From Massive Trajectory Data to Traffic Modeling for Better Behavior Prediction in a Usage-Based Insurance Context

Philippe Blais 1,*, Thierry Badard 1, Thierry Duchesne 2 and Marie-Pier Côté 3

1 Research Center for Geospatial Data and Intelligence, Université Laval, Québec, QC G1V 0A6, Canada; thierry.badard@scg.ulaval.ca
2 Department of Mathematics and Statistics, Université Laval, Québec, QC G1V 0A6, Canada; thierry.duchesne@mat.ulaval.ca
3 École d’actuariat, Université Laval, Québec, QC G1V 0A6, Canada; marie-pier.cote@act.ulaval.ca
* Correspondence: philippe.blais.5@ulaval.ca

Received: 21 October 2020; Accepted: 30 November 2020; Published: 2 December 2020

Abstract: Usage-Based Insurance (UBI) is an insurance framework that has made its appearance in the last few years. It allows direct measurement of the traveling of policyholders, hence the growing interest of the industry to better understand driving behaviors. UBI generates large data volumes, from which events can be extracted, like harsh brakes or accelerations. Still, these events are measured without contextual information, which limits their explanatory power. Traffic is one of these types of contextual information that may have great potential, but access to such data remains an issue. Solutions exist, like traffic data from external providers, but for insurance companies that conduct business over large areas, this could result in very large costs. This paper demonstrates that data from insurance companies acquired via UBI can be used to model traffic. A method based on link travel time is proposed and tested on four Canadian cities. Then, routes created with the model are compared with those created using Google Maps. The results show that for 38 routes with an average length of around 5 km, the difference between the travel time of the routes of the proposed model and Google Maps is as small as one second on average.

Keywords: traffic modeling; big data; usage-based insurance; travel time index; trajectory; prediction; driving behavior

1. Introduction

In general insurance, assessing the risk of a driver is of great importance in order to calculate the right premium. For many years, traditional characteristics like the age of a driver and the price of a vehicle [1] have been used to draw a portrait of drivers. These proxies are useful, but still limited [2], in that they tell who drivers are, but not what they really do on the road, what Baecke and Bocca [2] called the exposure. This is one of the reasons why Usage-Based Insurance (UBI) was developed. In recent years, many insurance companies turned to UBI to measure new variables and events like harsh brakes or accelerations and, also, the actual distance in kilometers driven per year [3]. These provide more precise and personalized data about policyholders, and since then, the industry has had a growing interest in understanding driving behaviors. Insurance companies now want to know why events happen, and for that, contextual information is needed. For example, bad weather conditions could explain why a driver drives slower than usual. Ma et al. [4] analyzed many context variables to be incorporated into insurance premium calculation [4]. Among those variables are traffic conditions that are provided by a smartphone app. However, not all insurance companies have access to such traffic data if they still work with probe vehicles that are not smartphone based. Many private
traffic data providers exist, but such a solution remains very expensive for insurance companies that may need traffic information nationwide. Thanks to UBI, it turns out that insurance companies now have something very valuable, that is the data they collect. GPS probe data collected from clients could be used to model traffic. In recent years, the use of UBI data has grown, but few studies directly used the GPS data of UBI to model traffic. Instead, as mentioned above in [4], the traffic is rather determined using a third party application. In addition to the already known characteristics of probe data, using UBI data would have the advantage of covering traffic over a large area, for example at the scale of a country. The data used in this study have also been sampled over a very long period of time, and as a result, a big volume of data is available and worth exploring for traffic modeling.

The goal of the research is to provide a methodology to model traffic using big GPS probe data from the automobile insurance field. This model could then be used to better understand driving behaviors. The paper addresses this problem. It is structured as follows: The first section reviews previous works that have addressed the problem of traffic modeling. The second section describes the proposed method that is based on link travel time estimation, and finally, the results are presented and compared using route travel times stemming from our model with the ones given by Google Maps.

2. Literature Review

The goal of traffic modeling is to estimate the traffic condition at a given time and location on a road network. Its foundations are based on the data used. They have diversified over the years with the arrival of new acquisition technologies. Bluetooth scanners, automated number plate recognition [5], and video footage from cameras [6] are a few of the numerous ways to estimate traffic conditions. In recent years, GPS data from probe vehicles or smartphones have taken the lead. The nature of this type of data is very often linked to transport services like taxis. This may be due to the fact that taxis are closely related to cities, which are not private organizations. As a result, the data are often shared freely, and so, they became an obvious resource for traffic modeling. The nature of the data has impacts on the methods and resulting models. The fact that they are related to cities often leads to methods that consider particular aspects of the city to which they belong. For example, Kong et al. [7] considered the off-peak period from 12:00 a.m. to 7:00 p.m. in Shenzhen, China. As for Stipanic et al. [8], whose data came from a smartphone application collected for Quebec City, Canada, the off-peak period was defined as 7:00 p.m. to 6:00 a.m. and 10:00 a.m. to 3:00 p.m. The data used in the present research work came from an insurance company that has clients across all of Canada, including cities with very different volumes of traffic. It is obvious that methods based on the characteristics of particular cities would need important adjustments.

A crucial component of traffic modeling with GPS data is the choice of a measure of traffic. In 2012, Castro et al. [9] presented a method based on traffic flow, which is the number of vehicles per unit of time. Suhas et al. [10] showed in a review paper that volume is the most common feature used in traffic prediction. However, methods based on the number of vehicles like flow and volume may not be suitable for GPS data because they are sparse. As mentioned by Olszewski et al. [11], “only a sample of vehicles is collected”, and data from the car insurance industry is no exception, as they only represent vehicles of policyholders from the company who provided the dataset.

In 2017, Puangprakhon and Narupiti [12] proposed a traffic model using GPS data. In this paper, the so-called sparse data rather refer to the low sampling frequency, such as every 15, 30, or 60 s. An interpolation method was proposed to calculate link travel time because, a consequence of low sampling frequency is that point locations do not match the beginning and ending points of links. This method was tested with only 15 vehicles running six road links, which makes the use of this method on the scale of a country hardly imaginable due to the high computational costs. The available data in the present research work were more than 10 Tb and did not need the use of interpolation because of the high sample rate, which was every second.

The volume of data also has an impact on the temporal resolution of the resulting traffic models. Olszewski et al. [11] concluded that a low traffic volume might lead to insufficient data for certain time
periods and that traffic should be estimated for no less than hourly time intervals. However, in the context of traffic modeling in order to understand driving behaviors, it is of great importance to have precise information of the traffic conditions, which might change within a one hour interval. Higher time resolutions of traffic are necessary to contextualize events related to driving behaviors for car insurance companies.

Finally, there are limits to the traffic modeling methods regarding the interpretation of congestion levels made with these models. The model proposed by Puangprakhon and Narupiti aims at estimating link travel time. Fusco et al. [13] on the other hand proposed a model based on machine learning for speed prediction using speed, the variance of individual speeds, and the number of vehicles as inputs. Those models predict travel time and speed values that are not translated into congestion levels, which are of great importance in the context of driving behavior understanding. They enable one to label events. That means placing events in a particular traffic context. Many kinds of traffic indices like the congestion index [8] or travel time index [11] exist to address this problem, but these are often based on the characteristics of particular cities. Stipancic et al. [8] mentioned that the threshold of congestion is 60% of the free flow speed for Quebec, Canada, but is 75% for Washington state. Sisiopiku and Rostami-Hosuri [14] used three travel time index classes; Xu et al. [15] used five; and Kong et al. [7] used six. For a nationwide traffic model, a more homogeneous definition of congestion levels is needed.

To summarize, methods for traffic modeling need adjustments to properly address the problem of traffic modeling using big GPS probe data from the automobile insurance field in order to understand driving behaviors.

3. Methodology

3.1. Additional Information about the Data

Before diving into the model description, more information on the data are presented. The data used in this research work came from a large insurance company in Canada. They consist of GPS points sampled at 1 Hz from probe vehicles of policyholders. Each GPS observation is associated with a latitude, longitude, timestamp, and speed in km/h. The observations are map-matched on Open Street Map (OSM) [16] using Open Source Routing Machine (OSRM) [17]. Direction is calculated using linear referencing. Links of the road network, having each an osm_id, have been split at each intersection, and also, no link is longer than 200 m, allowing for a high spatial traffic resolution. A new unique identifier is given to these split links, a link_id. Only osm_highway types matching the following are used to create the proposed model: motorway, trunk, primary, secondary, tertiary, with or without the suffix _link. These roughly refer to the main road axes, excluding residential roads. A complete description can be found here: https://wiki.openstreetmap.org/wiki/Key:highway.

3.2. Model Description

The literature review highlighted that traffic measures based on vehicle quantity should be avoided when working with sparse data like GPS probes. Features that are intrinsic to each individual vehicle are preferred, like speed and timestamp. Timestamp is far less used than speed due to some disadvantages related to its dependency on the locations of GPS points [8]. As pointed out by Puangprakhon and Narupiti, the location of data with low sample rates might not correspond to the beginning and ending of the link and requires an interpolation method. However, it is shown with the proposed method that timestamps can be considered with data having a sampling rate of 1 Hz to calculate travel time. Travel time is chosen because it is “easy to understand and widely employed in representing traffic status” [18]. The following section details how the traffic model is built. The first part explains how travel time is calculated at the link level and how traces are filtered. The second part explains how the travel time index is calculated. Finally, the third part shows how this method is scaled for the whole road network.
3.2.1. Link Travel Time Calculation and Filtering

To calculate travel time at the link level using timestamps, let us first see how it can be calculated in theory, with a hypothetical straight line \( l \) of length \( L \), with starting node \( d_{\text{start}} \) and ending node \( d_{\text{end}} \) (see the left side of Figure 1). The following equation gives the travel time:

\[
\text{tt} = t_{d_{\text{end}}} - t_{d_{\text{start}}},
\]  

where \( t_{d_{\text{start}}} \) and \( t_{d_{\text{end}}} \) are the timestamps at the corresponding nodes. Say now a car leaves a trace with \( n \) GPS observations \( p_1, \ldots, p_n \) map-matched on \( l \) (see the right side of Figure 1). Equation (1) can now be expressed as:

\[
\text{tt} = t_n - t_1,
\]  

where \( t_1 \) and \( t_n \) are the timestamps of \( p_1 \) and \( p_n \), respectively, the first and last observations. For the travel time of Equation (2) to be equal to that of Equation (1), two conditions must be met. First, the positions of \( p_1 \) and \( p_n \) must exactly correspond respectively to those of \( d_{\text{start}} \) and \( d_{\text{end}} \). Second, all observations from \( p_1 \) to \( p_n \) must be map-matched on \( l \). There cannot be any missing observations; otherwise, this could mean a car used a path other than \( l \) to go from \( d_{\text{start}} \) to \( d_{\text{end}} \), and therefore, the travel time is that of a distance longer than \( L \). In practice, \( p_1 \) and \( p_n \) never correspond exactly to \( d_{\text{start}} \) and \( d_{\text{end}} \), and missing observations can happen more or less often, because of wrong map-matching, bad GPS transmission, or cars using a road link that is not described by the road network used to map-match the data. Therefore, to make valid comparisons of travel time for different car traces on the same link, it must be ensured that the length of a trace described by the GPS observations, say \( \hat{L} \), corresponds as much as possible to the true length \( L \). That means \( p_1 \) and \( p_n \) must be as close as possible to \( d_{\text{start}} \) and \( d_{\text{end}} \), and all GPS observations from \( p_1 \) to \( p_n \) are map-matched only on \( l \). Two criteria are presented below to ensure these two conditions.

The first criterion requires that each car trace must be described by a continuous series of observations in time. Since the sample rate is 1 Hz, this means there must be one observation for each second between the timestamps of \( p_1 \) and \( p_n \). This can be expressed by condition \( C1: n - (t_n - t_1) = 1 \) where \( t_1 \) and \( t_n \) are the timestamps in seconds of \( p_1 \) and \( p_n \), respectively. The second criterion requires that \( p_1 \) and \( p_n \) must be in the first or last 10% of the length \( l \). This can be done using linear referencing where the position of \( p \) along \( l \) is \( \in [0, 1] \). Therefore, the second condition \( C2 \) is:

\[
\text{lin}_{p_1} \leq 0.1 \quad \text{and} \quad \text{lin}_{p_n} \geq 0.9
\]  

where \( \text{lin}_{p} \) is the position of \( p \) along \( l \) expressed as a proportion of \( L \).

Using the aforementioned criteria, the selection of GPS traces that lead to valid travel time calculations are shown on the left side of Figure 2. However, adjustment of the method is needed to avoid eliminating too many traces, especially ones having very few observations map-matched on \( l \), that is less than or equal to 10. That would be the case for fast speeding cars and short road links. In both cases, \( C2 \) would eliminate almost all traces even if no observation were missing in the trace (see the right side of Figure 2). An adjustment is also needed because the result of Equation (2) is an integer; it is the subtraction of two timestamps whose smallest unit of time is an integer in seconds. That means the shorter the travel time, the less accurate it is relatively to the “real” travel time. Consider a link \( l \) with length \( L = 100 \text{ m} \) and two cars leaving four GPS observations, leading to a travel time of three seconds. Say one car is traveling at 120 km/h and the other at 100 km/h. Even though their speeds are...
very different, their travel times are the same because they left the same number of GPS observations on \( l \). Although the above example may seem unlikely, the huge volume of data analyzed was large enough for such cases to occur. It was frequent enough to justify the use of an alternative travel time calculation for short links where cars travel at high speed. Therefore, the method must be adjusted to allow for rational (non-integer) travel times when having few observations.

Kong et al. [7] showed another way of calculating the travel time by using the speed \( v \) of the GPS observations and the length \( L \):

\[
\tau_{vl} = \frac{L}{v},
\]

(3)

where \( \tau_{vl} \) is the travel time based on the mean speed \( v \) of GPS observations map-matched on \( l \) and the length \( L \) of \( l \). Since \( C2 \) cannot be used to ensure \( p_1 \) and \( p_n \) are close to the start and end of \( l \); another criterion must be used. For that, the travel time calculated using Equation (2) can be compared to the travel time using Equation (3). The difference between the two should not be large. The purpose of converting travel time values from integer to non-integer is to have a more precise travel time. If the difference is large, that means a trip may have partially crossed \( l \). For example, say there are three observations. With Equation (2), a travel time of two seconds would be calculated. If \( l = 100 \text{ m} \), and \( v_{\text{mean}} = 30 \text{ km/h} \), the travel time with Equation (3) is 12 s. As \( 12 > 2 \), observations must be missing. Using this, the new condition that must be met is \( C3: \frac{\tau_{vl}}{\tau} < 2 \). A threshold value of two is selected as it empirically seemed to work well. It allows \( \tau_{vl} \) and \( \tau \) to be different, but not too much. The link travel time calculation and filtering algorithm is summarized in Figure 3, and an example of the resulting travel times is depicted in Figure 4.
3.2.2. Travel Time Index

To translate travel time into congestion levels, the travel time index (TTI) is used. Formally, it is the average travel time over the free flow travel time (FFTT). It can be formulated as follows:

$$TTI = \frac{TT}{FFTT}$$

where $TT$ is a travel time measure. For the calculation of FFTT, Chepuri et al. [19] used the minimum travel time observed for a road link. Using the minimum travel time may not reflect the expected travel time in a free flow situation well as drivers of the dataset used in this paper may have very different driving behaviors, which may not be the case for Chepuri et al. [19], whose data came from buses. Stipancic et al. [8] used the mean travel time of the off-peak period (10:00 a.m. to 15:00 p.m., and 19:00 p.m. to 6:00 a.m.) using data from Quebec City. However, the dataset of the present paper showed that congestion can occur in the evening, especially in downtown areas, maybe because of sporting or cultural events. Instead, for each road segment $l$ and over a given time period of observation, the median of travel times measured during the night time (from 22:00 p.m. to 5:00 a.m.) is used to calculate FFTT. The median is chosen as it is more robust to outliers.

For the number of classes and threshold values $f$ of the TTI classes, the proposed model adopts the ones of Google Maps (GM) and Xu et al. [15], as shown in Table 1. Four classes are used for further comparison with GM. Since GM does not provide any value for the thresholds of classes, the ones of Xu et al. [15] are considered. The final breakdown is shown in Table 1.

| Unknown < TTI < Unknown | Proposed Model |
|-------------------------|----------------|
| Fast 0.0 < TTI < 1.3   | free flow     |
| Unknown 1.3 < TTI < 1.6 | low congestion |
| Unknown 1.6 < TTI < 1.9 | moderate congestion |
| Slow TTI > 1.9 < 2.2   | heavy congestion |
| TTI > 2.2               |                |

Table 1. Travel time index (TTI) class intervals and names.
3.2.3. Scaling at the Road Network Level

So far, it has been shown how to calculate the travel time of a single GPS trace on a single link and how to calculate congestion levels with TTI. This section describes how to scale this at the road network level. Figure 5 presents the general scaling method. Each step must be done one after the other as information from one step is necessary for the following one. The input data are the GPS probe data and the road network. The final output is the road network with the median travel time and congestion levels for every link and every quarter-hour of a typical day of the week.

Figure 5. General scaling method. FFTT: free flow travel time.

Step 1 (Algorithm 1) involves the calculation of trace travel times at the link level. The inputs are three months of GPS probe data excluding weekends and the road network (RN). Like in many papers [8,21], it is assumed traffic patterns are similar for days of the week, and that is why travel times are averaged together. All trace travel times are calculated based on the method presented in Figure 3. This is done for each direction of every link of the RN. The output data are trace travel times in seconds, with the following labels: hour \( h \), quarter-hour \( q \), night identification (whether in interval [22:00 p.m., 5:00 a.m.]) \( N \), direction \( d \), and link id.

\[
\text{Algorithm 1 Trace travel times’ calculation.}
\]

\[
\text{input} : 3 \text{ months of GPS probe data excluding weekends} + \text{RN} \\
\text{output: Trace travel times with labels } h, q, N, d \text{ and link}_id \\
\text{for each link } l \in \text{RN do} \\
\text{for each direction } d \in \{-1, 1\} \text{ do} \\
\text{for each trace do} \\
\text{Calculate travel time according to method in Figure 3} \\
\text{Add labels } h, q, N, d \text{ and link}_id \text{ according to the time and location of the trip} \\
\text{end} \\
\text{end} \\
\text{end}
\]

Step 2 (Algorithm 2) aims at calculating the free flow travel time (FFTT) for each link and direction of the RN. The input data are the trace travel times with label \( N = \text{TRUE} \) and the RN. The output data are the links of the RN with the additional label \( \text{fftt} \) (seconds), the free flow travel time.

\[
\text{Algorithm 2 Link FFTT calculation.}
\]

\[
\text{input} : \text{trace travel times with label } N = \text{TRUE} + \text{RN} \\
\text{output: Links of the RN with additional label } \text{fftt} \text{ (seconds)} \\
\text{for each link } l \in \text{RN do} \\
\text{for each direction } d \in \{-1, 1\} \text{ do} \\
\text{for each trace do} \\
\text{Calculate travel time according to method in Figure 3} \\
\text{Add label } \text{fftt} \text{ according to the time and location of the trip} \\
\text{end} \\
\text{end} \\
\text{end}
\]
Algorithm 2 Free flow travel time calculation.

input: Trace travel times with label $N = \text{TRUE} + \text{RN}$
output: Links with additional label $\text{fftt}$

for each link $l \in \text{RN}$ do
  for each direction $d \in \{-1, 1\}$ do
    Calculate free flow travel time as the median of trace travel times of the present link $l$
    with direction $d$
    Add label $\text{fftt}$ to link $l$ and direction $d$
  end
end

Step 3 (Algorithm 3) finally calculates the median travel time and congestion level for every link at every quarter-hour of the day. Input data are the trace travel times and the links with new labels from Step 2. Output results are links with additional labels: hour $h$, quarter-hour $q$, median travel time in seconds $\text{med\_tt}$, and congestion levels $\text{cong\_lvl}$.

Algorithm 3 Median travel time and congestion level calculation.

input: Trace travel times and links from Step 2
output: Links with additional labels $h, q, \text{med\_tt}$, and $\text{cong\_lvl}$

for each link $l \in \text{RN}$ do
  for each direction $d \in \{-1, 1\}$ do
    for each hour $h \in \{0, 1, \ldots, 23\}$ do
      for each quarter-hour $q \in \{1, 2, 3, 4\}$ do
        Calculate median travel time $\text{med\_tt}$ of all trace travel times having labels $l, d, h, q$
        Calculate travel time index as $\text{TTI} = \frac{\text{med\_tt}}{\text{fftt}}$
        Classify $\text{cong\_lvl}$ as “free flow”, “low”, “moderate” or “heavy” according to the
        TTI classes of the proposed model in Table 1
        Add labels $h, q, \text{med\_tt}$, and $\text{cong\_lvl}$ to link
      end
    end
  end
end

3.3. Data Used

The methodology presented above was applied to the following dataset. It consisted of the data of drivers from Sherbrooke, Quebec City, Montreal, and Toronto in Canada. The data used were acquired in the days of the week during September, October, and November 2016. Table 2 shows the number of drivers, trips, and GPS points that were used. To deal with the huge amount of data, the library Sparklyr [22] was used to allow parallel processing on a Spark cluster from R code. A total of around two hours and 10 min were necessary to process the data. The 2018 road network of OSM [16] was used to map-match the data. This may potentially lead to some wrong map-matching since a road network is dynamic over time and the data were acquired in 2016. However, as the global evolution of the road network is generally assessed to less than 5–6% per year, this detail can be left out.
Table 2. Statistics of the data used.

|                | Sherbrooke | Quebec City | Montreal | Toronto       |
|----------------|------------|-------------|----------|---------------|
| Number of GPS observations | 229,589,252 | 702,197,751 | 3,437,191,916 | 4,047,701,407 |
| Number of trips      | 275,676    | 695,417     | 2,781,157 | 3,433,869     |
| Number of drivers    | 3554       | 8321        | 28,033   | 39,047        |

3.4. Validation Methodology

Two elements must be validated in the proposed methodology: the congestion levels (the TTI classes) and the estimation of the median travel times for each link for every 15 min of the day.

3.4.1. TTI Classes’ Validation

As was shown in the literature review, there is no ground truth for the number of classes, nor for the threshold values for each TTI class since these depend on the city for which the traffic model was developed. This represents a big challenge. Furthermore, the present model uses one set of TTI classes for a whole country for homogeneity purposes, which makes TTI classes’ validation with other models not really relevant. Still, a comparison could suggest if the choices made for the present model make sense or not. In line with the travel time validation presented next, Google Maps (GM) can be used. The colored links are a representation of the congestion levels that can be used for a map-to-map comparison. The purpose of this comparison is simply to spot some general traffic patterns like, for example, the location and moment at which heavy traffic occurs. It is important to use GM at a large scale, so that congestion location does not change because of cartographic generalization. It is also important to note that GM gives different maps depending on the day of the week. In our model (denoted M), all days of the week are averaged together. For a fair comparison, maps of a Wednesday are analyzed to represent a typical day of the week. To make use of the “historical” predictive model of Google Maps (not in real time), requests were made in the month of October 2020. More precision is given in the next section. Finally, it is understood that such qualitative comparisons have very limited validation power and that conclusions drawn from this should only assess whether the TTI classes are reasonable.

3.4.2. Travel Time Validation

To validate the median travel times that were estimated for each link \( l \) and direction \( d \), it is proposed to simulate routes going from an origin to a destination at given moments of the day on different types of roads according to the OSM highway types. Then, the calculated route travel times can be compared with those of GM, obtained by sending a request through the browser application [20]. This validation method has the following advantages. First, every route follows several road links of different types, which allows for the validation of a wide and diversified portion of the road network. Second, each route can be repeated to have departure times every 15 min of the day, which allows for a complete validation of the temporal coverage of the model on these routes, as much during congestion as during off peak periods.

Google Maps is used because it can provide the estimation of route travel times and because it is a free option that meets our need to validate traffic over a very large territory. Despite this, it is fair to argue about the legitimacy of its use. For example, route travel time estimations depend on the day of the week and month. To ensure a fair comparison with the proposed model, GM’s routes were created during a week of the month of October since the data used with the proposed model are from September to November. More precisely, routes were simulated in GM with a departure time varying from Monday 00:00:00 (19 October 2020) to Friday at 23:45:00 (23 October 2020) every 15 min. Since all days of the week are averaged together in the proposed model, the same was done for GM. At the time the requests were made (spring 2020), real-time traffic was affected by the pandemic that
had just started (1–2 weeks earlier), but according to our analysis, the “historical” (in opposition to real time) traffic model of GM was not yet affected. Indeed, requests made on the same day of the week in October 2016, 2019, or 2020 would give the same results. That explains the choice of making requests October 2020. Comparing routes with data dating back to 2016 with traffic level estimations of 2020 may seem like a strong limitation as traffic conditions might have changed greatly. However, using Google Maps has a strong advantage that no other traffic dataset has, that is covering the entire road network for all cities analyzed in this study, which is why it is used despite the aforementioned disadvantage. A final word must be given concerning what information is extracted from GM. When requesting not in real time (that is, given a specific departure time), GM gives a travel time interval, which may correspond to expected travel time and the worst case scenario travel time (95th percentile). The latter was just not considered.

GM is also used because it is generally accepted that it is a good traffic model, but it is not assumed to be a ground truth. That is why the validation is based on a comparison of both models, and not the calculation of residuals used when calculating RMSE or MAPE for predictive models. The comparison consists of calculating the difference between the prediction errors of both models, as shown by Equation (5):

$$E_{GM} - E_M = (T - T_{GM}) - (T - T_M) = T_M - T_{GM},$$

(5)

where $T$ is the (unknown) route travel time ground truth, $T_{GM}$ and $T_M$ are the route travel times estimated by GM and M, respectively, and $E_{GM}$ and $E_M$ are the respective route travel time prediction errors of GM and M. Equation (5) shows that the comparison is reduced to the difference between the estimated route travel times of both models. If the models give the same results, the difference should be zero. To measure if that is the case, a Wilcoxon signed-rank test is performed. It is a statistical hypothesis test used to compare paired samples for which the difference between the two groups is not assumed to be normally distributed. It aims at testing whether the difference between the two groups is significantly different. The null hypothesis $H_0: F_M = F_{GM}$ is compared against the alternative hypothesis $H_1: F_M \neq F_{GM}$, where $F_M$ and $F_{GM}$ are the travel time distributions of the proposed model and Google Maps, $T_M$ and $T_{GM}$. The $p$-value of the test can then be compared to a selected significance level $\alpha$, typically 0.05, and $H_0$ will be rejected if the $p$-value $\leq \alpha$.

It is important to note that with a large sample, a difference that is statistically significant may not be practically significant. With $H_0$, the difference must be exactly zero, so even a small difference of say 0.0001 s between M and GM could mean $H_0$ is false. In practice however, it could be concluded that such a difference would not be significant. To avoid the wrong conclusion with the Wilcoxon test, it is proposed to perform another test. Among all observations of $T_M - T_{GM}$, one-hundred of them are randomly sampled, and their average is calculated. This is performed 500 times. Then, the 500 average prediction errors are plotted in a boxplot to see if the distribution is symmetrical about zero and if the median is near zero.

4. Results

4.1. Travel Time

Figure 6 shows the distribution of travel times illustrated using boxplots as a function of the time of the day with intervals of 15 min for a link of Autoroute Félix-Leclerc in Quebec City. Colors in the background of the plot represent the congestion levels according to the TTI classes. FFTT is 4.41 s, and a total of 21,871 trips were used for this particular link. It can be seen that congestion occurs at the end of the afternoon. During the peak period, the median travel time reaches a heavy congestion level.

Figure 7 shows the travel time of a road of type secondary. As expected, it can be noted that there is a large difference in the number of trips used, which is 4129, compared to 21,871 for Figure 6 with a motorway road type. Again for this link, congestion occurs during the end of afternoon.
From Figure 6, one can note the progression of the median travel time over the course of the day. The median travel time gets higher and higher from 15:00 p.m. to the peak at 17:00 p.m. and then gets lower until the end of the rush hour.

As for Figure 8, which shows the travel times of Chemin Sainte-Foy in Quebec City, a completely different pattern can be noted. The median travel times oscillate from one time interval to the other. Median travel times are mostly in the highest TTI class. Furthermore, fewer trips were used for this link, and as a result, there is an absence of data for several intervals during the night, which confirms the conclusion of Olszewski et al. [11] about low traffic volumes. To overcome this issue, traffic information extracted from other data sources could be used, such as data from public transportation [23], but this solution would lead to more data integration labor. Instead, to fill this gap, more of the same data could be used since only three months of data were used and much more would be easily accessible. This particular link was specifically selected to show the impact of short links with traffic lights (at the end, given a direction) on the traffic detection. With the presented method, congestion does not clearly appear apart from free flow conditions for this type of link. In fact, only two situations are measured, which can happen when there is congestion or not. One is a car traveling at free flow because of a green light. The other is a car waiting for one traffic light cycle. When congestion occurs, cars wait for two, three, or more light cycles, but that cannot be measured on the link because the end of the waiting line is on another link. One can also note this particular phenomenon on Google Maps, where sporadic congestion strangely appears on links near traffic lights.

![Figure 6. Travel time boxplots as a function of the time of the day with TTI classes observed on osm_id 283052776 (fraction start-end: 0.35–0.51) (Aut. Félix-Leclerc (40) east, Quebec City).](image-url)
Figure 7. Travel time boxplots as a function of time of the day with TTI classes observed on osm_id 333516329 (fraction start-end: 0.84–0.91) (Quatre-Bourgeoys west, Quebec City).

Figure 8. Travel time boxplots as a function of time of the day with TTI classes observed on osm_id 427640838 (fraction start-end: 0.31–0.42) (Chemin Sainte-Foy west, Quebec City).

Figure 9 shows travel time on Bathurst Street near downtown Toronto. It can be seen that congestion is still present even at late hours. The figure also shows a different traffic pattern than that of Quebec City. For that particular link in Toronto, morning and evening rush hours happen later and for a longer period of time (from 8:00 a.m. to 13:00 p.m., and 18:00 p.m. to 23:00 p.m.).

To sum up, the results show that the number of trips on a link can be influenced by its hierarchical level (importance in the road network) and by time of day. More cars travel on motorways. Fewer cars
travel at night, which has an influence on the free flow travel time. The presence of traffic lights paired with short road segments make it hard to detect congestion with the proposed method.

![Image](image_url)

**Figure 9.** Travel time boxplots as a function of time of the day with TTI classes observed on osm_id 32803592 (fraction start-end: 0.16–1.00) (Bathurst Street south, Toronto).

### 4.2. Validation Results

#### 4.2.1. TTI Classes

Figures 10 and 11 show that in both M and GM, traffic occurs in the east direction on Autoroute Félix-Leclerc (40) and the north direction on Autoroute Robert-Bourassa (740). Other maps of the same location in GM’s model also showed that on Autoroute Félix-Leclerc, at the junction of Autoroute Robert-Bourassa, the highest level of traffic occurs from 16:15 p.m. to 16:30 p.m. on Wednesday. As for model M, heavy traffic occurs from 15:30 p.m. to 17:30 p.m.. On the other hand, still at the same location, low traffic starts at 15:15 p.m. and ends at 17:45 p.m. in model M and starts at 15:15 p.m. and ends at 18:15 p.m. for most of the days of the week in GM.

![Image](image_url)

**Figure 10.** Traffic map with model M at 16:30 p.m. over the junction of Aut. Félix-Leclerc and Aut. Robert-Bourassa, Quebec City.
Figure 11. Traffic map with model Google Maps (GM) at 16:30 p.m. for 21 October 2020, over the junction of Aut. Félix-Leclerc and Aut. Robert-Bourassa, Quebec City.

The models seem to show similar traffic levels. However, considering it is only at the highest level for less than an hour in GM and for more than two hours in M, it seems that the threshold for the highest traffic level is lower for model M. From these results, it is hard to conclude on the similarity of the traffic levels of both models, especially since for the most part, differences may be due to the data used for M being from 2016, whereas it is unknown for GM (the request was made in April 2020, for a map of 21 October 2020). The results rather suggest that the levels of traffic of the proposed model (the TTI classes) seem to be reasonable since the traffic map of the proposed model is in line with the GM traffic map. Maps of other locations have led to the same general traffic pattern findings.

4.2.2. Travel Time

Table 3 shows a summary of the 38 routes created. It can be noted that a fairly similar number of routes has been created for each city; each type of OSM highway is represented (except for trunk, which only represents less than 0.5% of the road network), and routes have a variety of lengths. The departure time of routes is every quarter-hour. That makes a total of 3648 different routes. Among them, sixteen were not valid because of a lack of data for some specific short links at specific times of the day with low traffic volume. Results are shown in Figures 12 and 13. The figures show the total route travel time, depending on the departure time, which is every 15 min.

Figure 12. Route travel time on Queen St W (east direction), Toronto (Road type: secondary, route length: 4.20 km).
Performing the Wilcoxon test leads to a median prediction error of \(-0.16\) min, which means model M underestimates route travel times by about 10 s compared to GM. The \(p\)-value is 0.007602, which is smaller than \(\alpha = 0.05\). Therefore, according to these results, both models are statistically different. However, it is important to note that the sample used in the statistical test has a size of 3632, which is pretty large, and it has enough power to detect very small differences. As has been mentioned in the validation method, this may not mean both models are that much different in practical terms. The other proposed validation test was performed. Prediction errors were calculated using random samples of 100 observations among the 3632 available, and the average prediction error was calculated. The 500 average prediction errors are plotted in a boxplot that is shown in Figure 14. It can be seen that the boxplot is fairly well distributed around zero and that the median average prediction error is 1.08 s. This test strongly suggests that in practice, both models yield similar travel time estimations.
5. Discussion and Conclusions

The research goal was to provide a methodology to model traffic using big GPS probe data from the automobile insurance field. In this paper, such data were used to model traffic using a method based on link travel time. It was shown that timestamps can be used for travel time estimation if data have a high sampling rate like 1 Hz. A validation method was presented and suggested that traffic predictions from the model are similar to those from Google Maps. In order to meet the research objective, the model was designed to better understand driving behaviors. It provides congestion levels based on a historical dataset and allows labeling and adding more context to car traces. The results also provide interesting findings from an insurance point of view. First, they show that although car insurance companies have sparse data, they have what is necessary to model traffic on major arteries and that external providers are not of absolute necessity. Second, although the data were produced by insured persons for whom driving behaviors were evaluated, the results being very similar to those of Google Maps suggest there is no bias in the data. For example, one might have expected drivers to drive slower than the rest of the vehicle fleet of the same road network, resulting in higher travel time estimation in free flow situations, but this was not the case. This is interesting for insurance companies as it reveals that drivers may not change their behaviors that much, even if they are being evaluated. Third, the proposed model estimated traffic on major arteries, but not on residential roads. Although traffic is less likely to occur for such types of road segments, no contextual information is provided, which would be interesting to analyze and incorporate in future research. To continue, the comparison with Google Maps’ model also revealed interesting findings. Google Maps is widely used, and it is reasonable to say that the amount of data used must be much larger than what was used in this research. Nevertheless, both models yield pretty similar results when estimating route travel times in a “not real-time” context. In this regard, it seems that not much data are needed in order to have good route travel time and traffic level estimations. To conclude the discussion, the findings also open the way to new research to better understand dangerous driving behaviors. For example, a harsh brake might be better understood because a car went from a link where the congestion level was at free flow to one where it was at a heavy level; or it could be confirmed that stop and go were due to traffic because of these newly added congestion level labels. This work could also open new research on how to take into account the contribution of traffic into insurance pricing, how to extract actionable features that could feed actuarial models, or how to establish driver profiles based on the traffic information provided by the model.

Author Contributions: Philippe Blais conducted the research, performed the experiments, and wrote the paper; Thierry Badard was the research director of the student (first author). Thierry Badard, Thierry Duchesne and
Marie-Pier Côté provided valuable advice during the research project and reviews of the different versions of the paper. All authors read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the Natural Sciences and Engineering Research Council of Canada (RDCPJ 515901-17) and a Canadian insurance company. P.B. was also supported by a grant from the Fonds de recherche du Québec-Nature et Technologie (FRQNT).

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Huang, Y.; Meng, S. Automobile insurance classification ratemaking based on telematics driving data. *Decis. Support Syst.* **2019**, *127*, 113156. [CrossRef]
2. Baecke, P.; Bocca, L. The value of vehicle telematics data in insurance risk selection processes. *Decis. Support Syst.* **2017**, *98*, 69–79. [CrossRef]
3. Verbelen, R.; Antonio, K. Unravelling the predictive power of telematics data in car insurance pricing. *Appl. Stat.* **2018**, *1275–1304*. [CrossRef]
4. Ma, Y.L.; Zhu, X.; Hu, X.; Chiu, Y.C. The use of context-sensitive insurance telematics data in auto insurance rate making. *Transp. Res. Part A Policy Pract.* **2018**, *113*, 243–258. [CrossRef]
5. Jie, L.; Zuylen, H.V.; Chunhua, L.; Shoufeng, L. Social and Monitoring travel times in an urban network using video, GPS and Bluetooth. *Procedia-Soc. Behav. Sci.* **2011**, *20*, 630–637. [CrossRef]
6. Phan, A.; Ferrie, F. Obtaining Dense Road Speed Estimates from Sparse GPS Measurements. In Proceedings of the 2008 11th International IEEE Conference on Intelligent Transportation Systems, Beijing, China, 12–15 October 2008; pp. 157–162. [CrossRef]
7. Kong, X.; Yang, J.; Yang, Z. Measuring Traffic Congestion with Taxi GPS Data and Travel Time Index. In Proceedings of the 15th COTA International Conference of Transportation Professionals, Beijing, China, 24–27 July 2015; pp. 1016–1027. doi:10.1061/9780784479292.346
8. Stipancic, J.; Miranda-Moreno, L.; Labbe, A.; Saunier, N. Measuring and visualizing space–time congestion patterns in an urban road network using large-scale smartphone-collected GPS data. *Transp. Lett.* **2019**, *11*, 391–401. doi:10.1080/19427867.2017.1374022. [CrossRef]
9. Castro, P.S.; Zhang, D.; Li, S. Urban Traffic Modelling and Prediction Using Large Scale Taxi GPS Traces. In *International Conference on Pervasive Computing*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 57–72.
10. Suhas, S.; V, V.; Katti, M.; V, A.P.B.; Naveena, C. A Comprehensive Review on Traffic Prediction for Intelligent Transport System. In Proceedings of the 2017 International Conference on Recent Advances in Electronics and Communication Technology, Bangalore, India, 16–17 March 2017. [CrossRef]
11. Olszewski, P.; Dybicz, T.; Jamroz, K.; Kustra, W.; Romanowska, A. Assessing highway travel time reliability using probe vehicle data. *Transp. Res. Rec.* **2018**, *2672*, 118–130. [CrossRef]
12. Puangprakhon, P.; Narupiti, S. Allocating Travel Times Recorded from Sparse GPS Probe Vehicles into Individual Road Segments. *Transp. Res. Procedia* **2017**, *25*, 2208–2221. [CrossRef]
13. Fusco, G.; Colombaroni, C.; Comelli, L.; Isaenko, N. Short-term traffic predictions on large urban traffic networks: Applications of network-based machine learning models and dynamic traffic assignment models. In Proceedings of the 2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), Budapest, Hungary, 3–5 June 2015; pp. 93–101. [CrossRef]
14. Sisisiopiku, V.P.; Rostami-Hosuri, S. Congestion quantification using the national performance management research data set. *Data* **2017**, *2*, 39. [CrossRef]
15. Xu, F.; Huang, Z.; Zhu, X. A New Design of Real-Time Traffic Index Model for Freeway. In Proceedings of the 5th International Conference on Advanced Cloud and Big Data, CBD 2017, Shanghai, China, 13–16 August 2017; pp. 302–307. [CrossRef]
16. OpenStreetMap Contributors. Planet Dump. 2018. Available online: https://www.openstreetmap.org (accessed on 1 December 2020).
17. Open Source Routing Machine. Available online: http://project-osrm.org/ (accessed on 27 April 2012).
18. Xu, M.; Guo, K. Utilizing Artificial Neural Network in GPS- Equipped Probe Vehicles Data- Based Travel Time Estimation. *IEEE Access* **2019**, *7*, 89412–89426. [CrossRef]
19. Chepuri, A.; Kumar, C.; Bhaneagaonkar, P.; Arkatkar, S.S.; Joshi, G. Travel Time Reliability Analysis on Selected Bus Route of Mysore Using GPS Data. *Transp. Dev. Econ.* **2019**, *5*, 1–15. [CrossRef]
20. Google Maps. Available online: https://www.google.com/maps (accessed on 2 June 2012).
21. Herring, R. Estimating arterial traffic conditions using sparse probe data. In Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems, Funchal, Portugal, 19–22 September 2010; pp. 929–936. [CrossRef]
22. Luraschi, J.; Kuo, K.; Ushey, K.; Allaire, J.; Falaki, H.; Wang, L.; Zhang, A.; Li, Y.; The Apache Software Foundation. Sparklyr: R Interface to Apache Spark; R Package Version 1.2.0. Available online: http://spark.rstudio.com (accessed on 1 December 2020).
23. Mo, A.; Pokonieczny, K.; Wilbik, A.; Wabi, J. Transport Accessibility of Warsaw: A Case Study. Sustainability 2019, 11, 5536.

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.