UNBRIDGED VERSION OF THE DIVISION REPORT 2018–06 (TR)

Canada Border Services Agency | Agence des services frontaliers du Canada

Science and Engineering Directorate

Border Technology Division

Predicting and Optimizing Border Wait Time Using Artificial Intelligence

Dr. Dmitry Gorodnichy

The publication of this article is enabled by the [Scientific Integrity Policy] released by the Canada Border Agency Services (CBSA) in December 2019 and is done in support of the Government of Canada's [National Action Plan on Open Government and Open Science].

Disclaimer

Opinions presented in this article are solely those of the author and do not represent the opinion of Government of Canada. The work presented in this article was done in 2016-2017 and may not reflect the latest information related to the subject.

Version #: 1.0		
Action	Name	Date
Prepared By	Dmitry O. Gorodnichy	2017-09-22
Reviewed By	Alexandre Desforges	2017-10-05
Prepared By	Dmitry O. Gorodnichy	2017-10-20
Reviewed By	Mehdi Tabib	2018-01-11
Prepared By	Dmitry O. Gorodnichy	2018-04-12

Executive Summary

Artificial intelligence (including machine learning, data science, advanced analytics) has been successfully applied in many social and economic spheres, including weather forecasting, targeted political campaigns, automated detection of targets and events, optimization of system performance, and maximizing profits of various industries. In our work it is applied to predict and optimize Border Wait Time (BWT), which is one of the key performance metrics for the Agency.

A novel scientific approach is developed to allow the Agency to predict and minimize BWT. The approach, made possible through the research and development conducted by the Science and Engineering Lab, consists of two stages. In the first (*Traffic prediction*) stage, hourly rate of vehicle arrival at the border is estimated from the historical data using traditional machine learning techniques such as regression and classification. In the second (*BWT prediction*) stage, BWT is estimated from the predicted traffic as a function of available resources (the number of lanes) using the queuing theory implemented through a computer simulation.

A conventional resource allocation formula used by the Agency such as the one that computes the number of open lanes as a function of the observed queue, LANES = f(QUEUE), is compared to a new recommended formula that computes the number of lanes as a function of *not only* the currently observed queue *but also* of the currently observed wait time: LANES = f(QUEUE,BWT). By using the historical data and proactively allocating lanes, the new formula is shown to be able to lower BWT by over 10 % (from overage of 38.7 to 34.4 mins per hour), while also reducing the total amount of resources by 2% (from average of 3.3 to 3.2 number of lanes used per hour) in a day in a medium size port. Extrapolated for a year, this yields 712 hours saving in wait time, 854 saving in lane-hours, and over a million of travellers experiencing faster service in a single port.

The results are obtained and demonstrated using the in-house developed traffic service simulation program, named *simBorder*. This program is written in R using open-source libraries and based on contemporary software engineering practicesIt is scalable, data-agnostic, and can be run on any Windows machine. As such, it offers a powerful and highly configurable alternative to commercial Resource Allocation Software. Apart from being free, highly configurable and scalable, it can also be integrated into other tools used by the agency, such as *Tailorable Operational Picture Software (TOPS),* within which it is currently being demonstrated to clients. The program can be further improved and extended to address other resource optimization problems of the Agency (such as related to PIK, NEXUS, and FAST programs), either as part of TOPS or as a standalone Windows application.

Theoretical background and practical recommendations related to measuring, predicting and optimizing BWT are presented. Directions for further research and development to advance Agency's capability to minimize BWT and resources are proposed. New metrics for measuring and reporting BWT are derived.

Introduction

Border Wait Time (BWT) is one of key performance metrics of the CBSA. Better BWT leads to trade facilitation and economic growth. BWT is directly linked to the expenses incurred by the Agency and travellers. For example, in 2003 the U.S. Department of Transportation estimated that the cost of delay while crossing the U.S.–Canada border to exceed $13.2 billion every year [1, 2]. In 2014, in response to the Treasury Board Secretariat Policy on Service, the Agency committed to establish a service standard for the air traveller primary processing service as part of its Service Management Strategy. BWT was also part of the Beyond the Border initiative that aimed at implementing automated BWT technology at 20 high-priority Canada-United States land border crossings. Hence there is the strong demand on the Agency to keep BWT low. Specifically, as per the TBS guidelines, the current BWT Service Standard requires the Agency to serve all vehicles within 10 mins from Monday to Thursday and within 20 mins from Friday to Sunday.

To meet this requirement, two challenges need to be addressed. The first one relates to the Agency's ability to correctly measure and report BWT, using either manual or automated technology. The second one relates to the Agency's ability to correctly predict BWT, thus making it possible to optimize it for various traffic patterns and resource constraints. This report presents the outcomes of research, development and analysis (RDA) conducted by the Agency's Science and Engineering Directorate that contribute to resolving both of these challenges.

We show that it is possible to predict BWT using a combination of Advanced Analytics (AA) techniques and queuing theory. The former is used (at the *Traffic prediction stage*) to predict the hourly rate of vehicle arrival from the historical data, while the latter is then used (at the *BWT prediction stage*) to estimate BWT from the predicted traffic and a resource allocation formula. Being able to estimate BWT for different resources formulas, it becomes possible to compare various resource allocation formulas to one another and improve BWT as well as the number of resources (the number of open lanes). The developed approach is implemented in a program, called *simBorder,* which can be used standalone or within other tools used by the Agency, such as *Tailorable Operational Picture System (TOPS).*

In the following, we present the insights gained from results of the analysis of the historical BWT data (Section 2), establish BWT metrics (Section 3), describe the novel BWT prediction approach (Section 4), introduce the *simBorder* program that implements this approach (Section 5), and finally show the simulation results (Section 6) and provide recommendations for next steps (Section 7).

2. Analysis of historical data

A. Overview of data

The BWT data for this work were provided by the Northern Ontario Region (NOR). They consist of hourly BWT records recorded from January 2010 till May 2014 (20,000 - 40,000 BWT records per port), from twenty six ports. All BWT records were measured and entered manually (i.e., without the use of automated BWT measuring technology. The following observations are made. More detail of this analysis are provided in Appendix A.

INSIGHTS

- BWT ranges significantly for all ports: from 10 minutes to over 3 hours. The shape of the BWT distribution is similar for all ports. However, several outliers (rare but extremely large BWT values) are also present in several ports.
- Additionally, all ports have many missing BWT values. The number of missing entries is almost the same as the number of all recorded entries combined.

In addition to BWT for commercial and traveller lanes, the following data were also provided for analysis: number of open lanes (LANES), number of cars (CARStotal.[1]), number of Passengers (PAXtotal), the percentage of these

[1] For simplicity of presentation, the variable names have been renamed, so CARStotal is referred to as CARS later in the report.

that have been referred to secondary (CARSReferralRate and PAXReferralRate), the number of full-time officers on duty (FTE_WORKED) – all also recoded for each hour of the day. The rate of US dollar (USD) was also provided. Except for the CBSA internal data related to staffing and referrals, these data are similar to those used in academic work [3], which were extracted from public BWT data warehouses [6].

These data came from different sources and as a result had rather small time overlap. Only seven ports had full overlap of key data variables and only for short period of time (less than three months, in spring - summer of 2013). Nevertheless, this amount of data was sufficient in order to gain critical insights related to the BWT improvement problem and develop a novel scientific approach to help the agency to address this problem. See Appendix A for more detail.

INSIGHTS:

- The shape of daily Conveyance patterns (CARS) are similar for all ports, and can be roughly approximated using the cosine functions with maximum at 2 pm and minimum at 2 am.
- In contrast, the shape of BWT and LANES daily patterns are very different for each port and cannot be easily approximated or predicted from one another.
- With approximately the same traffic patterns, some ports have much lower BWT than others.

B. Factors affecting BWT

The first step in examining the relationship between the variables of the historical data consisted in examining their correlations. The following observations are made. Appendix B provides more detail on this analysis.

INSIGHTS:

- LANES highly correlate with CARS (0.78). Correlation between BWT and CARS or LANES however is rather weak (< 0.3)
- BWT does not correlate at all with referral rates or FTE (< 0.05)
- BWT varies greatly by day of the week (WD), as do other variables, including the referral rates.

C. Examination of lanes as a function of car volume

The next research task was to understand the relationship between the number of allocated lanes (LANES), traffic volume (CARS) and BWT. Frontline officers have indicated that a linear relationship between LANES and CARS is currently commonly assumed, i.e., LANES = K * CARS, where K is constant.

To investigate this relationship, a linear regression was applied to compute the constant K = CARS/LANES as function of LANES for each month and week day. The following observations are made. See Appendix C for more detail.

INSIGHTS:

- The ratio K = CARS/LANES increases (from 10 to 60) with number of CARS and decreases with number of LANES.
- This shows that, historically, there are fewer LANES than are required to process large numbers of CARS. The ratio K is also shown to vary by month and day of week.

This indicated that for BWT minimization purposes, the lane allocation formula should be the function of not only CARS (as it is now), but also of BWT or queue length. This conjecture will be proven using simulations.

3. Analysis of BWT metrics

In order to build an algorithm for predicting and optimizing BWT, the first task is to define BWT metrics and the ways to measure it. Different ways of measuring BWT have been used at the Agency throughout the years, with different limitations and precisions. These are analyzed next.

A. Two ways to measure BWT: at Arrival vs. at PIL

According to its formal definition [4], BWT is the time it takes, in minutes, for a vehicle to reach the primary inspection lane (PIL) booth after arriving at the end of the queue.

As shown in Figure 3-1, there are two possible ways to "measure" BWT, which may produce very different results for the same traffic pattern:

- Way 1 (estimated at Arrival): BWT for *all* cars is *estimated* based on the queue length, by looking at the last arrived car. This is the most frequent way used by the Agency: done by using landmarks (when done manually) and by using inductive loops (when done using technology).
- Way 2 (estimated at PIL): BWT for *all* cars is *estimated* by taking measurement from some cars: done by asking drivers (when done manually) and by using Bluetooth or License Plate Readers (when done using technology).

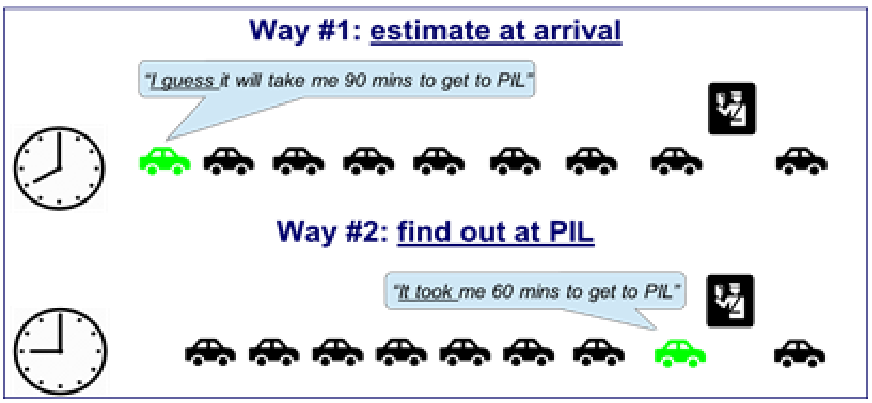


Figure 3-1 Two ways to measure BWT.

B. Three major limitations of BWT records

There are three inherited deficiencies in the currently recorded BWT data:

- Reported per hour, not per car: By its definition, BWT should report the average BWT per vehicle. In practice, however the BWT is recorded averaged per unit of time (hour). Therefore, it is not possible to induct the percentage of vehicles or travellers that experienced long or short BWT.
- Precision is not known: Daily BWT records contain a list of 24 numbers (one BWT measurement per hour). However it is not known how or when (at what exact time during this hour) these numbers were measured. Their credibility and correctness is also not known. Due to human factors and technology limitations, BWT will inevitably contain errors, the percentage of which has not been studied yet.
- Estimates are based on assumptions which often do not hold: Either of the above described ways of measuring BWT actually does not measure it, but rather provides an *estimate* of it based on certain assumptions, which in fact, often do not hold. For example, if the processing rate decreased (or number of lanes increased) during an hour then the reported BWT will be higher than actual. Similarly, if the car arrival rate has changed during an hour or the car whose BWT was used to extrapolate BWT for the entire hour had some particular issues, this will also skew significantly the BWT estimate.

INSIGHTS:

- The BWT data that was provided for the analysis were measured manually using either of two techniques (at PIL and at Arrival), however we do not know which technique was used when and how it was used.
- BWT is estimated on assumptions which may not hold, its precision is not known, it is reported per hour rather than per car. As shown in Figure 5-4, errors over 30 minutes in BWT values are possible, simply because of using different ways of measuring BWT.
- For reasons mentioned above, real-life data may not be used for evaluating BWT optimization strategies. In contrast, the data obtained using the simulation can be used to do that.

C. Advantage of analyzing BWT using simulation

Analyzing BWT using simulation offers three important advantages over traditional analysis based on real-life operational data.

- While the above described limitations of BWT records exist in a real-life operation, none of them exists in simulation environment.
- Furthermore, simulation allows one to measure BWT using all metrics both at Arrival and PIL.
- Finally, because simulation computes BWT per car (not per hour), it allows one to obtain the exact objective measures of the BWT performance - for all cars and travellers crossing the border, which is the eventual objective of border service performance evaluation.

INSIGHTS:

- In real operation, it is not possible to measure BWT precisely for each car. Instead, estimates may only be used for reporting BWT. Two main techniques for obtaining such estimates are extrapolation (At Arrival) and sampling (At PIL). Each one has its own limitations and errors.

- In contrast, in simulation BWT can be computed exactly for every car, which can be used to provide evidence on service performance improvements

D. Estimating BWT at given hour period from BWT per car

In order to compare simulation results, which compute *BWT per car*, to operational results, which record *BWT per hour*, three methods of increasing precision for estimating the latter from the former are developed:

- <u>By sampling (0-order estimation)</u>: e.g., on top of the hour. This mimics manual entries.
- <u>By 1st order statistics (1-order estimation):</u> e.g., Median, (truncated) Mean/Average. This mimics automated entries.
- <u>By histograms (2-order estimation):</u> X% of cars processed within N minute. This mimics the best practices followed by in EU in conducting the evaluation of eGates performance.

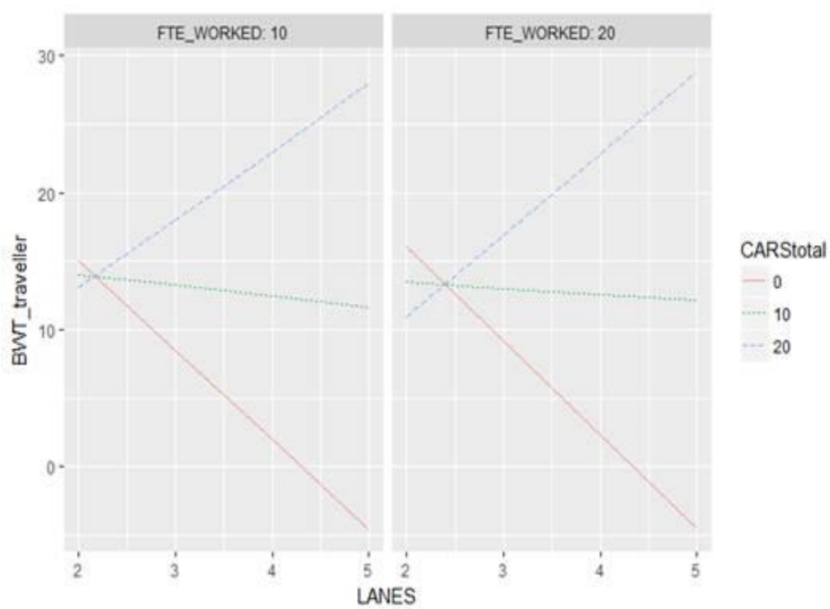
All these metrics are implemented in the BWT prediction methodology described next.

4. Methodology for BWT prediction

A. The fallacy of traditional predictive modeling approaches

Machine Learning (ML) techniques, colloquially called Predictive Modeling by data scientists, attempt to predict the output of the system for input data that has not been previously seen. The most popular of these include neural networks, decision trees, regression model and the boosting combination of them (when many techniques are run in ensemble and the best one is chosen depending on the input data). These were the techniques that we originally suggested for this work, however later found inappropriate for the reason described below.

Figure 4-1 shows the result of a ML technique (linear regression) that was applied to “predict” BWT from the historical number of Lanes, FTE, HH (hour of the day) and MM (month of the year). The level of significance of each regression term is also shown – all terms are over 99.9% significant. Without proper consideration to the nature of BWT problem, one may be trapped to make wrong conclusions from such “prediction”. As seen in the figure (blue top line) , the model “predicts” longer wait for larger number of lanes, which is what the model leant from historical data. This is however the opposite of what common sense is telling us. Time to introduce queuing theory!




```
Deviance Residuals:     Min      1Q      Median      3Q     Max
                       -4.84   -3.13   -1.04       1.06   40.01

Coefficients:     Estimate Std. Error z value            Pr(>|z|)
(Intercept)        4.229100   0.109430   38.65 < 0.0000000000000002 ***
LANES             -0.221182   0.028513   -7.76   0.0000000000000087 ***
FTE_WORKED        -0.028692   0.005116   -5.61   0.0000000204235935 ***
HH                 0.007158   0.000675   10.60 < 0.0000000000000002 ***
MM                -0.032495   0.000925  -35.13 < 0.0000000000000002 ***
WD                -0.025702   0.001141  -22.52 < 0.0000000000000002 ***
LANES:FTE_WORKED   0.008256   0.001330    6.21   0.0000000005446924 ***
```

Figure 4-1 Results of using regression to predict BWT: BWT depends significantly on each variable and increases with LANES for large CARS

INSIGHTS:

- Predictive modeling can be used for predicting the unseen traffic patterns from previously seen historical traffic data. However, it cannot be used for predicting BWT from historical BWT data, because historically, more lanes mean that there was larger BWT. Queuing theory needs to be used for this task.

B. Queueing theory

According to the queueing theory [5], BWT is a function of three variables: Arrival Rate (AR), Processing Rate of a server (PR) and the number of servers (S) and can be computed theoretically for a given set of AR, PR and S numbers. It is therefore important to realize that one cannot estimate BWT from PR only. One also needs to know AR to do that. This is further explained below.

If there is no waiting line (BWT = 0), then AR = PR. However if BWT > 0, then it means that more cars are arriving than can be processed, i.e., AR > PR, and that cars need to wait in a line. In the latter case, the wait lines will keep growing and BWT will keep growing until either a) the car arrival rate becomes smaller than the processing rate (AR < PR), or b) new servers are added and the system reaches a stable state (meaning that queue length remains approximately the same).

Wait Time Calculators, which allow one to compute (based on theoretical formulas) the number of servers required and the expected BWT for a given number of AR and PR, are freely available on the Internet [6]. These calculators however cannot be used for complex scenarios when AR dynamically changes with time, which is what happens at the border. Hence we moved forward to build a simulation program that will compute this information for us for any border configuration we need.

C. Two-step BWT prediction approach

The proposed approach for predicting BWT is defined as a two–step procedure:

- Step 1 (predict traffic): Estimate hourly car Arrival and Processing Rates (AR and PR) for the day from historical data using time-series prediction techniques such as exponential smoothing (a.k.a. "Holt-Winters approach), autoregressive integrated moving average (ARIMA), or naïve historical average.
- Step 2 (compute performance): Once hourly AR and PRs are estimated for the entire day, BWT is computed as a function of the dynamically changing number of LANES (i.e., lane allocation formula) using queueing theory and simulation.

The technical details on how these steps are implemented are outside of the scope of this report[3].

D. Performance metrics for BWT optimization

For evaluating the performance of simulated port, two key metrics are used:

---

[3] In brief, for Step 1, any traditional machine learning (predictive modeling) technique can be used to predict hourly PR from historical hourly conveyance data (CARS). In our work several such approaches have been tried, including exponential smoothing (a.k.a. "Holt-Winters approach), autoregressive integrated moving average (ARIMA), or naïve historical average. Their choice was not found to be critical. A procedure for estimating hourly AR is more complicated. It involves reverse-engineering, using historical data, in such a way that to produce PR that is similar to the one historically observed. For Step 2, queueing theory permits computing BWT theoretically, i.e., using a deterministic formula, as function of available resources. In our work however, we opted for using the simulation to do that, as it allows easier modification of scenarios and better visualization of the results obtained.

- Average hourly BWT, which is computed from BWT measured for each car at Arrival and PIL, using the "top-of-the hour" and averaging approaches (as described in Section 3)
- Average hourly allocation of Lanes (LANES), as selected by the program according to the lane allocation formula selected by the user, which is further described below.

Additionally, the program records several other performance indicators such as: average hourly queue length, number of cars arrived and processed, and the percentage of cars that were processed within BWT service standard (i.e., within 10 or 20 minutes).

E. Lane allocation formulas

First, we examine the lanes allocation formulas that are traditionally used in border operation, namely such that allocate LANES depending on queue length (QUEUE) or the traffic volume (CARS). Then we propose the improvement to those formulas by also taking into account the actually observed BWT and/or the queue length. :

Formula 1 (Reactive formula): LANES are computed as a function of observed queue (CARS):

LANES = F(QUEUE), e.g., LANES = QUEUE / 20 cars (1)

Formula 2 (Proactive formula - simplified): LANES are computed as a function of both QUEUE and BWT:

LANES = F(QUEUE, BWT), e.g., LANES = QUEUE / 40 cars + BWT / 20 mins (2)

Formula 3 (Proactive formula - generic): LANES are computed as a weighted linear combination of: car volume (CARS), observed queue (QUEUE), current BWT and acceptable BWT (targetBWT)

LANES = A/40*CARS + L/40* QUEUE + B/40*(BWT - targetBWT) (3)

By varying the constants (e.g., A = 1, L = 0, B = 1, targetBWT = 20), one is be able to compare multiple lane optimization strategies to one another. This can be done using the developed simulation program.

INSIGHTS:

- The key conjecture behind the new (Proactive) formula is that: the larger BWT, the more LANES should be added (in addition to those already defined by queue length) - so as to shorten BWT faster. The proof for this conjecture is obtained via simulation.

Analogy:

When fighting a fast-spreading fire, in order to extinguish it *faster*, the amount of water is a function of 1) the size of the fire, and 2) how fast fire is spreading. *Similarly, BWT can be minimized faster and with less resources, if the resources (LANES) are allocated proactively by taking into account the currently observed BWT.* This is the essence of the proactive approach for allocating LANES.

5. *simBorder* program

In the following we briefly describe the *simBorder* program (Figure 5-1), starting from the problems that needed to be addressed and finishing with the code that integrates it into the *Tailorable Operational Picture System (TOPS)* and the results obtained by applying *simBorder* to compare the efficiency of two different lane allocation formulas.

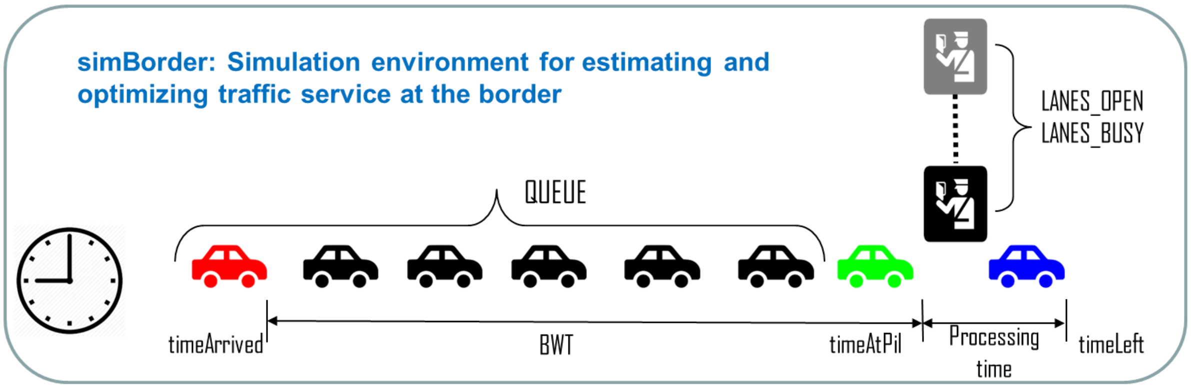


Figure 5-1 The simBorder simulation program generates virtual "Cars" and "Officers" in order to measure and optimize various border control metrics.

A. Problems to be addressed

The following problems related to building the intelligence behind the *simBorder* needed to be addressed:

- It must simulate Cars arrivals in such a way that it mimics the historical conveyance data.
- It must simulate the performance of Border Services Officer (BSO) so that it matches the feedback provided by BSOs
- It should include all possible "real-life" 24 / 7 / 365 scenarios (for each Time of day, Day of week, Time of the year)
- It must define "realistic" performance metrics and the way to measure them, including confidence intervals.

In addition, the following software engineering challenges needed to be considered:

- Code must be data agnostic, modular, scalable, and highly configurable, permitting the development and usage of multiple versions for difference scenarios.
- Because it is embedded into end-used run-time software (TOPS), it must not crash.
- Entire cycle for Software Development principles must be used, including use of source control and proper tools for debugging, forking and merging different variations of the code.
- Requires advanced and up-to-date knowledge of R and related R packages for statistical analysis and data visualization
- Requires expertize in Object-Oriented Programming in R so that the entire simulation environment can be implement as a Class (Object) which can be incrementally modified and interrogated by developer and user using intuitive get_variable() and set_variable() functions.

While most of these challenges have been overcome, some still require further work and refinement.

B. Simulating Car Arrivals and BSO work using Probability Distribution Functions

Two types of "events" need to be generated for simulation: Cars and Officers. The objective of the simulation is to be able to randomly generate the time that it takes for a new car to arrive ("CAR" event) and for an officer to process a car ("Officer" event) so that it mimics the historical data and matches the feedback provided for this work by frontline officers.

In order to do this, the following probability distribution functions (PDF) were tested: Poisson, normal, exponential, and Erlang(X), which is a sum of X several exponentials, with a range of various parameters, the most important of which are the mean expectation and deviation of the PDF. Figure 5-2 compares outcomes of several PDF to one another. Based on reverse engineering based simulations, Erlang(20) PDF has been chosen as the closest to the real-life scenario.

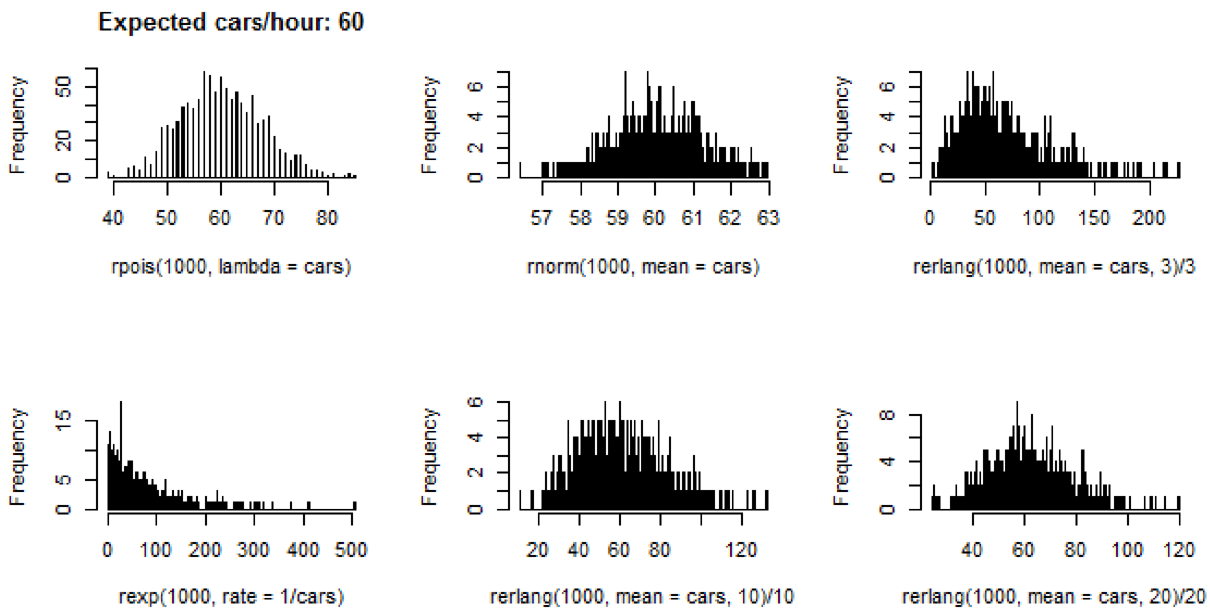


Figure 5-2 Probability distribution functions (PDF) that were tested for their applicability in simulating car arrivals and officers work: Poisson, Normal, Exponential, Erlang(20), Erlang(10) , Erlang(3).

For each PDF, the image shows the expected frequencies of possible outcomes for the simulated event. The event is the number of seconds between the cars (for car arrival simulation) and the number of seconds it takes an officer to process one car (for officer work simulation).

C. Three-step simulation:

The execution of the program is performed in the following three steps (see also Figure 5-5). The input to the program is marked by (*).

Step 1: Defines and monitors “resources”:

- Defines number of Lanes *, which can be changed at any time according to lane allocation strategy
- Defines processing times *, which is the min, max, average time it takes for an officer to process a car
- Follows each resource, computing all related statistics any time is it used

Step 2: Defines and monitors “cars”:

- Generates “cars” - according to time of day, day of the week/year *
- May assign “attributes” to “cars”: Number of passengers, citizenship, etc.*
- Creates “trajectory” for each car
- Follows each car along its “trajectory”, computing all related statistics

Step 3: Runs for 24 hours and aggregates all statistics about “resources” and “cars”:

- BWT/car, BWT/hour – computed using metrics described in Section 3-D.
- Resources used, Queue length

An example of the logged output produced by the simulation program is shown in Figure 5-3. It shows exactly what is happening in a simulated border crossing with every “car” and every “resource” at any moment of time. Based on these logged output, the *exact* BWT and resource allocation metrics are computed in each simulation run – for each hour and each car using any of six metrics described in Section 3-D (i.e., sampling, average, histogram at Arrival and at PIL). Because all of these metrics are exact (i.e., theoretically correct), one can compare these metrics to one another, as done in Figure 5-4, which shows the difference in average hourly BWT when two different ways of measuring BWT are used. As seen from the figure, such difference – now proved theoretically - can be quite large (over 30 mins).

```
      name timeArrived timeAtPil timeLeft BWT ProcessingTime
 1:  Car0     0.0131    0.0131  0.0413    0  1.70
 2:  Car1     0.0376    0.0413  0.0839    0  2.55
...
20:  Car19    0.4064     0.6490  0.6774  14  1.70
21:  Car20    0.4252     0.6774  0.7184  15  2.46
...
2710: Car2698   23.8    23.9   24.0    9   2.44
2711: Car2699   23.8     NA     NA     NA   NA
2712: Car2711   24.0     NA     NA     NA   NA
```

```
time    QUEUE LANESOPEN LANESBUSY

0.0131     0      1      1
0.0376     1      1      1
0.0413     0      1      1
...

23.9    10       3       3
24.0    11       3       3
24.0    10       3       3
```

Figure 5-3 Log output produced by the simulation:
shown are the metrics measured for “cars” (left) and for “resources” (right)

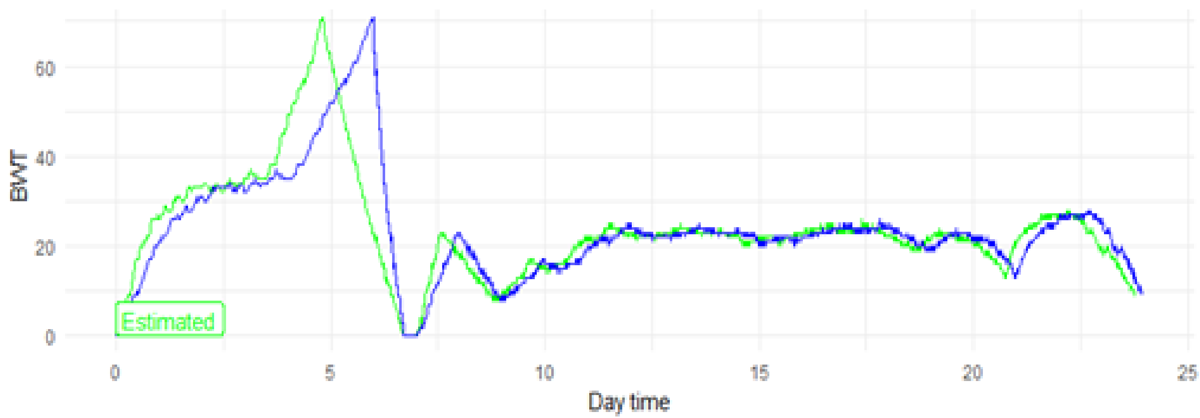


Figure 5-4 BWT per car computed using two different metrics: measured at Arrival (green) vs. measured at PIL (blue). A difference of over half an hour in reporting BWT using two different metrics is possible.

D. *simBorder* as part of Tailorable Operational Picture System (TOPS)

The *Tailorable Operational Picture System (TOPS)* is a data visualization technology that is currently being developed by the Agency and its partners through the support from Canadian Safety and Security Program (CSSP) managed by the Defence Research and Development Canada, Centre for Security Science (DRDC-CSS) [3]. It integrates a number of data sources and provides a flexible and easily understood picture of historical trends, and near real-time patterns related to traffic volumes, enforcement actions, staffing levels and other elements that affect border management operations. The current prototype contains data obtained from NOR and open sources to test, but is designed for complete national coverage.

In effort to further improve the Agency's understanding of how more advanced techniques can support frontline operations, the *simBorder* program has been integrated into the TOPS prototype.

INSIGHTS:

- A number of success stories related to BWT optimization have been reported by industry [2] and academia [3]. Now the Agency has its own tool, albeit in its early development, and the knowledge to address this problem, thanks to its in-house research and development capability.

The architecture and the output of *simBorder* within TOPS are shown in Figure 5-5. The parameters that need to be passed to *simBorder* by the operator in order to simulate port's operations are the following:

- Port ID and day, which are used to predict hourly arrival rate for the day,
- Lane Allocation formula (such as Eq. 1,2,3),
- "Hard" constraints, such as min/max number of lanes in a port and min/max/average time it takes an officer to process a car ,
- "Soft" constraints (targets), such as desired number of lanes and BWT.

The selection of these parameters is one of the most important steps to ensure realistic simulation. They need to be obtained in close consultation with end-users knowledgeable of port operation.

```
simTOPS <- cSimBorder$new()

simTOPS$setFormulaToAdjustLanes
   (timeStep=60, A=1, B=1, L=1)
simTOPS$setPort(POE= 4560)
simTOPS$setToday("2013-06-22",
   bHoliday=TRUE)
simTOPS$setConstraintLANES(
   MAX=6, MIN=1)
simTOPS$setConstraintProcessing
   Time(MIN=0.5, AVE=2, MAX=4)
simTOPS$setArrivalVolume(
   forecastedArrivalVolume)
simTOPS$setTargetLANES(trLANES)
simTOPS$setTargetBWT(targetBWT)

runSimulation(RANDOM_SEED=1,
    RANDOM_SIMULATIONS=10)

simTOPS$getBWT(round=5);
simTOPS$getLanes();
plotSimResults()
```

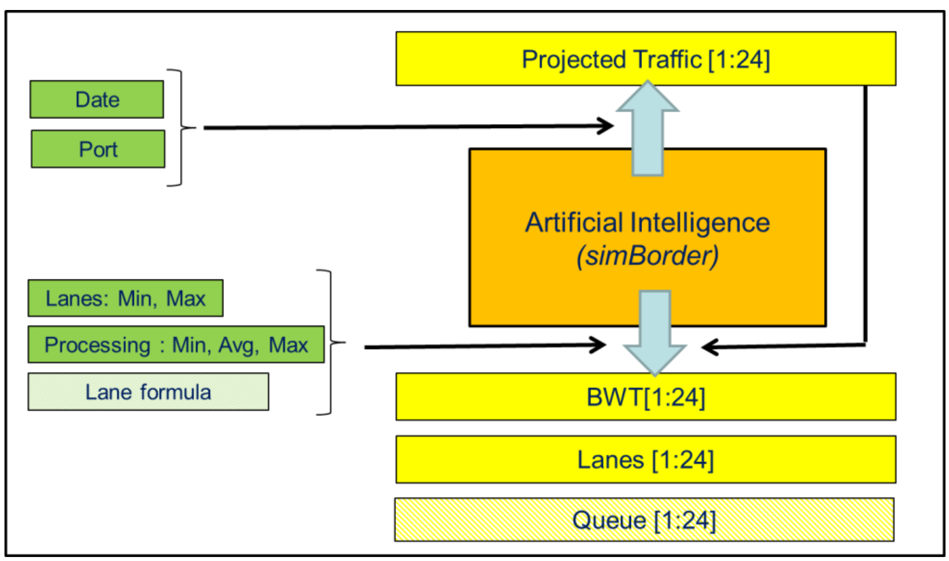


Figure 5-5, Architecture (top) and output (bottom) of the simBorder program within TOPS "Planning" window.

6. Experimental results

To demonstrate the value of the developed approach, we show the results of applying it to predict the performance in one port (Lansdown, ID = 4550) for one day (Sunday, 22 June 2013).[4] The simulation parameters (shown in Figure 5-5) of *simBorder* program are tuned so that to mimic the historical car arrival data on that day.

A historically used formula for allocating resources (Eq. 1), which computes the required number of open lanes as a function of the observed queue, LANES = f(QUEUE), is compared to the new recommended formula (Eq. 2), which makes the required number of lanes as a function of not only the currently observed queue but also of the currently observed border wait time: LANES = f(QUEUE,BWT), as per insights presented in Sections 2 and 4.

The result is shown in Figure 6-1, where the graphs show the measured *hourly* metrics, computed using both lane allocation formulas. As seen in the figure, Formula 2 (Proactive formula) performs better than Formula 1 (Reactive formula).

[4] Similar results are observed for other ports and days, but not shown in this report for briefness of presentation.

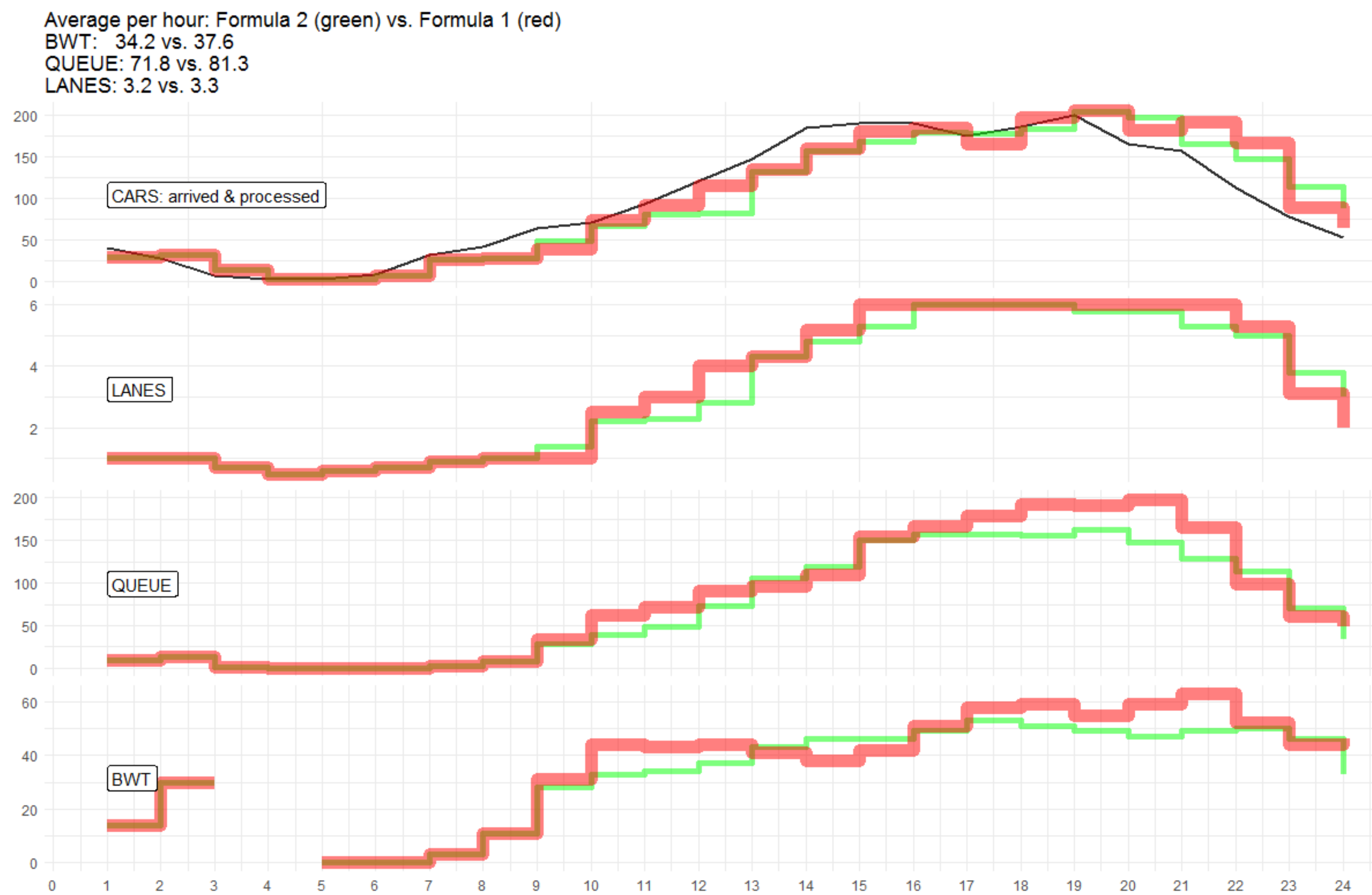


Figure 6-1 Simulation results: a snapshot for Sunday, 22 June 2013, Port 4560 (Lansdown). CARS processed, BWT, LANES, QUEUE - measured for Formula 1 (red) and Formula 2 (green). For the same hourly CARS arrival rate (black line), Formula 2 is seen to perform better than Formula 1.

INSIGHTS:

By using a new proactive formula for allocating resources, the following performance improvement is observed for a middle-size port (see Figure 6-1):

	Average BWT / hour	Average LANES USED / hour
Formula 1	37.6 mins	3.3
Formula 2	34.2 mins	3.2

For one day:

- 10% improvement in BWT and 2% improvement in LANES

Extrapolated for one year:

- 712 hours saving in wait time
- 854 saving in used lane-hours
- Over a million of travellers experiencing faster service.

A. Limitations of the results

The results presented above should be treated only as a guide and not as any actual proof of any particular port performance, since they are obtained using the simplest possible model to simulate the traffic service performance of the Agency – all lanes are the same, all travellers/cars are the same, nothing changes in the environment compared to the historical data that was used in predicting traffic patterns and officer throughput. In reality, the processing is much more complex - there are different types of lanes and different types of travellers and cars. Lane allocation rules may also be different.

Furthermore, the precision and reliability of historical BWT data, which was used in building the models, are also unknown. As mentioned in Section 3, BWT value may easily have a margin of over 30 mins, depending on the BWT metrics that were used. It was only recently, that a new BWT reporting methodology [8] (also developed by the Science and Engineering Directorate) has been approved by the Agency to permit more precise tracking of BWT performance. The BWT data obtained prior to applying the new BWT reporting methodology may substantially be off from the actual delays experienced by travellers and could not have been easily compared to one another.

Finally, reverse-engineering of Arrival Rate data from historical Conveyance (Processing Rate) may also lead to false assumptions. Ideally, Arrival Rate data should be measured first hand at the port, as should be all other input parameters (discussed in Section 5) that are passed to the *simBorder* for processing.

7. Recommendations for next steps

The limited duration of this project (half a year) and scientific resources allowed us only to “scratch the surface” of what is possible through the research and development applied to such an important operational problem as BWT optimization.

In the following the recommendations on how to further leverage and extend this work are presented. Two complementary directions for next steps are seen: one related to applying / deploying the recommendations and tools already developed; the other on adding more scientific rigour and technical complexity in developing more realistic simulations within the *simBorder* program.

A. Analysis of more historical BWT data

Up till now, the data from only seven NOR ports have been analyzed and during only a rather short period in spring-summer five years ago. Much could have been changed since then. Historical data from different regions and different seasons could uncover additional insights on the Agency’s performance and used for better BWT predictions and simulations. The visualization tools that were developed through this work (such as those used to obtained graphs shown in Appendices) can be used to analyze such data - for trends, abnormalities, performance improvement opportunity analysis, prior to their integration into the TOPS application.

B. Field trial of *simBorder*

To validate utility of *simBorder* program, it should be tried in the field. It can be installed as part of TOPS or as a standalone Windows application on a regular laptop. This would allow end-users to compare predicted results with the results observed and provide feedback to improve the quality of simulation. In such a way, the program’s interface, precision and functionalities can also be improved.

C. Field trial of new proactive lane allocation formula

This work shows that it is possible to improve service at the border by simply proactively allocating lanes, so that BWT is reduced *faster*. In doing that, the number of required allocated lanes can also be reduced. While obtaining the best optimization formula still requires substantial scientific research effort, the formula that has been described in this report (Eq. 2) that makes the number of open lanes (LANES) a function of not only the Processing rate (CARS) but also the currently observed BWT is expected to improve the performance right away. Implementation of the formula in the field is therefore recommended, provided that proper, scientifically correct, BWT reporting methodology is used for measuring and reporting BWT. Prior to its country-wide implementations, it can be first piloted in one port, for a period of at least three months.

D. Improvement of the BWT reporting methodology

Regardless of how much better the Regions will be able to improve BWT using the new proactive lane allocation formula, it will not be possible to measure and report any of it until the Agency improves its BWT reporting methodology. It has been only recently that a new, scientifically correct, BWT reporting methodology has been developed by the Agency, through the work of Science and Engineering Lab from the request of Programs branch [8]. This work however has not been finished yet, as there still a number of ways to improve the BWT reporting methodology. The results presented in this report should further facilitate doing it.

E. Analysis and optimization of other border control scenarios and programs

The in-house developed *simBorder* simulation program has a potential to offer a very powerful and highly configurable alternative to commercial Resource Allocation Software, such as those overviewed in [9]. Apart from being free, it allows CBSA to integrate the code into other end-user tools such as TOPS. The *simBorder* program may be further extended to address practically any resource optimization problem of the Agency. This includes the use of primary inspection kiosks (PIK), expedited commercial and traveller lanes (FAST, NEXUS), and the use of other Automated Border Control (ABC) technologies such as those studied in [10] that the Agency may be interested to consider as part of its border modernization strategy.

F. Better prediction and more realistic simulation

The *simBorder* tool that has been developed is already powerful enough to make useful recommendations to the border officers. However, it may still be improved and enhanced in a number of ways:

- Improving prediction of traffic patterns using a combination of short-term forecasting and long-term prediction
- Developing more realistic behaviours for its virtual objects (“cars” and “officers”)
- Obtaining more precise estimates and confidence intervals
- Extend the code to include more complex scenarios, such as:
    - Different types of travellers (commercial vs. travellers, domestic vs. international, low-risk / trusted vs. unknown risk, etc.)
    - Different types of traffic (different lanes configurations, different size Ports)
    - Different border control technologies (e.g., biometric-enabled, other automated technologies)
    - Different modes of operation (Air, Land)
- More robust, faster, easier to use code for the core part of the program
- Development of table-top end-user interfaces (using open-source programs such as Shiny)
- The use of relevant open source tools

Many open source tools are now being developed by the research community, which the Agency may use to enhance its own research and development capability in this area. Additionally various academic groups, including from the University of Ottawa , are interested in providing research contribution to the above described problems, which may also be leveraged by the Agency.

G. Optimal resource allocation to minimize BWT

Of particular research interest is the problem of finding the best resource allocation formula for each port and each possible border services scenario. This can be done theoretically using the queueing theory or via simulations. Mathematical optimization models from the Operations Research field should also be investigated.

References

1. Taylor, J.C., Robideaux, D., Jackson, G.C., 2003. The U.S.–Canada Border: Cost Impacts, Causes, and Short to Long Term Management Options. Technical Report. Grand Valley State University, Allendale, Michigan
2. CATSA - OSIR - 28 August, 2017, http://www.businessinsider.com/how-a-danish-startup-is-using-math-to-cut-airport-wait-times-in-half-2017-8 , How a Danish startup is using math to cut airport wait times in half.
3. Mengqiao Yu, Yichuan Ding, Robin Lindsey, Cong Shi. A data-driven approach to manpower planning at U.S.–Canada border crossings. Transportation Research Part A 91 (2016) 34–47
4. Ten science and technology projects to be funded through the Canadian Safety and Security Program (CSSP). 30 July 2015. https://www.canada.ca/en/news/archive/2015/07/government-canada-invests-canada-safety-security-1011829.html
5. Online resources on Queueing Theory:
   - Khinchin, A.Y., 1967. The Mathematical Theory of a Stationary Queue. Tech. Rep. DTIC Document.
   - Hall, R.W., 1991. Queueing Methods: For Services and Manufacturing. Prentice-Hall International Series in Industrial and Systems Engineering. Prentice, Hall, Englewood, Cliffs, NJ.
6. Online wait time calculators: http://www-ens.iro.umontreal.ca/ ~chanwyea/erlang/.
7. Public Border Wait Time Data Warehouses: http://www.cascadegatewaydata.com
8. D. Gorodnichy, *"Border Wait Time (BWT): Options for reporting* ". Presentation to the Advanced Analytics Sub-Working Group, May 2, 2017, D. Gorodnichy, *"Annex 1-Border Wait Time (BWT) - Options for Reporting.pdf*" prepared for Briefing Note "ADJUSTING THE HIGHWAY BORDER WAIT TIME REPORTING METHODOLOGY", August 2016.
9. Dmitry Gorodnichy. *Simulation tools for e-border*, CBSA S&E Memorandum, March 30, 2015, Memo DG20150327 – Simulation tools for e-border.pdf (http://sd01cuvap0008.omega.dce-eir.net/livelink/llisapi.dll?func=ll&objaction=overview&objid=1805812)
10. Dmitry O. Gorodnichy, "ART in ABC: Analysis of Risks and Trends in Automated Border Control. CSSP-2013-CP-1020 study". Executive Summary: http://cradpdf.drdc-rddc.gc.ca/PDFS/unc229/p803869_A1b.pdf . Full Report: http://cradpdf.drdc-rddc.gc.ca/PDFS/unc256/p804885_A1b.pdf

Appendix A: Graphical analysis of historical data

This appendix showcases a number of graphic analysis results obtained during the course of this study. The code has been written in R to automatically generate such graphs for arbitrary BWT and associated data at any port and any day. Such codes can be potentially integrated into TOPS for on-line analysis of any historical data recorded by TOPS. Additionally, they can be used for future analysis of historical data on any standalone computer.

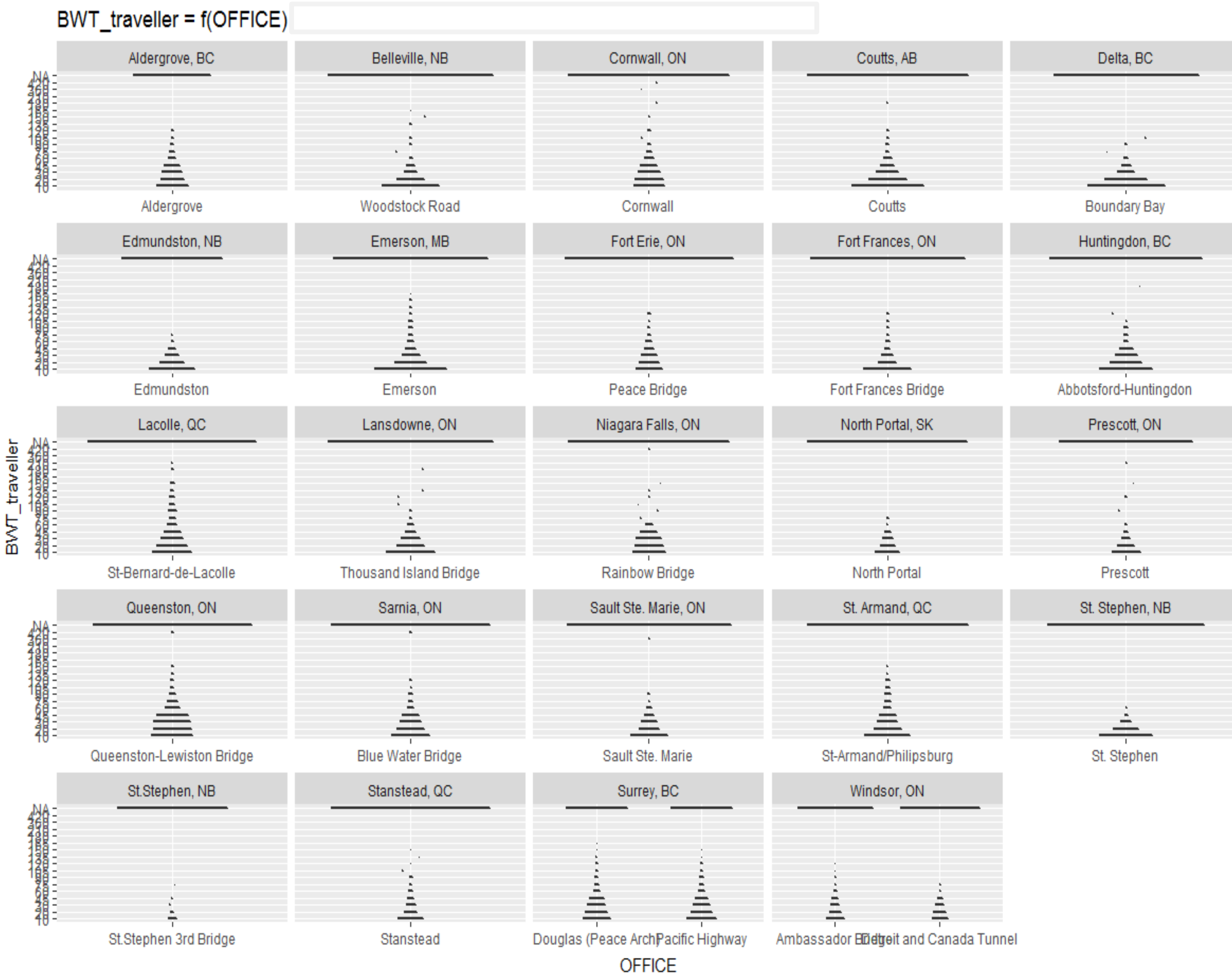


Figure A-1 The distribution of BWT values at all NOR ports.
Histograms show the relative number of instances (represented by the length of black horizontal lines) when BWT was within 10 (at the bottom),20,..,420 minutes , or not recorded (marked NA, top long line). No distinction between week-day and week-end records is made. Records also include data for NEXUS travellers.

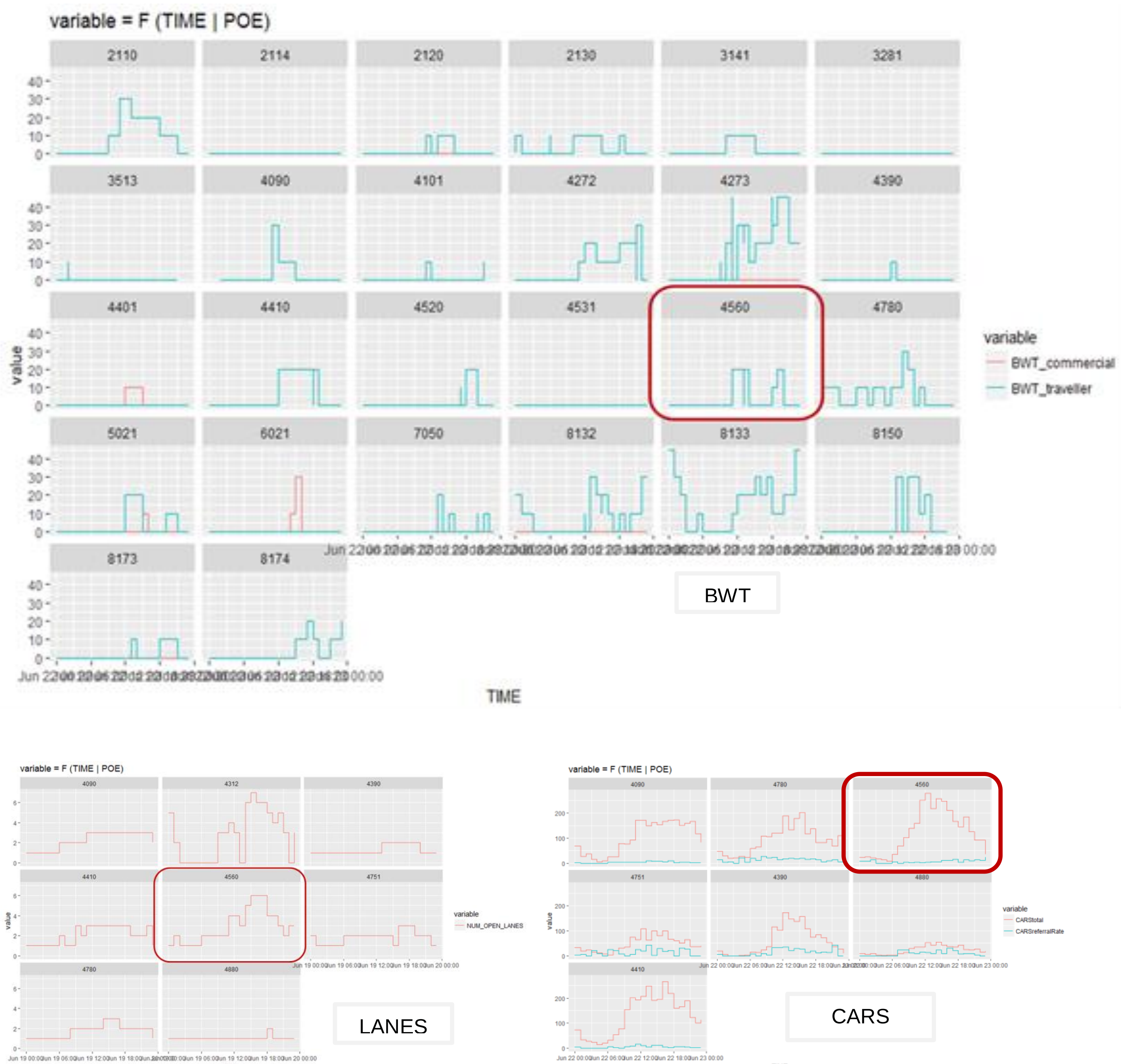


Figure 0-1 A one-day snapshot (Sunday, 22 June 2013) of BWT, LANES, CARS values for all ports in the dataset. Each grey box corresponds to a port. The port (Lansdown, ID=4560), which is used for presenting the simulations results, is marked in Red.

Appendix B: Correlation analysis of historical data

The first step in examining the relationship between the variables of the historical data consists in examining their correlations. The result of this analysis is provided in Figure 0-2, which shows the “bird’s eye” visual summary of the correlation analysis of all data from a single port (Prescott). Results from other ports are similar.

High correlations are marked in red. Data corresponding to BWT are marked by Blue rectangles. The variable names are renamed, following filtering and merging of all data components to allow easier data interpretation. Not all provided data variables are shown, but rather only those that have been found useful for the analysis.

The effect of LANES on key performance variables (total CARS processed, CARS referrals, the number of FTE and BWT) is also shown in Figure 0-3, which shows the counts of all variable values for different number of lanes.

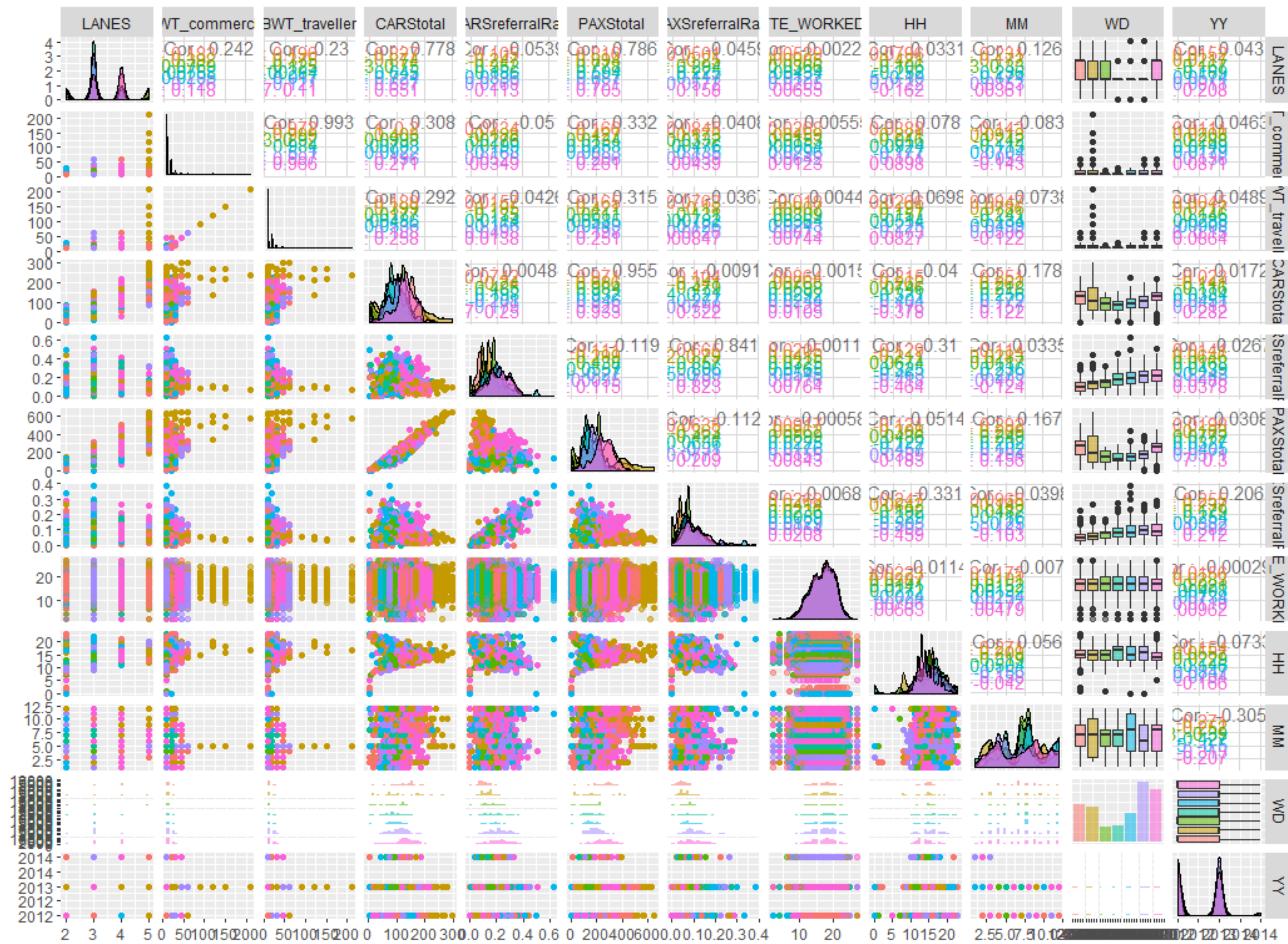


Figure 0-2 “Bird’s eye” view of data variables values and their inter-relationship (Prescott). The figure shows: variable value ranges (along the left vertical axis), data distributions (in diagonal), inter-variable correlations (in upper triangle - large correlations in grey are marked in red), and the affect day of the week (WD) on all of these variables (each colour represents one of seven days of the week). Coloured boxplots show the min,max,mean plus/minus standard deviation for each variable value as a function of WD. Overlapping numbers can be ignored.

A key research task of the project was to understand the relationship between the number of allocated lanes (LANES), traffic volume (CARS) and BWT. Frontline officers have indicated that a linear relationship with LANES and CARS is currently commonly assumed, i.e., LANES = K * CARS, where K is constant

To investigate this relationship, a linear regression was applied to compute the constant K= CARS/LANES as function of LANES as well as month and week day. shows corresponding results computed from all Sunday data in one port (Prescott). Results from other ports and days are similar.

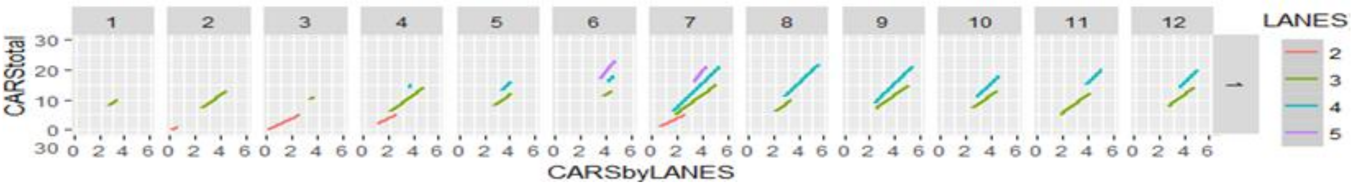


Figure 0-3 Analysis of Cars Per Lanes ratio (K= CARS/LANES) as a function of CARS, LANES and month.