JSC370-Final Report

Siyi Zhu

2025-04-29

1. Introduction

Every year, millions of people and vehicles cross the Canada border for work, tourism, and so on. These movements are critical not only for local economies but also for infrastructure planning. Understanding daily border traffic can help governments design targeted marketing for tourism, and forecast revenue from duty-free sales. Additionally, incorporating U.S. data extends our insights: comparing patterns on both sides allows policymakers to identify best practices, leverage U.S. operational strengths, and address Canada's crossover vulnerabilities. This binational perspective fosters more coordinated, evidence-based policies and infrastructure investments along the Canada–U.S. corridor.

To explore these rhythms, we use two complementary open-data sources covering January 2018–May 2019:
1. **Government of Canada "Traveller volumes by port of entry and month"**: It includes date, port of entry, region, mode of transportation, and the volume of crossings. Understanding border traffic is essential for transportation planning, security, and economic forecasting.
2. **U.S. Department of Transportation "Border Crossing Entry Data"**: This dataset provides complementary insights from the U.S. perspective, covering similar details such as port name, date, measure (vehicles, pedestrians, etc.), and crossing volumes.

The research question is: **To what extent do seasonal patterns, geographic regions, day-of-week (weekday vs. weekend), and transportation mode explain daily border-crossing volumes into Canada between January 2018 and Dec 2019? Additionally, how do seasonal patterns and geographic concentrations manifest on the U.S. side of the Canada–U.S. boundary between January 2018 and Dec 2019?**

Within this question, there are four hypothesis:
1. **Seasonality (Canada & U.S.)**: Peaks in summer and troughs in winter on both sides of the border.
2. **Weekday vs. weekend (Canada)**: Travel volumes drop on weekends.
3. **Transportation Mode (Canada)**: Land crossings dominate across the whole boundary.
4. **Regional Concentrations (Canada & U.S.)**: A small set of ports (both Canadian and U.S.) capture the bulk of crossing volume.
5. **Cross-Border Comparison (Canada v.s. U.S.)**: Although the U.S. side is expected to follow the same summer-peak/winter-trough pattern and port-concentration structure, the overall volumes and exact timing of those peaks may differ from the Canadian experience.

2. Methods

2.1 Acquiring Data

For "Traveller volumes by port of entry and month" from government of Canada, I acquire the data from by reading csv to load the dataset from 2018-01-01 to 2019-12-31. I retrieved the data by using `fread` and filtering in the specifc range of the date. For "Border Crossing Entry Data" from U.S. Department

of Transportation, I acquire the data from https://data.bts.gov/Research-and-Statistics/Border-Crossing-Entry-Data/keg4-3bc2/data_preview by calling API to laod data. Similarly, it filters to the same range of dates.

- **Canadian dataset ("Traveller volumes by port of entry and month")**
  I downloaded the CSV file from the Government of Canada Open Data Portal (https://open.canada.ca/data/en/dataset/1b1c2b92-b388-47d9-87d4-01aee8d3c3e4/resource/22653cdd-d1e2-4c04-9d11-61b5cdd79b4e) covering January 1, 2018 to December 31, 2019. In R, I used `fread()` to read the file and then filtered the `Date` column to the 2018-01-01–2019-12-31 window.
- **U.S. dataset ("Border Crossing Entry Data")**
  I accessed the U.S. Department of Transportation's API endpoint (https://data.bts.gov/resource/keg4-3bc2.json) using `jsonlite::fromJSON()` to flatten the response into a data.table. I converted the returned date strings into Date objects, then applied the same January 1, 2018–December 31, 2019 filter. This parallel filtering guarantees that both datasets cover the exact same date range for subsequent merging and comparison.

2.2 Data Wrangling and Cleaning

After importing the raw data, all date strings were converted into native `Date` objects—using `lubridate::ymd()` for the Canadian data and `as.Date()` for the U.S. data since they have different date formats in raw data. this can ensure the consistency. Then, I filtered both datasets from January 1, 2018 to December 31, 2019. Since we also need to compare two countries' differences, I merge them together where non-numeric characters were stripped from the Canadian `Port of Entry` and cast to an integer `PortCode`, and the U.S. `port_code` was likewise converted to integer. All NA values are removed and all extreme outliers at the 1st and 99th percentiles were winsorized to mitigate data-entry errors. Finally, an inner join on `(PortCode, Date)` yielded a single, clean `data.table` ready for analysis.

2.3 Data Exploration

We reshaped and summarized the merged dataset using **`dplyr`**, **`tidyr`**, and **`data.table`**, grouping by day, month, region, and mode. Static visualizations—including time-series plots, boxplots, and stacked bar charts—were produced with **`ggplot2`**. For interactive exploration, we used **`plotly`** to create hoverable seasonal-trend charts and **`leaflet`** to map the top-30 ports with scaled circle markers. Port-coordinate lookups and inline tables were facilitated by **`stringr`** and **`tibble`**, and summary tables (e.g., mean daily volumes by mode or region) were formatted for PDF output with **`kableExtra`**. This toolkit allowed rapid iteration on both high-level patterns and granular breakdowns.

2.4 Modeling & Tests

I applied a series of predictive models and classical hypothesis tests corresponding to the hypotheses formulated earlier. For the **Trend of Border Crossings Over Time**, I created an interactive line graph and applied a **Mann–Kendall test** to evaluate whether a monotonic trend existed over time without assuming linearity. The **Trend of Border Crossings by Region** and the **Trend of Border Crossings by Transportation Mode** were visualized using an interactive bar graph and an interactive line graph, respectively. Both analyses were complemented with **one-way ANOVA** tests to assess whether significant differences existed in mean daily volumes across Canadian regions and transportation modes. The **Trend of Border Crossings by Region and Transportation Mode** was represented through an interactive grouped bar graph and analyzed using a **two-way ANOVA** model to investigate the main effects and interaction between region and mode. Seasonal shifts in mode composition were displayed with an **interactive stacked bar graph**, and statistical significance was assessed via a **Pearson's chi-square test** on

a contingency table of transportation mode by meteorological season. Additionally, I created an interactive map to illustrate the geographic concentration of crossing volumes across different census regions.

For the U.S. analysis, I first visualized monthly border crossing patterns using an interactive line graph. To explore cross-border dynamics between Canada and the U.S., I created another interactive line graph depicting monthly trends side-by-side and applied a **cross-correlation function (CCF)** analysis to investigate potential lead–lag relationships between the two countries' crossing volumes. An additional interactive map was developed to show the geographic concentration of border crossings across both Canadian and U.S. regions.

In predictive modeling, I developed several parametric and semi-parametric approaches. Three **linear models (LMs)** were fitted: a simple linear trend model, a quadratic time trend model, and a categorical model using month, region, and mode as predictors. In parallel, three **Generalized Additive Models (GAMs)** were constructed, progressively incorporating smooth terms for time and additional categorical predictors. GAMs extend linear regression by integrating **smooth spline terms (`s()`)**, allowing the model to capture complex seasonal and temporal variations without needing to pre-specify polynomial degrees. Model selection was guided by **Akaike's Information Criterion (AIC)**, seeking a balance between model fit and complexity. To further validate predictive performance, I manually conducted **5-fold cross-validation**: the dataset was partitioned into five folds, the full GAM model was fitted to training folds, predictions were made on testing folds, and evaluation metrics including **Root Mean Square Error (RMSE)**, **Mean Absolute Error (MAE)**, and **R-squared (R^2)** were computed. This approach provided an unbiased assessment of out-of-sample predictive accuracy.

3. Results

3.1 Canada Border Crossing Trends

The daily trend of border crossings into Canada from January 2018 to December 2019 is illustrated in Figure 1 on the website under Home page. The interactive line graph reveals strong seasonal patterns, with prominent peaks during the summer months and declines during winter periods in both years. It is clear to show that total crossings by day increase from 2018-01 to 2018-08, and start to decline afterwards, reaching bottom value in 2019-01. Then, another similar pattern appears in the next year which increases from 2019-01 to 2019-08 and decline again. It supports hypothesis 1.

This visualization is tested by using a Mann-Kendall trend test and has shown the result in Table 1 below. It has a Kendall's tau of 0.05 and a two-sided p-value of 0.03. The p-value is smaller than 0.05, which shows that it is statistically significant and reject the null hypothesis that there is no monotonic trend over time. So, after rejecting it, we know that this statistical result strengthens the conclusion that the observed seasonal variations are systematic rather than random and supports the hypothesis I set before that border crossings exhibit a regular seasonal cycle and have peaks usually during summer months.

Table 1: Mann–Kendall Trend Test Result for Total Border Crossings

Tau	P.value
0.05	0.03

After exploring the whole trend over time, we put attention into regional effects. Figure 2 on the website under Home page uses an interactive bar graph to show the total border crossings by Canadian region over the same period. From the graph, we easily find that Southern Ontario and the Pacific Region has the largest number of total daily crossings, where each over 40 million. Conversely, Atlantic region will have the lowest total daily crossing volumes. This result also visually shown in the interactive regional concentration map shown in Figure 11 under Regional Concentration Map page. The top 2 regions will have larger circles to

correspond to their volumes on the map. In Figure 11, it strongly show that a small set of ports will capture the main crossings, which supports hypothesis 4.

For the trend by Canadian region, I also conduct a one-way ANOVA to further verify the observation. The ANOVA output shows in Table 2 as shown below. It shows a very large F-value which is equal to 2826.51 and a very small p-value which can reject the null hypothesis that the crossing volumes are the same across all regions. Therefore, from the test we know that there are regional differences in bordeer crossing volumes, which confirms hypothesis 4 made before and aligns with what shown in both Figure 2 and Figure 11.

Table 2: One-Way ANOVA Results for Total Border Crossings by Region

Term	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
Region	6	2.32e+12	3.87e+11	2826.51	<0.001
Residuals	5103	6.99e+11	1.37e+08	NA	NA

For the factor of transportation mode, I plot an Interactive line graphs showing the distribution of border crossings by transportation mode as shown in Figure 3 on the website under Home page. We can easily see that land crossings dominate through the two-year range, and air dominate following, while marine and rail will contribute less. Besides, in Figure 11 which is the map in Canada, the busiest port shown is Pearson International Airport (Toronto) with over 18 million crossings, followed by Trudeau International Airport (Montreal) with over 13 million. So, it also aligns with the result that air will be the second highest crossing volumes in Canada.

To verify my finding and hypothesis, I conduct a one-way ANOVA as shown in Table 3 below. It has an output of a large F-value of 6797.1 and a very small p-value that is lower than 0.05 which rejects the null hypothesis (assumes all transportation modes have same crossing volumes). The test result verifies that the transportation modes will have different crossing volumes which has the trends shown in Figure 3 and overall supports hypothesis 3.

Table 3: One-Way ANOVA Results for Total Border Crossings by Transportation Mode

Term	Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
Mode	3	1.25e+13	4.16e+12	6797.1	<0.001
Residuals	2915	1.79e+12	6.13e+08	NA	NA

Next, I further investigate the trends grouped by two factors: season and transportation mode, and region and transportation mode. In Figure 4 under the same Home page, I plot the seasonal composition of border crossings by transportation mode across four seasons in stacked bar plot. The figure indicates that land crossings dominate in all four seasons, which same as the patterns in Figure 3, while other transportation mode may fluctuate among four seasons. Besides, the stacked bar plot shows that the total volume of crossing is much higher during summer compared to other seasons, and winter has the less volume of crossings, which also same as the pattern in Figure 1.

Then, I conduct a Chi-squared test shown in Table 4 which has the result that p-value is really small, which rejects the null hypothesis that assumes the transportation mode is independent of season. So, we know that the transportation mode factor is dependent and associated with season factors, which also aligns with what we see in the Figure 4.

Table 4: Chi-square Test Result for Transportation mode and Season

	Chi.squared	df	p.value
X-squared	3265731	NA	<0.001

In Figure 5 under home page, I plot the trend grouped by both region and transportation mode. Similarly, the land crossings still dominate across all regions except Greater Toronto Area Region, Prairie Region and Querbec Region. Those regions will have air be the highest corssings, which may because the people in those regions would like to choose to cross borders by air. Marine and rail will continuelly be small in all regions as same in the trend by region.

In this case, I choose to conduct two-way ANOVA which has the null hypothesis that region and transportation mode are independent with each other. The result shown in Table 5 shown that p-value is very small so that it will directly reject the null hypothesis and verify the trends we observed previously and suggest that transportation mode will have association with region on affecting the total crossings.

Table 5: Two-Way ANOVA Results for Region and Transportation Mode

Term	Sum.Sq	Mean.Sq	F.value	Pr..F.
Region	7.59e+11	1.27e+11	4020.0	<0.001
Mode	2.15e+12	7.17e+11	22783.9	<0.001
Region:Mode	2.06e+12	1.29e+11	4090.8	<0.001
Residuals	5.29e+11	3.15e+07	NA	NA

3.2 Seasonal and Daily Patterns in Border Crossings

From Figure 6 on the website under Monthly/Daily Patterns page, there is a bar plot illustrates average daily crossings by month, showing clearly that crossings peak is during summer months similarly as mentioned before. In addition to seasonal variation, differences between weekday and weekend traffic is shown in Figure 7 on the website under the same page. It plots in boxplot and we can easily see that the medium in weekends are much larger than that in weekday. Meanwhile, the interquartile range of weekend is larger than that of weekday, illustrating that weekend crossing volums are more variable.

I also conduct a t-test to compare weekday and weekend crossing volumes, which the null hypothesis is that crossing volumes are same in weekends and weekdays. For the result in Table 6 as shown below, the p-value is almost 0, which is much smaller than 0.05. We can reject the null hypothesis can conclude that the crossing volumes are different in weekday and weekends, verifying what we observe in interactive box plots. Overall, we can use this result to support hypothesis 2 we made before.

Table 6: T-test of Daily Border Crossings between Weekdays and Weekends.

Mean (Weekday)	Mean (Weekend)	p-value
250580	305063	0

3.3 Modeling (GAM v.s. LM)

After observing the trends under different factors, I turn into explore the relationship between border-crossing volumes and time, Generalized Additive Models (GAMs) were fitted and compared against several linear models. Three GAM specifications were tested: a basic smooth over time (GAM-basic) shown in Figure 8 on the website under Modeling page, an extension adding a seasonal month factor (GAM-month) in Figure 9, and a full model including month, region, and transportation mode (GAM-full) in Figure 10. Three linear models (LMs) were fitted: a simple linear trend model, a quadratic time trend model, and a categorical model using month, region, and mode as predictors. Since linear models only fits linear well, I didn't plot them but use AIC to directly compare and evaluate them with GAMs models plotted before.

AIC has been shown in Table 8 on the website under the same page as GAM models. From the table, we can find that gam_full which is the GAM model with month, region and transportation mode will have much smaller value in AIC, indicating that this is the best model to fit. This full GAM model incorporating

month, region, and transportation mode achieved the lowest AIC among all models, indicating that these three factors are key drivers of daily border-crossing volumes. This result supports Hypothesis 1 (seasonality), Hypothesis 3 (mode dominance), and Hypothesis 4 (regional concentration), confirming that volume patterns are primarily shaped by seasonal cycles, modal differences, and geographic location—rather than simply progressing over time.

Finally, I use 5-fold cross validation to verify the predictive accuracy by separating them into train dataset and test dataset. The result has been shown in Table 7 below. The RMSE and MSE is relative low in metrics compared with the total number of observations, meaning that the error of prediction is low. R^2 results indicates that there may have any other external factors such as weather, economic environment and so on that not include in the model affects the results.

Table 7: 5-Fold Cross-Validation Results for GAM Model

Metric	Value
Average RMSE	3229.07
Average MAE	1662.66
Average R^2	0.19

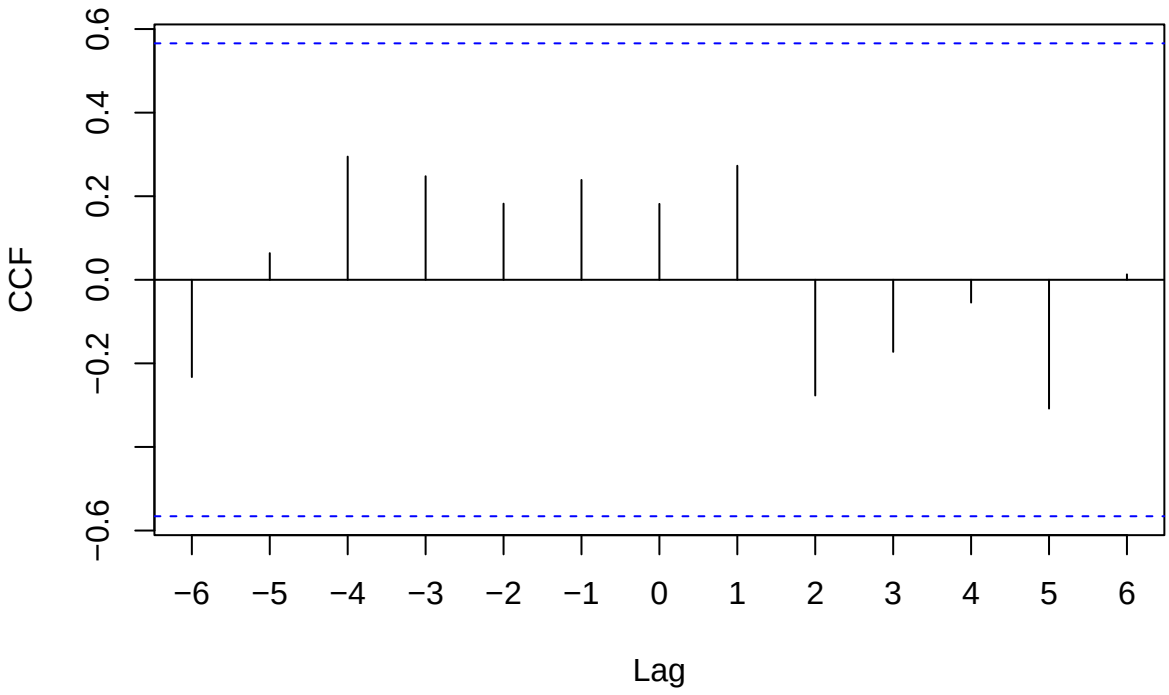
3.4 Seasonal Patterns and Geographic Concentrations on the U.S. Side

The seasonal trends on the U.S. side of the border exhibit patterns similar to those observed in Canada. For seasonal trends in month, as shown in Figure 12 one website under US Analysis page, the trend shows that it will reach the highest peak around July and August, and start to decline afterwards until November will have a little increase. It supports hypothesis 1 that summer is the peak season of border crossings.

In Figure 13, I directly plot line graph by comparing differences in the trend of Canada and U.S. monthly volumes. The difference is that, Canada will have a very significant trend in the rise in June, while U.S. has a much steadier increase compared to Canada. Similarly, Canada has a sharp decline after October, while U.S. will decrease steadily. It indicates that U.S. will have a smaller seasonal fluctuations relative to Canada. So, it supports hypothesis 5.

I also compute cross-correlation function (CCF) between Canadian and U.S. monthly crossing volumes as shown in Figure 15 below. All CCF lags are within the blue dotted confidence interval lines. This suggests that there is no statistically significant cross-correlation between the two time series at any lag. This suggests that while Canada and U.S. border crossings show visually similar seasonal trends, their month-to-month fluctuations are not strongly synchronized at a statistically significant level. In other words, although the two sides share broad seasonal patterns, short-term monthly variations appear largely independent.

Figure 15: CCF of U.S. vs. Canada Monthly Volumes



Finally, the regional concentration of Canada and US has been shown in Figure 14 on the website under US Analysis page. The map shows that the total volume in US will be larger than Canada in almost all overlapping ports. Meanwhile, similar as Canada, the map highlights that a small number of ports dominate the total crossing volume, which supports hypothesis 4. Besides, U.S. and Canada both has similar patterns in border crossing that the activity will not distribute evenly to each port. Instead, it will mainly focus on several main ports.

4. Conclusions and Summary

4.1 Findings & Conclusion

This study investigated how **seasonal patterns, geographic regions, day-of-week effects, and transportation modes** explain daily border-crossing volumes into Canada between January 2018 and December 2019, while also exploring how seasonal and spatial dynamics manifest on the U.S. side.

All the findings in **Results** part has strongly supported the hypothesis stated at the beginning of the report: - **Seasonality** is a major driver of border crossings. Both Canada and the U.S. exhibit **clear peaks during summer** (July–August) and **declines during winter** (December–February), though the **seasonal variation is sharper in Canada**. - **Day-of-week effects** are significant: **weekend crossings** are consistently higher than weekday crossings, verified by a t-test with near-zero p-value. - **Transportation mode** strongly influences crossing volumes, with **land crossings overwhelmingly dominating** across seasons and regions. - **Geographic concentration** is prominent: a **small set of ports** captures the majority of border crossings on both sides of the border, particularly near **Southern Ontario–Michigan** and the **Pacific corridor**. - Although seasonal structures are similar, **month-to-month fluctuations**

between Canadian and U.S. crossings are not statistically synchronized, as indicated by the cross-correlation function (CCF) analysis.

From a modeling perspective, the **GAM-full model**, which incorporated month, region, and transportation mode, achieved the **lowest AIC** and reasonable cross-validation metrics (**RMSE 3,229**, **MAE 1,663**, **R^2 0.19**). While the explained variance was modest, the GAM model successfully captured the dominant seasonal and regional patterns.

Overall, in a bigger picture way, I can find that the strong seasonal pattern suggests the government that they can optimize security by scaling more operations in summer, especially investing more revenues in land-based infrasturcture since most people will use transportation mode by land. Meanwhile, if the policy-maker can make more policies to attract people cross border to come to Canada in winter season, it can lead to increase iin economics since now winter has the lowest total daily crossing volumes. Besides, the slight mismatch in seasonal peaks between Canada and the U.S. indicates opportunities for more synchronized cross-border event planning or tourism marketing. This allows policy-maker to target interventions more precisely, improving traveler experience and operational efficiency at key border ports by learning advantage of policies made in U.S.

4.2 Limitations and Future Work

1. **External factors** like weather, policy changes, global special situation and so on will be exteral factors that affect the reality. So, in the model we mentioned before, the R^2 will be a little bit large and indicate there are other factors that may affect the variance.

2. The study focused on **aggregate crossing volumes**, but did not distinguish among traveler types such as coming Canada for work, study, trip and so on, which could show different temporal patterns.

3. The study is mainly focused on from 2018-01 to 2019-12, which is limited in a specific time range. This time range is before COVID-19. After the covid, the trend may a little bit different from this range of time since global economics have all been striked due to the covid.

For my future study, I can apply similar models to a much closer time range and compare the trend difference before and after the covid. Also, I can find some otehr datasets to show the consumer type and transportation type to have more detailed information in investigation.