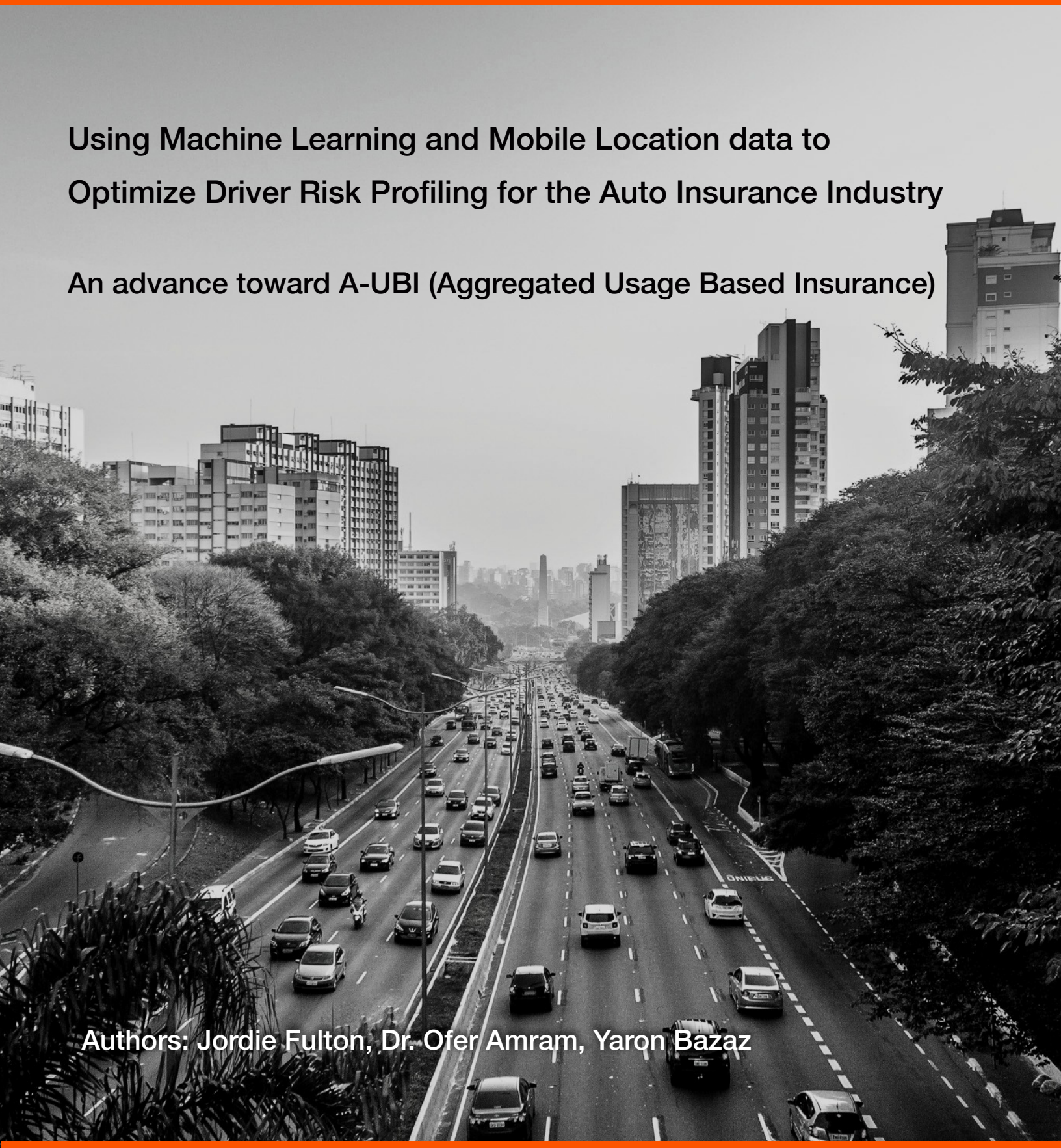
Downtown.AI

July 2021

Using Machine Learning and Mobile Location data to Optimize Driver Risk Profiling for the Auto Insurance Industry

An advance toward A-UBI (Aggregated Usage Based Insurance)



Authors: Jordie Fulton, Dr. Ofer Amram, Yaron Bazaz

Abstract

Within the insurance industry, driver risk profiles are used to assign rates such that risk exposure is minimized, profitability is increased and costs to drivers are further reduced.

In this report, we share the results of Downtown.AI's R&D on the potential use of machine learning and anonymized mobile location data to improve driver risk profiles. We suggest that by analyzing the aggregated and anonymized road behavior of a large sample of the population, we can generate differential risk profiles for very granular geographic areas. In this research we describe four variables we used to measure exposure to risk: VKT (mileage), acceleration, speed and crash risk.

We propose the term "**A - UBI**" for Aggregated Usage Based Insurance. The A-UBI method enables better understanding of driver behavior and exposure to risk - with no apps or sensors to install, and no infringement of client privacy. The new method can improve personal line auto and commercial lines auto risk models.

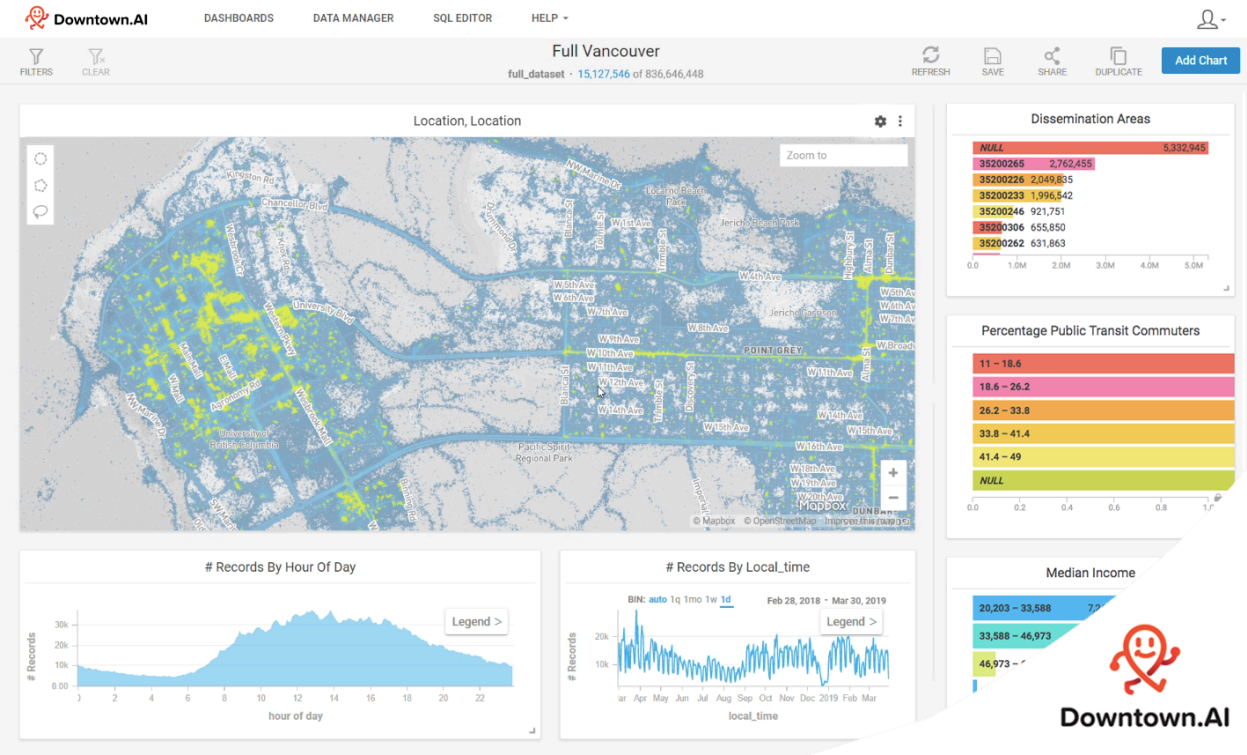
Importantly, this technique allows insurers to set rates in a way that is not only more profitable, but also more fair to their customers.

About Downtown.AI

Downtown.AI leverages machine learning and mobile location data to analyze, map and accurately predict the multimodal traffic of entire countries' populations, including all modes of urban transportation (cars, public transit, pedestrians, micro mobility and ferries).

We aggregate anonymized location data from half a billion mobile devices (adhering to GDPR and CCPA standards), and integrate it with a variety of supporting datasets. We then use proprietary machine learning algorithms and dedicated GPU clusters to rapidly query trillions of data points to generate deep analyses and forecasts which would not be otherwise possible. By analyzing the real-world traffic patterns of entire populations, we enable mobility operators to predict demand, we assist cities to orchestrate traffic, and show transportation and urban planners how they can optimize capital placements.

Downtown.AI's operational dashboard. The real time platform enables clients to visualize, analyze and predict the entire multimodal traffic in a given geography.



Introduction

One of the more complex tasks in the vehicle insurance industry is to determine risk profiles for drivers.[1] An accurate determination of driver risk profile allows insurance companies to assign rates in a way that minimizes their exposure to risk, increases profitability and further reduces costs to drivers.[2] Therefore, insurance companies analyze and model large amounts of data at both the individual and area level, including driver behavior with the aim of optimizing risk.[3]

The increased availability of high-resolution spatial data generated by mobile phones provides a unique opportunity to enhance the granularity of these models in order to better assess driver risk. This report introduces the use of anonymized mobile phone data as an improved method of estimating driver risk profiles for the car insurance industry. This new approach will provide measures that can assist in establishing improved individual driver risk profile by removing risk bias based on population profiling. By doing so, this approach will move the car insurance field closer to a user-based risk profiling. [3]

Current Approaches to Assessing Driver Risk

Currently insurance companies use several indicators to build a driver risk profile.[4] Most of these indicators pertain to individual driver characteristics, such as historical driving record, gender, employment status/type, make/model of vehicle, marital status, and place of residence.[3] Table 1 describes how each of the variables contributes to the assessment of driver risk.

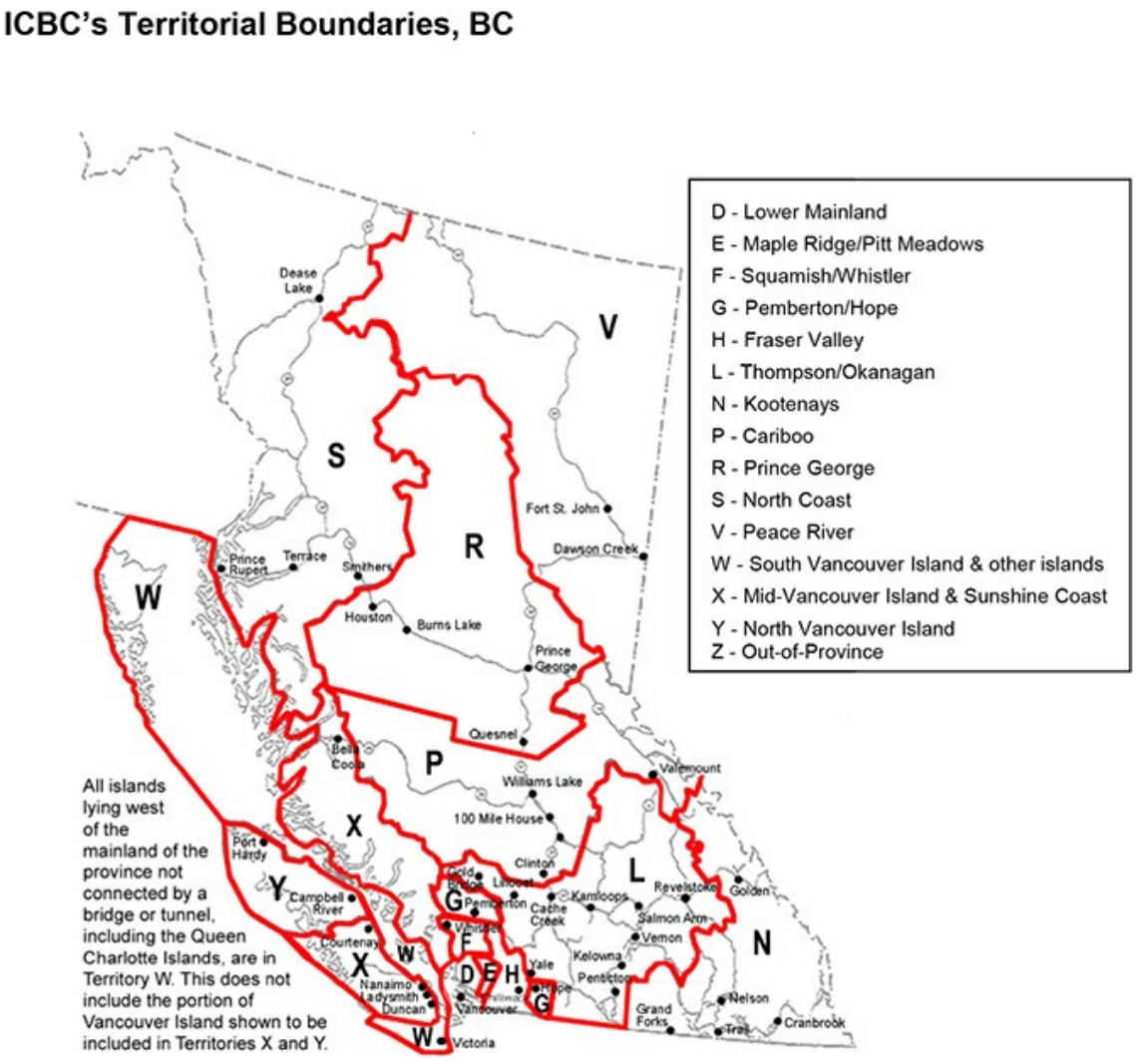
Table 1. Variables Contributing to the Assessment of Driver Risk [5]

Car Characteristics	Age, Manufacturer, Value and Safety Features
Type of Coverage and Deductibles	Liability Limits, Uninsured Motorist, Collision and Whether the Deductible is High
Driver profile	Age, Gender, Marital status, Place of residence and driving record
Car Usage	Commute for Work or Pleasure

While most of the information derived from these variables is extremely valuable in predicting driver behavior and risk, there are several limitations associated with the use of the 'place of residence' variable. Currently, the 'place of residence' variable is formulated using historical information on accidents and insurance claims that occurred within the Zip Code (in the US) or another similarly large geographical area associated with the driver's place of residence.[3] For example, in British Columbia, Canada, areas that are located in urban areas are assigned higher levels of risk because they are associated with higher risk of accidents and larger numbers of claims. However, as these areas can be fairly large in size, this method provides a broad estimation of risk at best. (Fig-

ure 1) A more accurate approach would be to accident risk based on the **aggregated behavioral / usage patterns of drivers** within more granular geographic areas, rather than historic claims profile of the larger urban area within which they reside.

Figure 1. Driver risk zones used by the Insurance Corporation of British Columbia (ICBC). British Columbia (944,735 km²) with its 5 million residents divided to only 20 risk zones.



Usage-Based Insurance

To better understand driver risk and tailor insurance to personal behavior, insurance companies have introduced a new model of insurance, called Usage-Based Insurance (UBI).[6] With UBI, drivers pay per kilometer/miles driven, so this type of insurance tends to benefit drivers who travel less. In addition, drivers are rewarded for safe driving habits through reductions in the cost of their insurance premiums.[6] However, as part of UBI, insurance companies also require the driver to install a tracking device on their car, along with an app that tracks driver habits, such as location (even when they are not driving). This requirement introduces some privacy concerns which, not surprisingly, are a deterrent for many drivers. In fact, a 2016 Pew research center survey found that while many Americans were "willing to share private information in exchange for tangible benefits", only 37% of the survey respondents were willing to give their location.[7]

Anonymized smart phone data

An estimated 85% of adults in the US use a smartphone.[8] The exponential increase in smartphone use over the past 15 years has brought with it the increased potential to collect large amounts of anonymous mobile phone data (both spatial and non-spatial). This data is typically collected passively through smartphone applications ('apps') that use location as a key feature and is becoming increasingly available from commercial vendors for marketing, commercial and academic research purposes.

To protect privacy and adhere to regulations such as GDPR and CCPA, aggregated mobile location data is anonymized to conceal personal attributes. Therefore smartphone data typically does a poor job of representing personal attributes (e.g., characteristics of smartphone users) but provides good information on the device's location

throughout the day. Nevertheless, this type of locational data – when processed by dedicated machine learning algorithms that segment modes of transportation and correlate with demographic models – has the ability to generate new information about driver behaviour. For example, by combining GPS coordinates with corresponding time stamps, Downtown.AI can calculate how much time a mobile user spent at a specific location, their mode of travel and their speed. Moreover, using machine learning techniques to analyze this data over time can provide insight into driver behaviour, risk and patterns, help identify the location of high-risk areas based on driver behavior, and ultimately assist in better estimating driver risk behaviour by enhancing current driver risk models.

Objective

In this research report, we present a new approach to the assessment of driver behavior which uses mobile location data in addition to traditional variables in the determination of driver risk profiles. This approach is intended to improve the accuracy with which driver risk profiles are calculated and has the potential to reduce costs for both insurance companies and drivers. More importantly, this approach protects driver privacy by leveraging large amounts of aggregated crowd data rather than individual mobile location data, and offers the drivers rates which more accurately reflect their actual risk by considering geography in a very granular way.

Methodology

Mobile phone data from over 50,000 mobile devices during March 2018 to 2019 were used for this project. This data represents approximately 4% of the adult population of mobile phone users in greater Vancouver, Canada. Previous research from both the US and Canada has shown that this data represents population demographics with 95% accuracy. Excluded from this data are those aged 18 and younger, and those aged 80 years and older.

Data Preparation

Table 2 shows the data attributes. The mobile phone ID is a unique identifier that allows tracking of individual mobile phones over time. To prepare the data for analysis, we first removed points that did not meet a minimum usage threshold. This allowed us to generate GPS tracks that could be analyzed for speed, acceleration and bearing. Using machine learning algorithms, we were able to create categories for different types of movement (vehicular, pedestrian, cyclist, etc.) from GPS point trips. This created a subset of the data that included only vehicular trips. More specifically, by calculating trip speed and acceleration, the algorithm excluded pedestrian traffic, and public transportation from the dataset. Finally, the GPS points associated with each trip were snapped to the road network in greater Vancouver.

Table 2. Example of smartphone data with location and timestamp

DeviceID	Latitude	Longitude	UTC DATE	UTC TIME
ID12545241	47.66156	-117.405	1/19/2018	21:41:44
ID12545242	47.66168	-117.405	1/19/2018	01:11:53
ID12545243	47.66155	-117.405	1/19/2018	12:32:04
ID12545244	47.66148	-117.405	1/19/2018	22:45:15
ID12545245	47.66147	-117.405	1/19/2018	17:42:21
ID12545246	47.66137	-117.405	1/19/2018	13:12:31
ID12545247	47.66136	-117.405	1/19/2018	15:02:49
ID12545248	47.66159	-117.406	1/19/2018	08:53:54
ID12545249	47.66169	-117.406	1/19/2018	17:43:30

Variables

Vehicle Kilometers Travelled (VKT): the distances between each point within each track are summed, on a per user basis. The result is proportional to the total actual VKT of that user over the study period.

Average Speed: each point-to-point distance is divided by the difference in time between the two points, generating an instantaneous speed. These speeds are then averaged over the nearest three points before and after each point (if present) to create a smooth average speed at any given time.

Absolute Acceleration: based on the instantaneous speeds above; the absolute value of the point-to-point difference in speed is divided by the time elapsed, generating an instantaneous acceleration.

Road Risk: The road risk behaviour variable represents the number of times a driver traverses a section of roadway on which there are many accidents. This variable was generated by taking the path represented by each user's data points over the study period and counting the number of times that it passes over a location where an accident has been reported in the past five years. This variable differs significantly from simple aggregating the number of crashes in a user's home geographic area as many users (ie. commuters) regularly traverse areas far from their home location.

Insurance Corporation of British Columbia (ICBC) Crash Data: The locations of all reported car crashes in BC for the years 2016 through 2020 were obtained from ICBC's online portal. This data is publicly available.

Analysis

We used Uber's Hexagonal Hierarchical Spatial Index mapping system to estimate each of the variables (VKT, average speed and acceleration) for each driver.[9] The location of the hexagon represents the zone within which each driver in the dataset resides. Place of residence is calculated using our property algorithm which estimates the driver's "home zone" using aggregated points at one location during nighttime hours. We also aggregated the ICBC crash data and the estimated road risk within each hexagon and then assessed the relationship between road risk and historical crash data.

Results

Approximately 50,000 devices, yielding between 300 and 5000 location points per day per device, were used for this analysis. Over a one-year period, the total estimated GPS points used for this analysis numbered over 9 billion. Figures 2-4 show the different variables that were created using our algorithm as well as their spatial distribution. These figures are screen shots from our interface that allow near real-time analysis of the data points as well as the ability to generate customized reports based on different time periods, locations and variables of interest.

Figure 2

Map showing VKT across greater Vancouver generated from mobile phone data. Hexagons in red indicate areas where drivers reside and tend to spend more time driving.

Figure 3

Map showing driver speed across greater Vancouver generated from mobile phone data. Hexagons in red indicate areas where drivers reside and tend to speed more frequently.

Low **High**

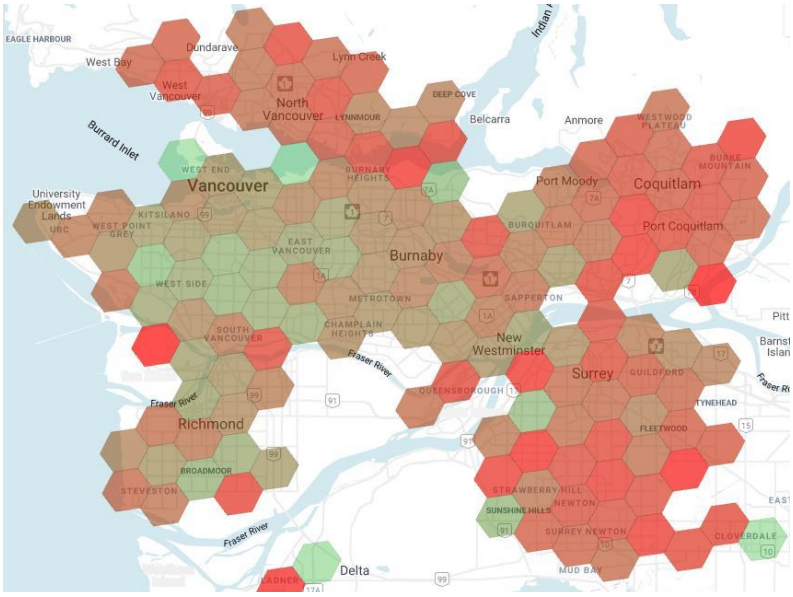


Figure 4

Map showing driver acceleration across greater Vancouver generated from mobile phone data. Hexagons in red indicate areas where drivers reside and tend to accelerate more frequently.

Low **High**

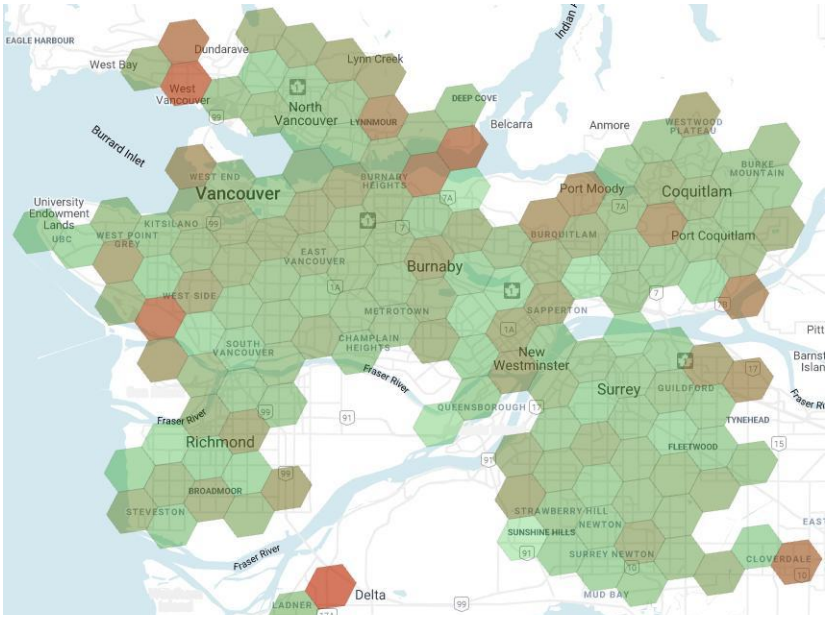


Figure 5

Map showing crash risk areas across greater Vancouver generated from mobile phone data. Hexagons in red indicate areas where drivers reside and have higher level of crash risk.

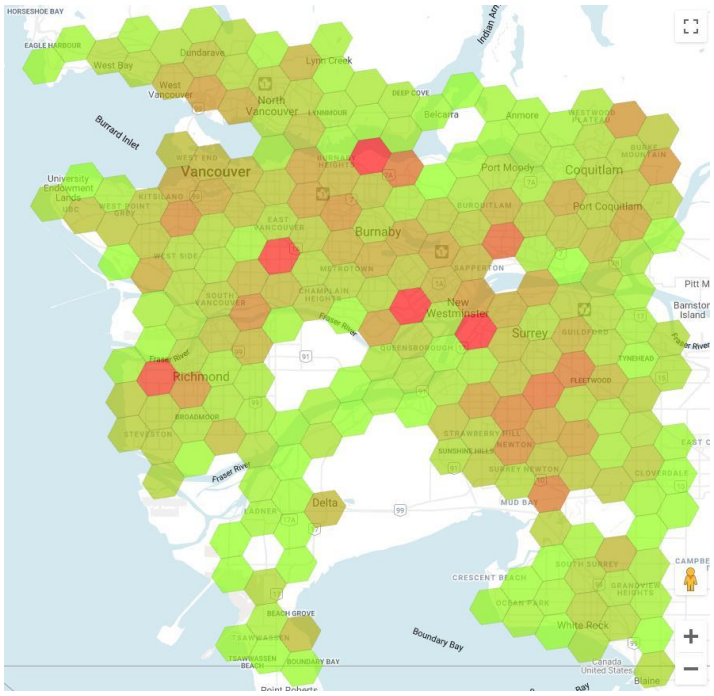


Figure 6 and 7 shows the relationship between aggregated insurance claims counts and average VKT and crash risk estimations in eight postal forward sortation areas in greater Vancouver. The insurance claims counts were obtained from the Insurance Corporation of British Columbia (ICBC) over a 5 - year period.

Figure 6

Show the relationship between aggregated insurance claims data and average VKT estimation from approximately 50000 mobile devices in eight areas within greater Vancouver.

VKT: Vehicle Kilometers Travelled

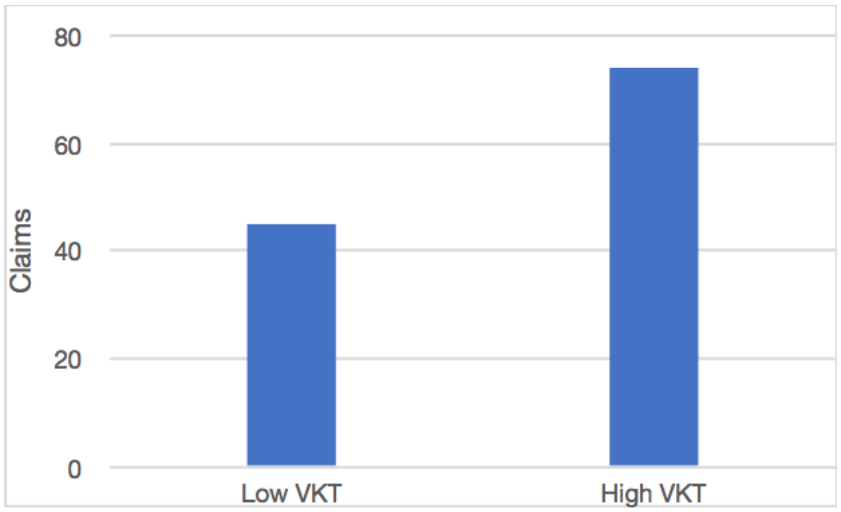


Figure 7

Show the relationship between aggregated insurance claims data and average crash risk estimation from approximately 50000 mobile devices in eight areas within greater Vancouver.

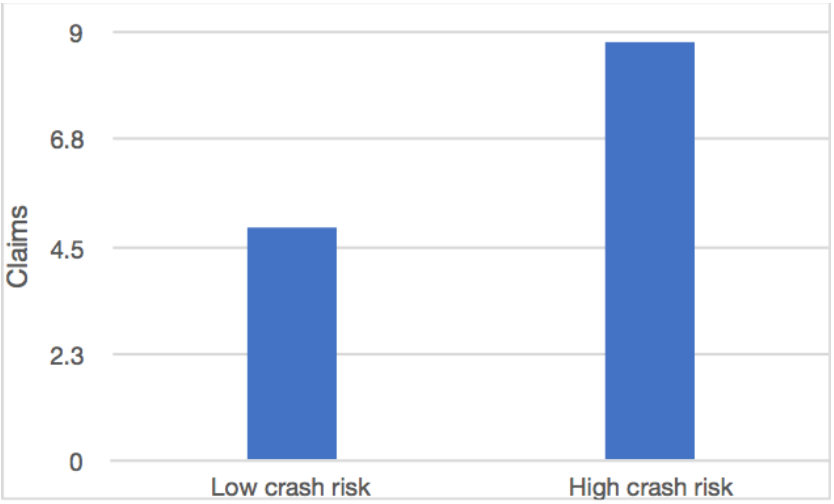
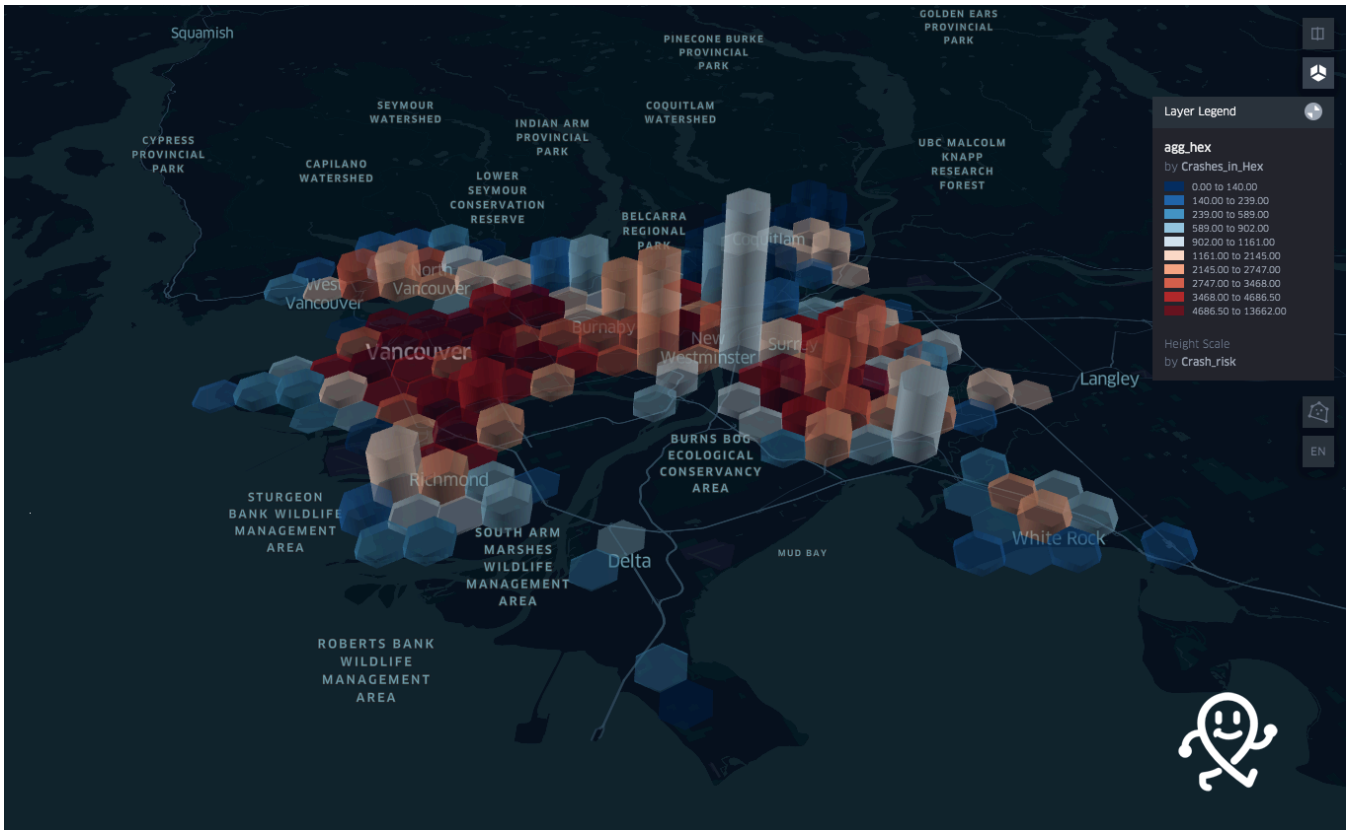


Figure 8 shows the relationship between aggregated reported car crashes and crash risk estimations in greater Vancouver. There is no clear correlation between areas with a high number of reported car crashes and those where estimated crashes are high based on driver risk. This indicates that most crashes occur in areas away from drivers' home locations, and that important data is being ignored if one considers only crash rates in the area of the drivers' home addresses.

Figure 8

Showing the relationship between aggregated ICBC crash data and average crash risk metric from approximately 50,000 mobile devices in eight areas within greater Vancouver.



Discussion

This report introduces the use of anonymized mobile phone data as an improved method of estimating driver risk profiles for the car insurance industry. Using a combination of spatial analysis and machine learning algorithms, we created new variables representing driver risk behaviour which, together with currently used variables (age, driving history, etc.), will allow insurance providers an improved means of assessing driver risk. As far as we know, this is the first use of anonymized mobile phone data for this purpose.

In this report, we describe four measures (VKT, acceleration, speed and crash risk) that can be used to improve the calculation of driver risk profiles. These measures were generated at a higher spatial geographic resolution than is currently utilized within the industry. Currently, insurance companies are using data at a regional level (or other large geographic areas) to drive decision making regarding high-risk areas. This in turn creates a situation where individual drivers are linked to the risk profile for an entire region, regardless of the risk profile for their specific neighbourhood. The approach proposed here overcomes this issue by assigning each driver to a much smaller geographic area, which allows rates to be structured to more accurately reflect each driver's true risk. In addition to the variables presented here, spatial analysis based on mobile phone data can yield additional indicators that can assist in estimating driver risk. For example, using historical data on areas where multiple accidents have occurred, we can identify the areas from which drivers' trips originate. This information can then be provided to insurance companies to assist them in developing more individualized risk profiles.

This study also revealed an interesting relationship between reported ICBC car crash locations and estimated driver risk. Figure 8 indicates that while ICBC reported crashes

are most heavily concentrated within the downtown Vancouver area, estimates of driver risk are currently based on home location, which is typically outside of this area. This indicates that drivers are more likely to experience a car accident if they drive to the downtown area. This should lead insurance companies to charge higher premiums based on the crash risk of the areas to which drivers travel rather than on where they reside.

In this report, we also examine the relationship between VKT, the key variable created in this study, and claims data from eight forward sortation areas in greater Vancouver. The results, based on five years of data, show that VKT is positively correlated with claim data and indicate that our newly created variable has the potential to predict high risk areas. However, given that we were only able to assess this relationship within eight zones, further analysis is needed to assess the predictive power of this variable.

References

[1] A. B. Ellison, M. C. Bliemer, and S. P. Greaves, “Evaluating changes in driver behaviour: a risk profiling approach,” *Accident Analysis & Prevention*, vol. 75, pp. 298–309, 2015.

[2] D. A. Cather, “Cream Skimming: Innovations in Insurance Risk Classification and Adverse Selection,” *Risk Management and Insurance Review*, vol. 21, no. 2, pp. 335–366, 2018, doi: https://doi.org/10.1111/rmir.12102.

[3] P. M. Ong and M. A. Stoll, “Redlining or risk? A spatial analysis of auto insurance rates in Los Angeles,” *Journal of Policy Analysis and Management*, vol. 26, no. 4, pp. 811–830, 2007, doi: https://doi.org/10.1002/pam.20287.

[4] O. Tufvesson, J. Lindström, and E. Lindström, “Spatial statistical modelling of insurance risk: a spatial epidemiological approach to car insurance,” *Scandinavian Actuarial Journal*, vol. 2019, no. 6, pp. 508–522, Jul. 2019, doi: 10.1080/03461238.2019.1576146.

[5] “Risk Profiling in the Auto Insurance Industry,” *Gracey-Backer, Inc.*, Mar. 14, 2017. https://www.graceybacker.com/risk-profiling-auto-insurance-industry/ (accessed Feb. 27, 2021).

[6] M. Juang, “A new kind of auto insurance technology can lead to lower premiums, but it tracks your every move,” *CNBC*, Oct. 06, 2018. https://www.cnbc.com/2018/10/05/new-kind-of-auto-insurance-can-be-cheaper-but-tracks-your-every-move.html (accessed Feb. 27, 2021).

[7] 1615 L. St NW, Suite 800Washington, and D. 20036USA202-419-4300 | M.-857-8562 | F.-419-4372 | M. Inquiries, “Auto trackers not worth car insurance discounts, most say,” *Pew Research Center: Internet, Science & Tech*, Jan. 14, 2016. https://www.pewresearch.org/internet/2016/01/14/scenario-auto-insurance-discounts-and-monitoring/ (accessed Feb. 27, 2021).

[8] 1615 L. St NW, Suite 800Washington, and D. 20036USA202-419-4300 | M.-857-8562 | F.-419-4372 | M. Inquiries, “Demographics of Mobile Device Ownership and Adoption in the United States,” *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/fact-sheet/mobile/ (accessed Feb. 27, 2021).

[9] I. Brodsky, “H3: Uber’s Hexagonal Hierarchical Spatial Index,” *Uber Engineering Blog*, Jun. 27, 2018. https://eng.uber.com/h3/ (accessed Feb. 27, 2021).

