Transportation Big Data Simulation Platform for the Greater Toronto Area (GTA)

Islam R. Kamel[1](✉), Hossam Abdelgawad[1,2], and Baher Abdulhai[1]

[1] Civil Engineering Department, University of Toronto, Toronto, Canada
islam.kamel@mail.utoronto.ca,
hossam.abdelgawad@alumni.utoronto.ca,
baher.abdulhai@utoronto.ca
[2] Faculty of Engineering, Cairo University, Giza 12631, Egypt

Abstract. This paper presents how big data could be utilized in preparing for smart cities. Within this context, smart cities require intelligent decisions in real time, while processing large amount of data. One big component that relates to smart cities in ITS applications is using artificial intelligent techniques that rely heavily on simulation environments for the evaluation and testing of ITS strategies. In this paper, we present a model for the GTA transportation network. While the model enables big data transportation applications to run in real time, its building process implied intensive work with big data. Within this paper, we show the structure, the calibration, and the outputs of the model. Moreover, some applications, which use the proposed model, are presented. These big data applications are a step towards the smart city of Toronto. Finally, we conclude with some thoughts of future work and the next generation of big data models.

Keywords: Big data · Smart city · Traffic simulation · Intelligent Transportation Systems · Greater Toronto Area

1 Introduction

Big data is typically characterized by three dimensions, widely known as 3Vs: *v*olume, *v*elocity, and *v*ariety as first defined by [1]. As such, dealing with big data implies processing very large volumes of fast generated data coming from multiple sources. Since then researchers have been expanding the definition of the basic 3Vs to include many other Vs, including: *v*eracity, *v*isualization, and *v*alue of data [2]. Additionally, the quality, readability, and contents of data are essential as the data itself. Having a lot of data generated at high speed from different sources is worthless, if it is of a low quality. This why research emphasizes the need to measure the ***v***eracity of data by determining how much one can trust the data. Even trustworthy data may be useless, if it remains in its raw format without being processed and visualized. *V*isualization is always seen to be a challenge when dealing with big data as it is challenging to summarize several variables with a lot of details in a graph or chart. The ***v***alue of data is another basic dimension in characterizing big data. Although the data in its raw format may not be of a great value, rigorous analysis (descriptive, exploratory, and statistical) is what gains the data its value.

A. Leon-Garcia et al. (Eds.): Smart City 2015, LNICST 166, pp. 443–454, 2016.
DOI: 10.1007/978-3-319-33681-7_37

Another dimension, which is somehow related to the *v*ariety or *v*eracity of data, is the *v*ariability of data [3]. It defines the variance in the meaning of data. The meaning of data should be extracted from the whole context around it, just like words which have different meanings according to the context they are used in. In summary, to date, various definitions of big data result in 7 Vs as follows: *V*olume, *V*elocity, *V*ariety, *V*eracity, *V*isualization, *V*alue, and *V*ariability.

This increasing interest in big data is witnessed across many fields worldwide. In two years (from 2011 to 2013), the generated data worldwide was almost doubled (from 1.8 to 4 ZB) [4]. These massive volumes of data exist almost everywhere around us; generating a lot of opportunities and challenges. Examples include technological industries and social networks, such as: *Facebook* deals with 30+ PetabyteS of user-generated data and receives 100 TB of data daily; *YouTube* handles more than 48 h uploads every minute; and *Twitter* estimated the number of tweets in early 2012 to be 175 million tweets daily. Examples also include other industries, such as retail companies: *Walmart* handles every hour more than 1 million customer transactions and imported them into databases estimated to contain more than 2.5 PB of data [5].

Big data in transportation is no different from the above trends. Collected data and traffic network conditions, especially those indicating traffic conditions and public transit status, is growing rapidly within the field of Intelligent Transportation Systems (ITS). For instance, traffic loop detectors in the Netherlands generate approximately 80 million records daily [6]; while the Greater Toronto Area (GTA) alone generates around 35 million traffic loop detection readings [7]. Crowd sourcing traffic data generate a significant portion of data in transportation (e.g., Waze: the largest community-based traffic and navigation mobile application [8]; and Roadify: the largest mobile application for transit information [9]). These data sources provide real-time traffic and transit information to enable travellers make better real-time informative decisions.

The intelligent use, analysis, and visualization of the previous data sources, combined with information and communication technologies (ICT) aim at building smarter cities. While there is no specific definition to the term smart city; it could be characterized by: (1) efficiently using the existing infrastructure, such as roads, by harnessing artificial intelligence techniques and data analytics; (2) effectively engaging with residents through e-Participation to improve the decision making process; and (3) intelligently responding to any changes or system disturbances in real-time.

From a transportation perspective, efficient and fast traffic simulation modelling of the transportation network enables real-time simulation of the system and application of smart artificial intelligence techniques to the transportation system. Not only that building and calibrating these simulation models require big data, but also these models themselves can generate a significant amount of data with various frequencies, and varieties; i.e., another source of big data.

To that effect, the motivation of this research is to shed some light into the big data sources that could be harnessed to build large-scale simulation platforms to enable ITS applications and smart cities; and to illustrate how simulation models themselves could be the source of generating big data. These models – along with the data used to build and calibrate them, and the data to be generated from their use in ITS applications – form a key component of smart cities as they replicate the dynamic and stochastic transportation system in a simulation environment.

2 Big Data in ITS

Big data in transportation and smart cities is related to the use of artificial intelligence and digital technology in order to improve the performance of the existing services and to achieve the optimal usage of the infrastructure. This is exactly the definition of the ITS. In this section, we show how big data is a core part of ITS using a few examples. For instance, the advances in ICT and automobile manufacturing sectors have motivated the ongoing research on connected vehicles. With the number of vehicles reaching 1.18 billion vehicles in 2013 (25 % increase from 2006) [10], the emergence to the smart connected vehicles becomes a must rather than a choice. In fact, most modern smart cities have the minimum required intelligent infrastructure to be compatible with these connected vehicles. According to [11], a connected vehicle today has about 40 microprocessors and hundreds of sensors, and generates more than 25 GB of data per hour. Considering only 1 % penetration of the 1.18 billion vehicles above could easily translate to 280 PB of data per hour. With this potential massive set of data, a number of questions arise: Could it be processed and analyzed in real-time? What are the travel patterns and sensory information that could be inferred from this data? Could this data be integrated with other sources such as weather and social media? The answer to these questions could open up new possibilities and applications that would easily change the ITS and transportation systems all together.

Another recent example of big data in transportation is the naturalistic driving study, which is part of the Strategic Highway Research Program 2 (SHRP 2) experiments. In this single study, large volumes of data from more than 3000 drivers in the United States were collected to better understand drivers' behaviors and hence improve the safety on the roads [12]. The collected data includes: driving, driver, crash, and roadway data obtained by cameras and sensors installed in the vehicle and attached to the driver. In 2014, the collected data was estimated to exceed 4 PB from multiple sources with about 5.4 million trips covered representing about 3,958 vehicle-years of data.

As discussed earlier, simulation platforms are integral part of enabling ITS applications in both real-time and off-line testing and modelling of smart cities. The next section discusses a case study of developing an enabling big data simulation platform for the Greater Toronto Area (GTA).

3 Big Data Simulation Platform for the GTA

Although developing a traffic simulation model to test what-if scenarios is becoming a best practice and known among the research community; we approach the simulation platform in this paper as an enabler for big data applications in traffic and ITS. In this section, we present a big data simulation platform for the GTA, which includes: a simulation-based dynamic traffic assignment (DTA) model, a set of analytics, and of a lot of data collected from multiple sources within the region. This platform is an important step towards the smart city of Toronto; the digital city that utilizes all available data to improve the traffic and safety conditions on its roads. Currently, this platform is being used in many applications, such as minimizing evacuation travel time in case of severe disasters, implementing dynamic congestion pricing over some corridors in the GTA, and deciding optimal investment decisions for regional transportation planning projects. In this section, we present the structure of the proposed model, how it was built, and its outputs. Thereafter, the calibration of the model is presented, showing the different criteria used to adjust the parameters of the model. Finally, the section concludes with a brief discussion on the challenges and data issues faced during building, calibrating, and using the model.

3.1 Model Input Data: Volume, Variety, Veracity and Value

A number of data sources have been used to build the DTA model. It started with a comprehensive GIS database from Land Information Ontario (LIO) warehouse as shown in Fig. 1. The model covers more than 7000 km^2 and consists of 26,446 links, including all highways and major arterials in the GTA, and 14,228 nodes, including 830 signalized intersections.

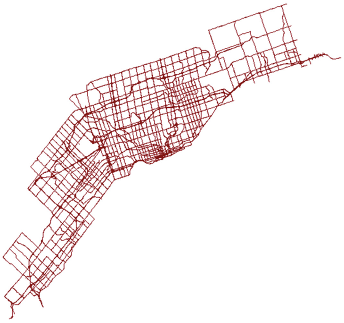


Fig. 1. The GTA network obtained from LIO

Although that the ***v***alue of the GIS database was great to form the basis for the model geometry, it suffered from a number of missing items that accentuate the need to further refine the data and complete the missing data. The following are tasks to illustrate the effort required to refine the data for such big area: encoding locations and timing of 800+ signalized intersections; defining location of thousands of off and on ramps to the highway network; preventing U-turn movements at the signalized intersections; and checking the number of lanes and geometry of more than half the network.

3.1.1 Travel Demand and Traffic Zones

After the network geometry was prepared, the travel demand data has to be prepared and fed into the simulation model. For such big area, the number of trips travelled across the region is captured through the Transportation Tomorrow Survey, which is one of the largest surveys in North America [13]. The travel demand is then translated to form of trips from origins to destinations (OD matrices). To reflect on the volume and frequency of the demand data, an OD matrix with 2.25 million cells (1500 origins to 1500 destinations) has to be generated each 15 min, for the AM peak period (from 6 AM to 10 AM); resulting in 36 million OD cell records fed into the simulation model. During this period, about 2 million trips traverse the GTA network and their individual traces are stored in a min per min basis.

Quality and *v*eracity of the demand data are vital to ensure the accuracy of the simulation model. To that effect, two issues arose after extracting the TTS demand data. *First*, the number of generated trips every 15 min exhibited significant flip-flopping demand pattern due to the fact that travelers have a tendency to report their departure time on a half-hour basis, instead of using the exact time 15 min interval. To resolve this issue, the OD matrices were filtered through a mathematical procedure to generate a smooth demand curve while maintaining the same total number of trips as shown in Fig. 2.

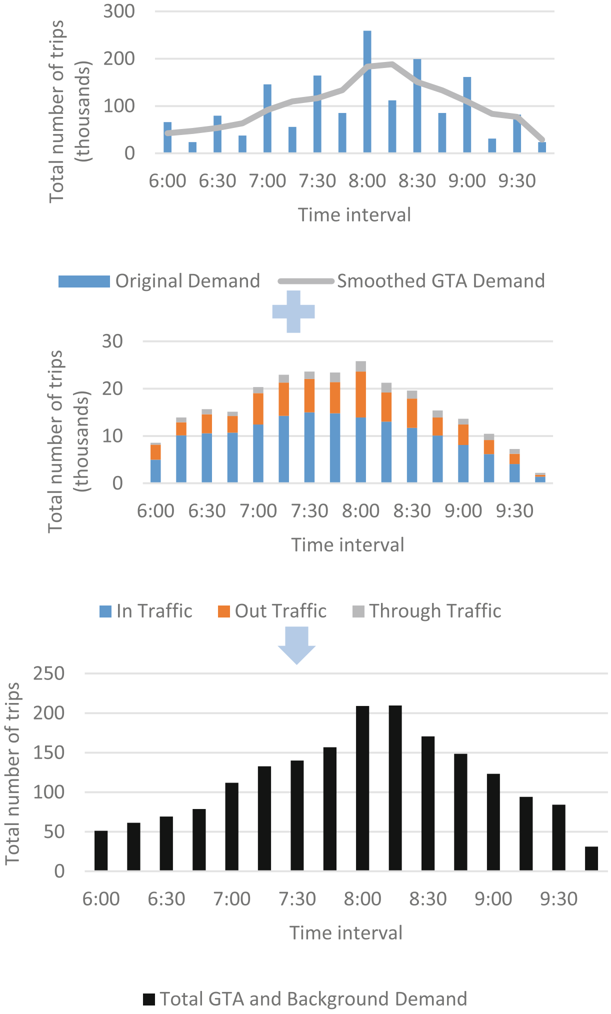


Fig. 2. The GTA travel demand during the AM peak (Color figure online)

Second, the TTS demand matrices did not consider the fact at 6 AM there is traffic in the network, the so-called background demand. This background demand shall include three types of trips: (1) from outside the GTA to the core of the network, (2) from the GTA to outside the network, and (3) from and to zones outside the GTA but passing through some routes within the network. To resolve this issue, a model of the Greater Toronto and Hamilton Area was simulated to get information about vehicles that meet the above conditions and the simulation results were analyzed to find the path for each vehicle, which can be traversed to find out the number of vehicles/trips in each category. Figure 2 shows the background demand during the AM peak and the total number of extra OD trips added to each time interval. About 260,000 trips have been added to the original demand. The total smoothed demand of the GTA, including these trips, is shown in Fig. 2, and estimated as 1.9 million trips.

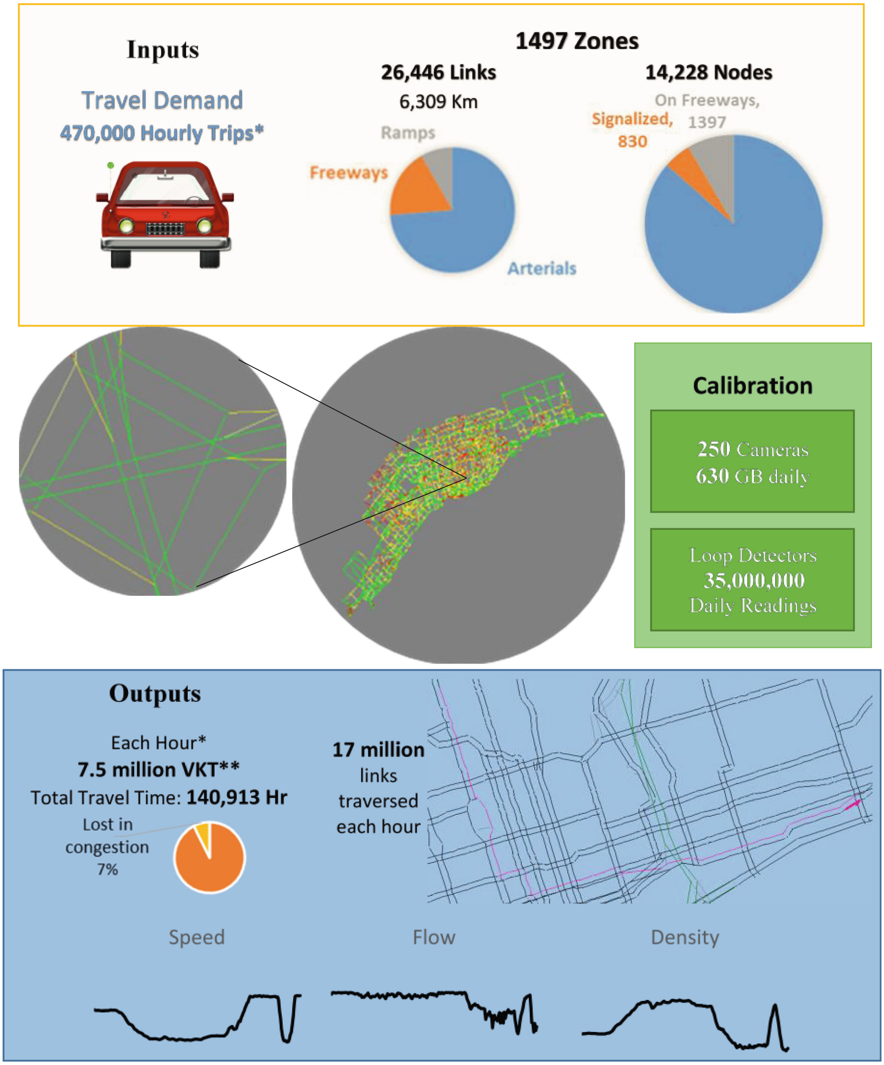
The above input data examples clearly demonstrate the *v*olume and *v*ariety of the required data to develop the basic geometry and demand of the simulation model; but what is probably more important is to reflect on the additional significant effort exerted to deal with issues related to the *v*eracity and *v*alue of the data to the application at hand.

3.2 Model Output Data: Volume, Visualization, Velocity and Variability

Using the simulation modelling platform, various types of outputs can be generated: at the network-level, e.g., total travel time, travelled distance, and time lost in traffic; and corridor and facility level, e.g., speed over a link, queue length at an intersection, and travel path of a vehicle. Simulation outputs vary in ***v***olume and ***v***elocity. For instance, while the total travel time is only a single value for the whole network produced at the end of the simulation, traffic counts over links are reported minute by minute for each single link. In the GTA model, this translates to 1.6 million readings for only one measure that characterizes traffic each hour of the simulation (i.e., traffic counts or flows on links). Considering other key performance measures; such as average speed, free flow speed, traffic density, toll revenues, and queues of vehicles; this data could easily result in 10 million readings each hour; 40 million readings for the morning rush hour; and 240 million readings for a day of simulated traffic data for the GTA. Additionally, detailed vehicle trajectory data are available for each vehicle. The spatio-temporal paths of vehicles generate more than 17 million node-arrival time pairs each hour, and are updated almost second by second. A sample of vehicles' paths is shown in Fig. 3. Although the large ***v***olume and ***v***elocity of this fast-produced traffic data poses a number of challenges associated with extracting useful information and making informative decisions in real time, it also introduces an opportunity for data analytics and visualization techniques to translate this humongous set of data to meaningful information in an intuitive manner, especially for real-time ITS applications.

3.3 Model Calibration and Validation: More Vs

After defining the model inputs, building the model, and extracting model outputs and measures, the calibration and validation of the model cannot be overemphasized. The calibration process consisted of multiple interrelated steps, including: a comparison



* During AM Peak

** VKT: Vehicle Kilometer Travelled

Fig. 3. Inputs and outputs of the GTA model, and calibration data

between observed and simulated measures, such as traffic speed and volume, to evaluate the accuracy of the model; and tweaking some specific parameters, e.g., traffic flow model parameters, OD matrices, and network geometry-related parameters, to improve the calibration results. Figure 3 shows the various inputs and outputs of the model, and sources of data used in calibration of the model, e.g., loop detectors with readings updated every 20 s for each single lane, camera feeds from about 250 different locations, hundreds of daily Twitter reports, and google maps.

Two sets of data were used during the calibration process. *First*, the simulated hourly volumes at 177 locations over different highways, e.g., HWY401 and the Gardiner, were compared against data collected from loop detectors. The simulated volumes were plotted against the average observed volumes. Due to variations and seasonality of observed data, a wide range of weekdays (Tuesday, Wednesday, and/or Thursday) across a number of months (September, October, or November) were considered in the calibration process. In addition to the average of the observed volumes, the observed volume variations were estimated at each location to build an envelope that represents an estimated interval with a 95 % confidence. Figure 4 shows a scatter plot of the observed and simulated volumes classified by highway and travel direction.

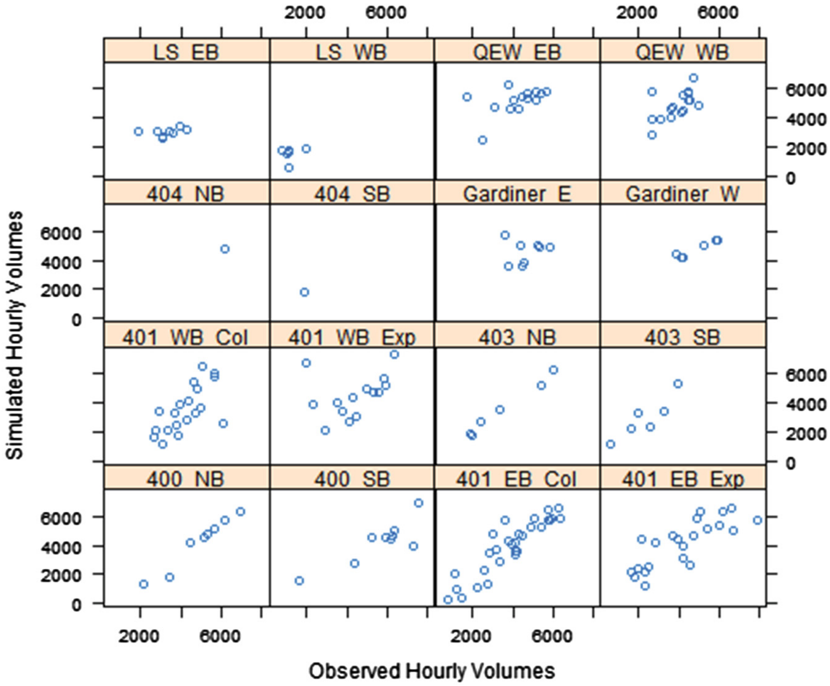


Fig. 4. Scatter plot of the observed and simulated hourly volumes

Second, observed speeds are compared against simulated speeds. Like the traffic volumes, the observed speeds were obtained from detectors every 20 s. The comparison was held between the simulated speeds, and the average observed speeds and their envelopes which are defined as intervals contain the average speed at confidence level of 95 %. The results showed that simulated speeds follow a similar pattern as this of the observed speeds as shown in Fig. 5.

Afterwards, some OD trip demand tables were adjusted to ensure calibrated speeds and flows/volumes. More than 1300 trips travelling during the AM peak were analyzed to obtain their origins and destinations. These OD pairs were adjusted (increased/decreased) based on how much they affect the traffic flows and speeds across

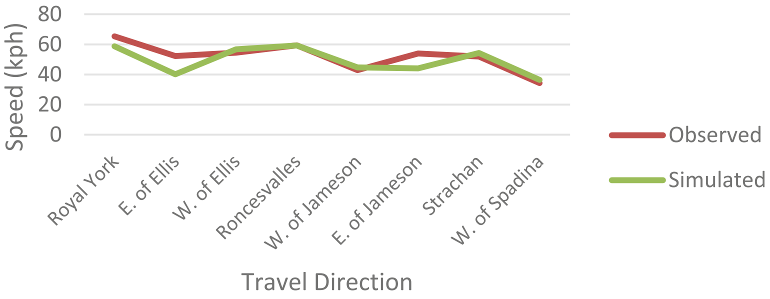


Fig. 5. Lake shore Blvd EB: observed and simulated speeds (Color figure online)

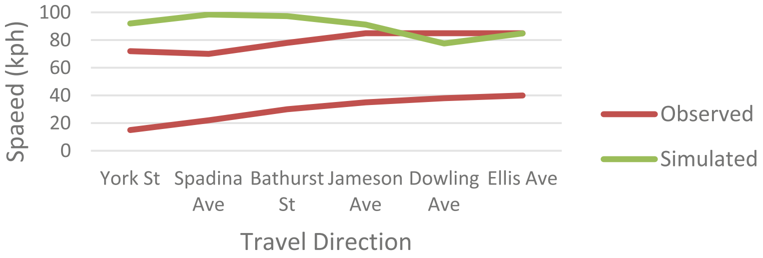


Fig. 6. Gardiner expressway WB: speeds before demand calibration (Color figure online)

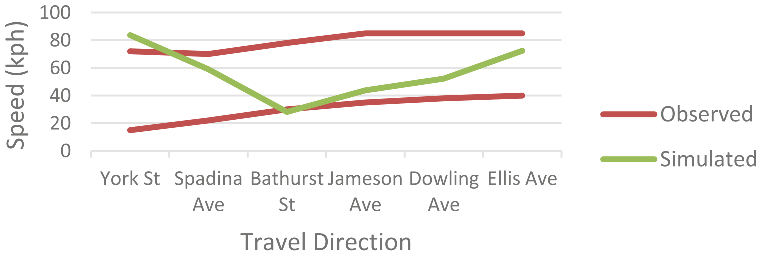


Fig. 7. Gardiner expressway WB: speeds after demand calibration (Color figure online)

key corridors throughout the network. Sometimes additional demand imposes more traffic and therefore creates realistic congestion such as the case of the Gardiner WB. The simulated speeds before and after demand calibration are plotted with the observed speed envelope in Figs. 6 and 7, respectively.

3.4 Data Issues and Challenges

As discussed above, a number of data challenges were faced while dealing with this big data simulation platform for the GTA. The GTA network contains thousands of links and nodes, and millions of vehicles. Not only the *v*olume of data, size of the network, and *v*ariety of data sources were the key challenges; but also the ***v***eracity and ***v***alue of the data used to build and calibrate this model. Similar challenges appeared during the calibration and validation process, such as large volumes, high speed, and different sources of data. During the calibration, historical loop detectors data from 2010 to 2012 were used. These loop detectors cover the GTA highway network and produced about 40 million readings a day. These large datasets has been filtered to get only traffic speed and volume during weekdays within months: September, October, and November, to match the travel demand fed into the simulation.

As a common problem in big data, not all readings are always reported correctly. Hence, a threshold has been set to determine the maximum number of missing readings which maintain the data quality at an acceptable level. The threshold was set such that a detector will be excluded either if it is down for more than a minute, i.e., three successive readings are missing, or if it has more than 5 % of its daily readings missing. On the other hand, the few gaps within the readings of the acceptable detectors were supplemented using a simple moving average procedure taking into consideration values of their neighbors. Moreover, for each acceptable detector, the speed-volume relationship is checked to verify that it follow the regular speed-volume relation.

4 Applications

The proposed platform is currently being used in some real applications and will be used in more applications in the next few years. In this section, we give a brief overview on the applications that have used and are currently using the proposed GTA model. While some of these applications can run in real-time, such as the congestion pricing framework; others work offline, such as the emergency evacuation platform. On the other hand, the model is used as a use case to test and validate some other applications and platforms, such as the platform presented in [14] which studies the robustness of networks. These applications deal with big volumes of very frequent data as they contain the GTA model within their core.

4.1 Emergency Evacuation

In [15], authors presented a multi-objective optimization framework for emergency evacuation and tested it with the city of Toronto as a use case. This platform minimizes the in-vehicle travel time, the at-origin waiting time, and the fleet cost in the case of mass evacuation. It has been tested for evacuating more than 1 million persons from Toronto based on travel demand extracted from TTS-2001. A lot of changes have occurred through the last decade, and the TTS-2001 dataset became obsolete. Hence, the evacuation platform is currently running using the new simulation GTA model presented in this paper. The new application run aims to produce up-to-date results match the current population of Toronto and to test a different implementation scheme of the evacuation platform where it runs on a virtual infrastructure.

In the current version of the evacuation platform, the optimization is done by Genetic Algorithm (GA) which runs on different virtual machines on the NSERC strategic network for Smart Applications on Virtual Infrastructure (SAVI). It is a network that provides various applications with a flexible platform where they can run and redistributed over different virtual machines [16]. Each generation within the GA includes performing 20 simulation runs and analyzing their results. These results are in order of and millions of data records. Within each simulation run, the GA produces new pattern for travel demand by assigning different departure time for each vehicle. That includes making trip plans for millions of vehicles simultaneously.

4.2 Robust Network Design

Robust network design is widely used in many fields studying the optimal design of different network layouts that can handle the impacts of any disturbance in the network, such as weather conditions and incidents in transportation networks, excessive noise in communication networks, cyber-attacks on computer networks, and even disturbances in financial networks. The framework presented in [14] studied different robust network designs especially for transportation networks. The simulation test network started with a proof-of-concept network, and the research now has been extended to improve the performance of the simulation in [17] and was tested using the proposed GTA model. This paper presents new solutions to speed up the DTA simulation of transportation networks, especially for large-scale applications. Hence, the authors used our big data GTA model as an example of large transportation networks to test the efficiency of their algorithm.

4.3 Congestion Pricing

Dynamic congestion pricing is an ongoing research in the transportation group at University of Toronto. Unlike other congestion pricing policies, researchers at University of Toronto are working on an optimization platform that finds the optimal dynamic toll on roads to maximize the social welfare [18]. Unlike the emergency evacuation application, the dynamic pricing platform is intended to be a real-time application, which makes intensive analysis and optimization on larger volumes of data within timing constraints. In addition to the optimization, it analyzes the simulation results to get information about vehicles paths, traffic volumes and speeds, travel time, and toll revenues. Using the GTA model, the application sets dynamic tolls on the GTA's roads, while analyzing about 10 million of data records per hour.

5 Next Generation ITS and Big Data

The transportation applications illustrated above are just examples of how big data could be utilized given current data needs and simulation requirements. Such applications, which deal with big data nowadays, cannot be compared to the future applications of ITS and the expected sizes of data they will deal with. As expected by many experts, the future revolution in ITS will be the emergence of autonomous (driverless) vehicles. These self-driving vehicles are seen to be the next generation of the connected vehicles. Although some technologies already exist and some samples of these vehicles are currently under test, this is a tiny fraction of what will be available in the future.

While connected vehicles are still driven by humans and only use communication technologies to improve their efficiency and safety, driverless vehicles will be fully automated. Instead of the 25 GB generated by each connected vehicle today, the generated data from autonomous vehicles will be hundred fold greater.

Like the existing transportation applications, these future projects will soon require trustworthy models that can deal with these expected big volumes of data. Those models will be the future of existing models, such as the GTA model presented in this paper. They have to deal with very dynamic environments including tremendous volumes of data on both micro and macroscopic levels.

References

1. Laney, D.: 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group Inc., Stamford (2001)
2. Quinn, E.: Discovering big data's value with graph analytics. White Paper, Enterprise Strategy Group (2013)
3. Understanding big data: the seven V's. http://dataconomy.com/seven-vs-big-data/
4. Podesta, J., Pritzer, P., Moniz, E.J., Holdren, J., Zients, J.: Big Data: Seizing Opportunities, Preserving Values. Executive Office of the President, Washington (2014)
5. A comprehensive list of big data statistics. http://wikibon.org/blog/big-data-statistics/
6. Daas, P., Loo, M.: Big data (and official statistics) (2013)
7. Browse traffic loop detectors by list (ONE-ITS). http://128.100.217.245/web/etr-407/trafficreports2
8. Waze mobile. https://www.waze.com/
9. Roadify. http://www.roadify.com/
10. Statista Inc.: Number of passenger cars and commercial vehicles in use worldwide from 2006 to 2013 (in millions). http://www.statista.com/statistics/281134/number-of-vehicles-in-use-worldwide/
11. Hitachi Data Systems: The Internet on Wheels. Technical report (2014)
12. Campbell, J.L.: SHRP2 naturalistic driving study: data collection is complete - now what? In: Northwest Transportation Conference (2014)
13. Data Management Group at UofT: Transportation Tomorrow Survey. http://dmg.utoronto.ca/transportation-tomorrow-survey/tts-introduction
14. Koulakezian, A., Abdelgawad, H., Tizghadam, A., Abdulhai, B., Leon-Garcia, A.: Robust network design for roadway networks unifying framework and application. IEEE ITS Mag. **7**(2), 34–46 (2015)
15. Abdelgawad, H., Abdulhai, B., Wahba, M.: Multiobjective optimization for multimodal evacuation. J. Transp. Res. Rec. **2196**, 21–33 (2010)
16. Kang, J.-M., Lin, T., Bannazadeh, H., Leon-Garcia, A.: Software-defined infrastructure and the SAVI testbed. In: 9th International Conference on Testbeds and Research Infrastructures for the Development of Networks & Communities (Tridentcom) (2014)
17. Koulakezian, A., Graydon, W. E., Abdelgawad, H., Chiu, Y.-C., Abdulhai, B., Leon-Garcia, A.: Speedup of DTA-based simulation of large metropolises for quasi real-time ITS applications. In: IEEE 18th International Conference on ITS (2015)
18. Aboudina, A., Abdulhai, B.: Win-win dynamic congestion pricing for congested urban areas. In: ITS Canada - ACGM (2012)