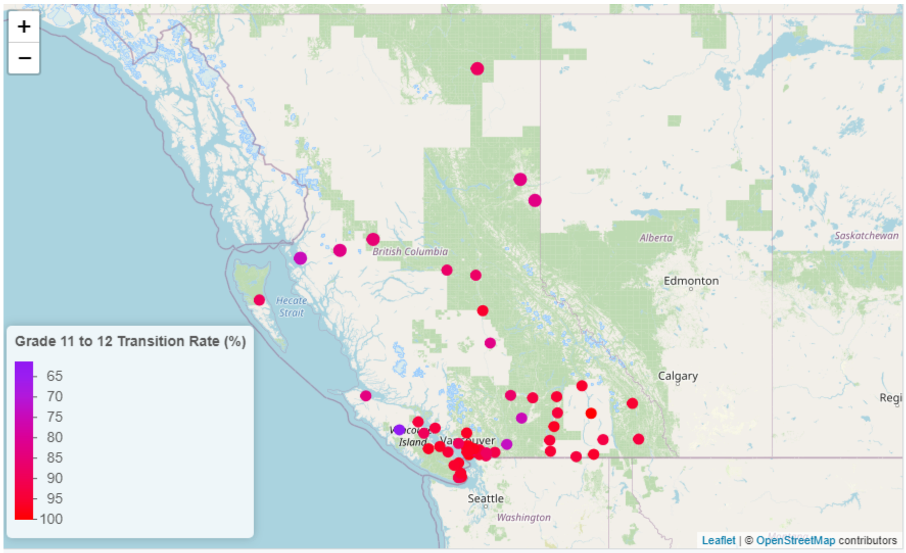

Figure 4: Map displaying the location of the board offices for each district. The color of the points is based on the rate of successful transitions from Grade 11 to 12 in 2020.

As we will be training a machine learning model on a dataset at the school level, it is worth doing some exploratory data analysis on that dataset. Note that as the dataset that we will use for the machine learning model is the result of merging several datasets, and as some datasets cover a different range of years, the machine learning dataset will only cover years from 2006 and 2016. Figure 4 on the website is an interactive plot for this dataset, and it shows how the number of full time educators in a district and the average class size for a grade group in a school may be related to the grade transition rates. One observation we can make is that the lower grade transition rates are often coming from observations for relatively high grades (like grade 11) and from schools with a relatively relatively low number of students for a given grade group.

Machine Learning

Prior to feeding our data in, we removed province specific data—for instance, we removed school names and district names. This reduces overfitting, so our model can be somewhat generalized to schools outside British Columbia. Our machine learning dataset has data from 2006 to 2016–as such, my train test split involved me taking data from the final 3 years (2016, 2015, 2014) and placing them into the test dataset. Everything else was placed into the training dataset. By doing this, we are in essence predicting the “future” based on our current information. I chose to this for 3 years because this makes things roughly a 70-30 train-test split, which is typical train-test split in machine learning. Another important thing to note is that all schools examined here are public schools, as we could not access the relevant information for independent schools.

Or at least, that was the original plan. When constructing our bagged and random forest models, we realized that there were computational constraints, and we were not able to produce the models using the full train dataset (which had 107885 observations). As such, I ended up sampling 2500 observations from our train and test data sets, and used these samples for model creation.

Obviously, this does make a difference in terms of results. We were able to make a decision tree with the full datasets (but not a random forest), and the mean square error on the full test data set was 11.7754406. On the other hand, for our decision tree based on the two samples, our test mean square error was 21.4133868. This is a very significant difference in performance, so we should take our results with a grain of salt while also knowing we can do quite a bit better when we have more computational power.

We ended up making 5 models: a regression tree, a bagged model, a random forest, a gradient boosting model, and an extreme-gradient boosting model. Their attributes are described in the table below.

Table 3: Test MSE and Variable Importance of each of our models

Model	Most Important Predictor	Second Most Important Predictor	Test MSE
Regression Tree	Grade Level	Number of Students of the Grade in the School	21.41339
Bagging	Grade Level	Number of Students of the Grade in the School	19.34337
Random Forest	Grade Level	Number of Classes in School	18.83897
Gradient Boosting	Grade Level	Number of Students of the Grade in the School	23.51652
Extreme Gradient Boosting	Number of Classes in School	Number of Students of the Grade in the School	18.64410

Note that for our extreme gradient boosting model, the `grade` predictor was actually split into 11 different indicator predictors, each one for the grades between 1 to 11, so its importance was not as high in the model. As such, one could argue that the grade level of students is a very important if not the most important factor to account for when we are considering the grade-to-grade transitions for students of a certain grade from a certain school at a certain year. The second most important predictor for most of our models is the number of students of our grade of interest in our school of interest. This makes sense to me, as from our data exploration from earlier, we clearly saw how high school students (grade 8-11) were less likely to have a successful grade transition. Furthermore, the number of students of a grade of interest is a proxy variable for the general size of a school. As large public schools tend to be in big cities whilst small public schools are in relatively rural locations, it makes sense that a smaller number of students for a given grade could indicate a lower likelihood of transition.

Conclusions and Summary

In conclusion, through exploratory data analysis, we could see that over the years, the rate of students successfully transitioning from grade-to-grade has increased over the years. That said, there is still a gap in transition rates between students from public schools and students from private schools, students with special needs and students without, as well as Indigenous students and non Indigenous students. These gaps are decreasing, though, which is a good sign for the future. Another thing we found is that students in high school are significantly less likely to transition compared to students in lower grades, likely due to the rebellious nature of teenagers. From our machine learning models, by looking at the most important predictors, we saw that the grade of a student and the size of a school is related to the grade transition rate. We just explained the reason why the grade would matter significantly, so I will not repeat myself–as for the size of the school, it can be explained to be a proxy variable of whether or not a school is rural or not.

In the future, we could also look at data on various exam results and assessment statistics provided by the British Columbian government and relate them to our findings in this study.