Fall 12-14-2017

U.S. - Canadian Border Traffic Prediction

Colin Middleton
Western Washington University

U.S. - Canadian Border Traffic Prediction

Colin Middleton

December 14, 2017

1 Introduction

1.1 Setting

The goal of this project is to predict the flow of traffic arriving at the Peace Arch and SR 543 border crossings from the south based on current traffic flow data along Interstate-5. This model is to use data from northbound traffic on Interstate 5 from milepost 251.81 to the Canadian border crossing as well as northbound traffic along SR 543 from the Interstate 5 exit at milepost 275.12 to the Canadian border crossing. Traffic flow data is collected from sensors located along Interstate 5 and SR 543 and stored every five minutes as a count of how many cars passed the sensor in that five minute period. Sensors collect data twenty-four hours a day, seven days a week, except for days when maintenance is performed, in which case the data reads zeros for the whole day.

At the border crossings there are enough border officials to open all ten booths at the Peace Arch crossing and all ten booths at the SR 543 crossing, but the border compound would prefer to have as few officials operating booths as possible while maintaining a minimal queue of cars crossing the border. There are other important tasks in the border compound that border officials not operating a booth need to accomplish.

1.2 Question

The task of this model is to predict traffic flow at the Canadian border crossings on Interstate 5 and SR 543 based on historical and current data from those two locations, historical and current data from sensors along Interstate 5, or both. This model must allow sufficient time for officials at the border to react to predicted traffic flow by opening additional booths or closing currently open booths – a thirty minute lead time is deemed appropriate. If possible the model should provide an early prediction of traffic flow that may be inaccurate (an early warning), but whose accuracy increases as the traffic pattern approaches the border. The initial model is meant to only apply to weekend traffic during the summer (July - September) and so only data from these times is used in testing and analysis. This stipulation is added because it is assumed that weekend traffic is similar independent of which weekend is observed, but weekend and weekday traffic may differ significantly. The same assumption was made for summer and winter traffic.

2 Mathematical Approaches

2.1 Assumptions

Before each prediction method is described, it is prudent to explain several assumptions and concepts that are common to most of the models. In order to predict which cars are headed to which border crossing, Peace Arch or SR 543, a prediction technique was tried which assumed that the ratio of cars going to the border crossing on SR 543 over the number of cars going to the Peace Arch crossing is constant. Some assumption about the relationship between the traffic at Peace Arch and SR 543 crossing is necessary because of the requirement that the model must give a thirty minute lead time for border officials to react to any change in traffic flow that will arrive at the border. This requirement necessitates collecting data from sensors south of the SR 543 exit off of Interstate 5, since data from sensors at or north of this location would not provide sufficient lead time. This means that the available data for prediction does not indicate which border crossing the observed cars are destined to reach. This constant ratio assumption was tested using data from July 5, 2014 and the ratio was found to have a mean of 0.8417 with standard deviation of 0.9942. Because this assumption was inaccurate, methods for the prediction task were tested against either the sum of cars arriving at both border crossings or from one sensor location to another along I-5 (when appropriate).

Another difficulty was finding the appropriate time-gap between sensors. Cars' speeds differ from driver to driver and from one time to the next, so there is no consistent time it takes a car to go from one sensor to another along the road. The method used for finding the most appropriate time-gap was to assume different time-gaps in increments of five minutes, since the data is listed by five minute increments, then compare the data collected from one sensor to another and determine which assumed time-gap produced the least sum of squares of differences of traffic flows between the sensors.

Some of the tried prediction methods only begin predicting what happens in the current day once a few data points have been collected. This gives the models some information on which to base their predictions about what the day will be like.

2.2 Average Method

The simplest approach was to create an "average day" at each border crossing by taking traffic flow data from many historical days and averaging the number of cars for each five minute increment. The predicted number of cars at time t is the average number of cars at time t in the past. This approach would work well if each weekend day had about the same flow distribution as the last. This method is likely currently employed at the border crossings to determine shift scheduling. Only historical data from the border crossings is needed for this method.

2.3 Weather Model

Inspired by weather prediction models this method creates a database of historical traffic flow data and compares the current day to that database to form a prediction. The model takes the data from the current day and finds similar days in the database of historical data (meaning the sum of the differences in the known data points of the current day vs a historical day is below some threshold), then averages those similar days to create a prediction for the rest of the day. If this method were to be implemented a useful addition would be to add current days that are significantly different from all the days in the database to the database in order to increase the match likelihood for any traffic pattern for the current day. Current and historical data from the border crossings is needed.

2.4 Principal Component Analysis

The method of Principal Component Analysis, or PCA, is a linear algebra technique that is fundamentally similar to the Weather Model idea. Here each day is treated as a 288 entry vector and the current day is predicted using the linear combination of vectors (similar to historical days) in a database that best recreate the start of the current day – the part that is known. The same linear combination is then carried out for the entries past the known current day and this combination becomes the prediction for the current day. The database here consists of vectors that attempt to account for as much variation in historical data as possible in as few vectors as possible. Basically the database attempts to be a basis for the set of historical data in as few vectors as possible. This method also requires current and historical data from the border crossings. An attempt was made at incorporating data from other sensors by doubling the length of the vector with data from another sensor and then searching for appropriate linear combinations. This addition requires current and historical data from at least two sensor locations.

2.5 Ratio Method

The Ratio Method assumes that for every five minute increment there exists a constant ratio between the traffic flow sensed at any sensor location and the traffic flow sensed by any downstream sensor in the corresponding five minute increment. The corresponding five minute increment is the most appropriate time gap between the two sensors which is found by using the sum of squares method (see 2.1). For example, if at Station A it is known that twenty cars passed between 9:10 am and 9:15 am and that the most reasonable gap time to use from Station A to Station B is 10 minutes then we can predict how many cars will pass Station B between 9:25 am and 9:30 am based on the known ratio during this specific time of day for these specific sensors. If the ratio here were 0.5 then the predicted number of cars for Station B would be 0.5 x 20 = 10 cars.

This is a nice method because the average ratios between two sensors for every five minute increment must be calculated only once using historical data. Using the data from the current day, a simple multiplication of the current data by the appropriate ratio yields the prediction. This method requires historical data from all the sensors involved in the model (which can vary in number and location) and data from the current day from the upstream sensors used for prediction (which can also vary in number and location, but at least one sensor is required).

2.6 Improved Ratio

The problems with the Ratio Method are that it predicts traffic flow only for the next period, not for the whole rest of the day and that while the historical data becomes smooth and regular when averaged to find the sensor ratios (see Figure 2), the predictions are much more erratic. This is because the predictions are based on incoming data which are erratic which is most likely due to cars going different speeds from one another and different speeds at different times. In the Improved Ratio method an attempt is made at smoothing the predicted traffic flow and predict further in advance. This is done by creating an inaccurate but smooth model of the day using linear combinations of orthogonal sine and cosine functions fit to average historical data (see Figure 3) and then taking the average of that model and the erratic prediction from the Ratio Method.

The smooth function can be used to predict the rest of the day, though inaccurately. Since the Ratio Method only predicts traffic flow at a certain lead time and no further, even an inaccurate prediction could assist in planning. As traffic moves toward the border, the Ratio Method would predict outcomes at the border so the Improved Ratio prediction should become more accurate. This model uses all the data the Ratio Model requires plus some matrix computation to build the sine and cosine model.

2.7 Randomness Checking

The poor results of the attempted prediction methods (see 3 Results) seem to indicate that there is a large amount of randomness inherent to the traffic data or at least there are factors which cannot be accounted for using only the traffic data provided. Some examples of other possible factors are the U.S.-Canadian exchange rates, prices of goods, nearby entertainment events, and weather. An investigation was done of how much randomness the data actually contains.

To investigate randomness, the data of the combined border traffic flow was grouped into bins of similar traffic days and then the traffic on the same days in each grouping were analyzed at an earlier sensor to see if similarity in traffic behavior can be traced backward from the border sensors to previous sensors. If the data had a large amount of randomness from sensor to sensor, then the groupings of similar traffic days at the border would produce bins with no distinct difference from one another (essentially randomly assigned) at a previous sensor. Similarity of traffic days was determined by normalizing each day - dividing each entry by the magnitude of the day as a vector quantity - then averaging the entries for each day, then computing the quartiles for this data set, which became the bins for the groupings. The normalization step was taken because the bins were intended to find days with similar traffic distributions without regard to volume - just similarly shaped distributions.

<u>3 Results</u>

3.1 Assumptions

The first assumption made - that the ratio of cars traveling to the SR 543 border crossing divided by the number of cars traveling to the Peace Arch crossing is constant, does not appear to be valid. Even when comparing the above ratio data point by data point, as in the ratio method, for July 5, 2014 and July 6, 2014 the mean error between the ratios was 0.503 which is very high considering the mean ratio for July 5, 2014 was 0.842.

The assumption that a lead time of thirty minutes must be provided for the border officers to react could probably be reduced. Thirty minutes was chosen rather arbitrarily since no specifics regarding this aspect of the project were given or determined. The likelihood of finding a usable prediction method would increase dramatically if the lead time was reduced because the noise produced by on-ramps and off-ramps on I-5 would be reduced if a sensor further north could be used for predictions. Note that the shorter lead time would likely not improve any of the methods tested in this project - most methods here were tested with lead times shorter than thirty minutes. A thirty minute lead time necessitates using sensors south of Bellingham to predict cars arriving at the border. This could pose serious problems if Bellingham adds significant noise to traffic patterns on I-5. Also, the noise produced by on-ramps and off-ramps is hypothesized and no assessment of noise production was done here.

3.2 Average Method

To test the Average Method an “average day” was created by averaging data from Saturdays and Sundays from July 2, 2016 to September 25, 2016 from a single sensor. The average was created for each five minute increment, that is for the increment 12:00 am to 12:05 am the “average day” shows the average of all the days in the data set for that specific time period. This “average day” was found to have an average standard deviation for all the time positions of 11.21 cars per period, meaning for each five minute period the uncertainty in the number of cars predicted could easily vary up or down by 14.06 cars. This variation is so high that the Average Method would not make usable predictions.

3.3 Weather Model

The Weather Model was tested using Peace Arch crossing data from April 11, 2015 as an example current day and using the first ten entries as known, then comparing them to all the days in January, 2014 at Peace Arch. Choosing a threshold of 75 for the sum of variances for the first ten entries (the known part of the current day) between the current day and each of the days in the data set. Five days had sums of variances less than 75 so the prediction was created by averaging these five days. The absolute error between the prediction and the actual day (April 11, 2015) had a mean of 105.83 cars per period. This is obviously much too high.

3.4 Principal Component Analysis

The method of Principal Component Analysis (PCA) was tested using the 5th, 6th, 12th, 13th, 19th, and 20th of July, 2008. The macro for Openoffice Calc, OOoStatV0_5 created by David Hitchcock, statistics associate professor at the University of South Carolina, was used to perform the PCA. A sample current day was used and the vectors created using the PCA of the six days above were combined in a linear combination to most closely resemble the first half of the current day. The same linear combination was used to create a prediction for the second half of the day. The mean absolute error was 47.56 cars per period, an improvement but again much too high.

An attempt to include data from previous sensor locations was made by adding the entries from the second location on to the original vector, creating a 576 entry vector. This extended version of PCA was tried on weekend data from August for sensors at the border and sensor AIR25857 using August 13, 2016 as the current day. The average error produced was 93.39 per period. The original PCA prediction had a much lower error.

3.5 Ratio Method

The Ratio Method, and especially the Improved Ratio, were the most promising methods among those tested. Because of this they were the most extensively tested. For the original Ratio Method test the average ratio of each increment was found using weekend days from July 2, 2016 to September 25, 2016. The average standard deviation for each period was 0.30 for the ratios from sensor GRA26611 to the border crossings (see Figure 1). Using the randomly selected day of July 10, 2016, the Ratio Method prediction produced an average difference from the actual border traffic flow of 11.11 cars per period. This is a much lower error than any of the previous methods produced.

3.6 Improved Ratio

After finding that the Ratio Method is the most accurate method tested, adjustments were made in an attempt to improve upon the basic Ratio Method. Using data from sensors AIR25857 and SLA26001 (see Figure 1), the Improved Ratio method was tested. The prediction produced by the Ratio Method for July 24, 2016 for the second sensor was averaged with a best fit sine-cosine function of three terms plus a constant using the first 144 entries as known outputs. The result of the average of these two models produced an average error from the actual traffic of 10.83 cars per period. The success of this result is limited by the fact that these sensors were very close to each other so only one ratio map was needed, while several successive sensors were used when testing the Ratio Method.

3.7 Randomness Checking

To test randomness the weekend data of the combined border traffic flow of July, August, and September of 2016 was grouped into bins of similar traffic days (see 2.7 Randomness Checking). The four resulting bins were used to group days at sensor GRA26611 (see Figure 1) and several two tailed t-tests were done on the average of the entries for the days in each grouping. Only the 100th, 150th, and 200th entries were compared for each day since these data points appeared to be important times to characterize what type of day the data set describes (see Figure 2). This is because knowing those points will give the general shape of the traffic distribution and a rough idea of the peak volume for the day.

The t-tests between the average of each grouping (95% confidence with a null hypothesis that the averages of the two compared groups for each period tested was zero) revealed that for the 100th period groups one through three were not statistically different, but group four was statistically different from the other three. For the 150th and 200th entries no differences were found except that for the 200th periods group two and group four were found to be statistically different.

The same kind of t-test was done on the groupings of the combined border data and the results were very similar; for the 100th period groups one, two, and three were not statistically different, but group four was statistically different from the others. Here the 150th and 200th entries showed no statistical difference between groups.

4 Conclusion

The seemingly simple task of predicting border traffic using data from cars along Interstate 5 proved to be more difficult than anticipated. Traffic analysis is a relatively undeveloped area of math, but models do exist and are currently used that describe traffic velocity and flow. These developed models were not used because the task appeared to be simple enough that a simpler technique would be more appropriate. After attempting to predict traffic flow using the several simple techniques described above, the Average Method produced an average error of 11.21 cars per period, the Weather Model produced 105.83, PCA had 47.56, Ratio Method had 11.11, and Improved Ratio had 10.83. The Ratio Method and the Improved Ratio appear to produce the least errors, but still not significantly better than the basic Average Method. The Weather Model and PCA both produced very high errors.

The randomness of the traffic data was investigated and the results showed that grouping similar traffic days at the border does appear to produce similar groupings at an earlier sensor. In 3.7 it was found that group four was statistically different than groups one through three at the border and at the previous sensor. It was also found that groups one through three were not statistically different at the border and they were not statistically different at the previous sensor, either. An interesting aspect of the results was that the difference of group four only appeared in the 100th period, except for the 200th period on the previous sensor. This exception could be due to the bins being created in such a way as to not find similar traffic days precisely enough. Perhaps some of the days in group two should have been sorted into group four. There is most likely some level of randomness in the data - though the results of 3.7 appear to indicate that it is small - and this randomness could also have contributed to the anomaly. This suggests there should be some method of predicting border traffic accurately. None of the tried methods appear to be the accurate method that is desired.

5 Suggestions for Further Research

5.1 Peace Arch vs SR 543 Relationship

If an accurate prediction technique is found the task remains of finding a better assumption to replace the constant ratio of cars going to Peace Arch vs SR 543. An alternative to finding a better way of relating these two values is to make their relationship more predictable. This might be done by increasing signage from I-5 to the SR 543 border and by more accurately measuring wait times at each border crossing then posting those wait times on the existing electronic wait time signs along I-5. These actions should cause people to choose between the Peace Arch and SR 543 crossings based on wait time alone and not lack of familiarity with the route to the SR 543 crossing or lack of confidence in the posted wait times. If cars do choose which crossing to take only based on wait times, the ratio of traffic volume from one border over the other will become much more predictable.

5.2 Lead Time

An important aspect of this prediction task is to provide ample lead time for the border officials to react. This project was approached by attempting to find a prediction method first and then changing which sensor to use for the input data depending on how much of a lead time is required. If a method could be found using this approach it would be more general, but an easier task would be to establish a lead time and then focus on finding a method. If a more reasonable lead time were found then the task of prediction could focus on the few appropriate sensor locations.

5.3 Target Accuracy

All of the methods tried were deemed unsuccessful because they produced average errors of at least 10 cars per period, but this threshold was also chosen rather arbitrarily. Logically it seems that errors higher than that threshold would cause the queue at the border to grow or shrink too drastically, but perhaps if there are enough checking booths open an unexpected extra 8 cars per minute, say, may not affect the queue. It would be very useful to know how accurate these prediction methods need to be in order to be effective.

5.4 Fast Lane

One interesting suggestion for the project is to investigate if prediction is easier if only the fast lane of traffic is used. The idea here being that there may be noise in the slow lane from cars exiting and entering I-5, but the fast lane may be more consistent meaning more cars in the fast lane are heading to the border.

5.5 Saturday Different than Sunday

An important assumption made at the beginning of this project was that the traffic flow on Saturdays is similar to that on Sundays. It seems logical that weekend data (Saturdays and Sundays) may be different than weekday data, but no effort was made to quantify this difference or to investigate whether Saturdays and Sundays have significantly different traffic characteristics. It may be useful to investigate both of these assumptions.

5.6 Binning Based on Previous Day

In the tried methods any binning step was carried out using the first few entries (usually 144) of the current day. The previous day may be an adequate indicator about how the current day will turn out. If this is the case then binning, say for the Weather Model, could be carried out using the previous day's data - look for historical days that had distributions similar to yesterdays, then for each of those similar historical days see what happened the next day to predict the current day.

6 Tables and Figures

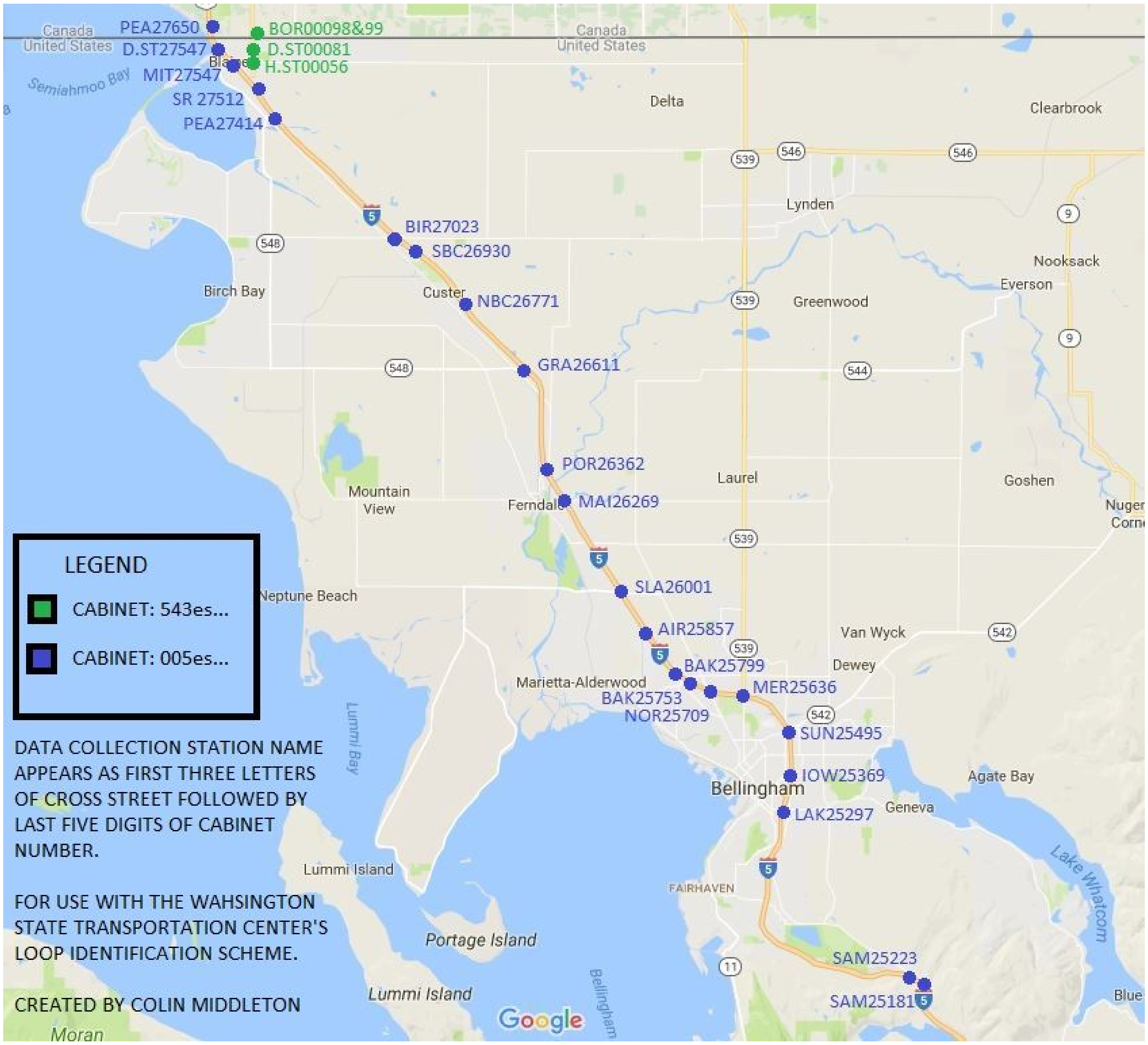

Figure 1. Map of Interstate 5 and Data Collection Sensors. Access was obtained to data from all sensors shown above by the blue dots (I-5) and green dots (SR 543) from 2014 to present.

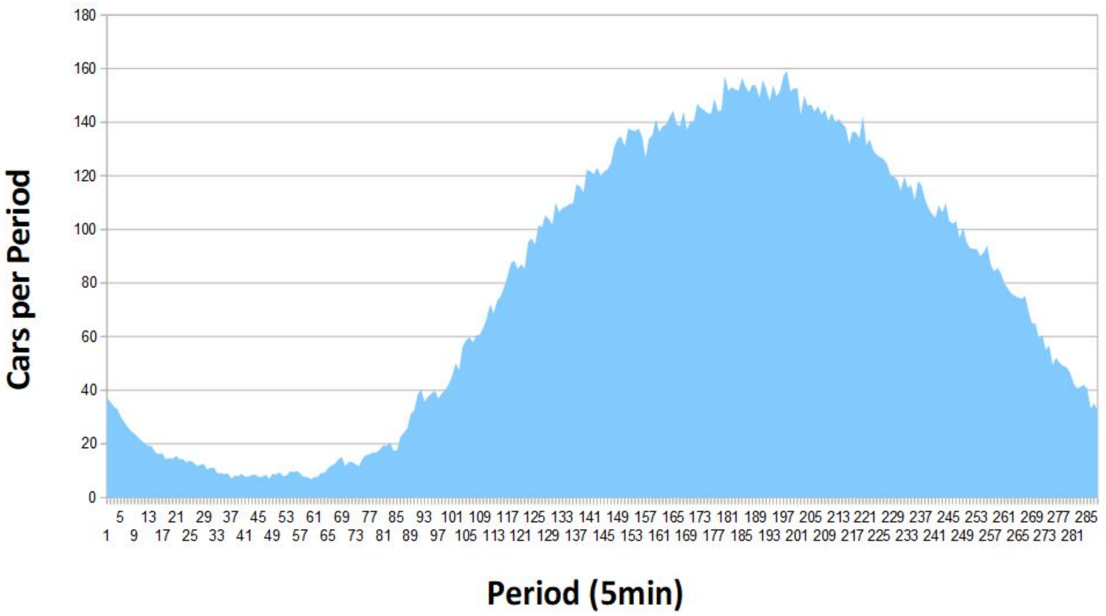

Figure 2. Average Distribution of Traffic Flow. This particular average distribution is for sensor SLA26001 (see Figure 1) for weekends (Sat. and Sun.) in July-September, 2016, and is a typical distribution for a daily distribution though much smoother.

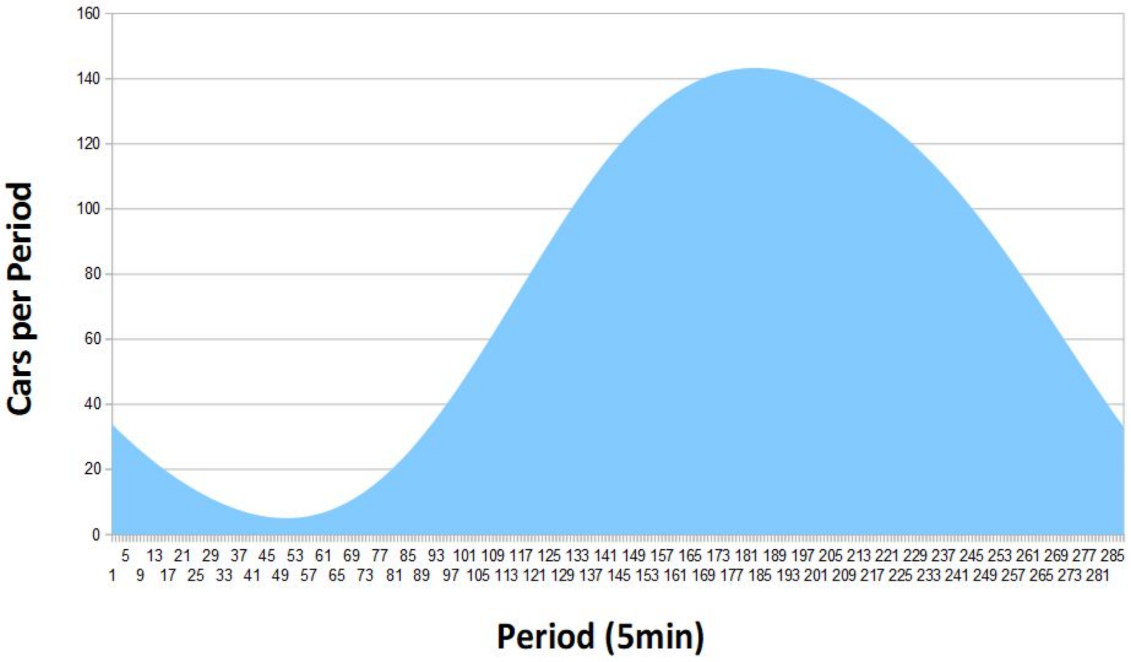

Figure 3. Function Graph Approximating Average Traffic Flow. This particular function approximates the average traffic flow for Station SLA26001 (see Figure 1) for Weekends in July-September, 2016. Compare this approximation to the actual average shown in Figure 2.