GPS Trajectory Feature Extraction for Driver Risk Profiling

Johannes Paefgen
University of St. Gallen
Dufourstrasse 40A
9000 St. Gallen, Switzerland
johannes.paefgen@unisg.ch

Florian Michahelles
ETH Zurich
Scheuchzerstrasse 7
8092 Zurich, Switzerland
fmichahelles@ethz.ch

Thorsten Staake
ETH Zurich
Scheuchzerstrasse 7
8092 Zurich, Switzerland
tstaake@ethz.ch

ABSTRACT
In this paper we develop a method and relevant feature constructs for the measurement of accident risk exposure from a large sample of real-world GPS data that includes accident and accident-free drivers. For trip frequency and accumulated driven distance features, an evaluation of their discriminatory power is given based on computational results. In our conclusion, we briefly discuss suitable classification approaches and limitations arising from external validity considerations.

Author Keywords
Usage-based Insurance, Feature Extraction, Accident Risk, Spatio-temporal Trajectories, Driver Behavior

ACM Classification Keywords
H.4.2. Types of System: Decision Support.

General Terms
Algorithms, Economics, Human Factors, Management, Measurement.

INTRODUCTION
Ubiquitous sensors allow for a more precise estimation and monitoring of insurance risks. In automotive insurance, usage-based premium schemes that employ remotely collected data to adapt to actual measurements of risk exposure have been under consideration for several years. For this application, the derivation of risk-related information from spatio-temporal vehicle trajectories is a critical prerequisite. While researchers have pointed out significant macroeconomic benefits of adaptive premium schemes for society [1,2], its operational realization has not yet received much attention from academia so far, though it holds several interesting problems that justify a scientific treatment of the matter. Besides privacy considerations, technology costs are a major hindrance in the implementation of usage-based premium schemes. In our research, we explore to what extend a stand-alone GPS sensor unit suffices for the cost-effective profiling of accident risk amongst drivers.

Based on a comparatively large sample of 1500 drivers over 2 years obtained from a telematics provider in northern Italy, which includes a significant amount of accident drivers, our paper discusses a data aggregation and feature extraction approach for driver risk profiling. While an abundance of literature exists on antecedents of traffic accidents, our focus lies on the method and constructs required to derive a reliable and empirically grounded implementation for this specific application. Ideally, risk profiles can be mapped to a linear scale in a bounded interval from which premiums can be directly calculated. Towards this end, we propose a consistent set of trajectory features and discuss some preliminary computational results based on the available dataset.

UbiComp' 11, Sep 17 – Sep 21, 2011, Beijing, China.

RELATED WORK
Before the availability of in-vehicle sensor technology, conventional methods of accident analysis were largely limited to driver questionnaires and accident reports that reproduced the circumstances under which accidents occurred. The advent of GPS sensor technology has enabled new means for the investigation of individual mobility [3-5]. It allows for a continuous monitoring of mobility patterns, which are proven to exceed the accuracy of other data collection methods [6]. Nevertheless, empirically verified risk exposure metrics that extend the situational definition above remain sparse in the literature. The principal feasibility of driver assessment based on GPS-measured distances, velocity and acceleration was demonstrated in a laboratory setting by [7]. As part of the Commute Atlanta program, spatio-temporal activity patterns were first employed to explain crash involvement in a field study. Here, the same three parameters where extracted with a refined Kalman-filter based estimation algorithm leading to the conclusion that greater mileage, higher speeds and frequent, hard deceleration are correlated with accident risk [8]. In a more recent publication based on the same field study, the authors question their results due to a limited sample size and the inconsistent identification of accident drivers from questionnaires; however they emphasize the

general potential of their approach and call for an increased research effort, specifically mentioning usage-based insurance in their motivation [9].

Based on this brief literature survey, we conclude the following. Firstly, due the low probability of accident occurrence in a driver population, large GPS samples are required in order to observe a representative amount of accidents within the data. Secondly, as GPS-based spatio-temporal patterns are a relatively new source of accident risk exposure metrics, an exploratory approach is justified for feature extraction. And thirdly, the integration of different metrics to combined exposure metrics that exploit the full richness of GPS data remains an open challenge for the research community.

DATA AND METHODOLOGY

Our premise is that vehicles for which an accident was observed in the sampled period exhibit different trajectory features than accident-free vehicles. For both, GPS measurements were collected with on-board data recorders in continuous operation. In order to reduce data transfer costs and memory demands of the overall database, high resolution GPS measurements were aggregated on the device level in intervals of 2000 meters estimated driven distance and then transmitted. Furthermore, such data points were generated upon change of the vehicle ignition status. A data point consists of a unique device ID, a timestamp, latitude and longitude of vehicle location corresponding to that timestamp, driven distance and time since last stored data point, together with an indication of the ignition status of the vehicle and a road type parameter distinguishing urban, arterial, and highway facilities. The low resolution of generated data points is a limitation inherent to the infrastructure employed for data collection and prohibits us from extracting certain accident risk exposure metrics. However, we assume that the information density of the low resolution data still suffices to explain a significant fraction of variance between accident and no-accident drivers. Also, note that driven vehicle distance is captured much more precisely than stipulated by the transmitted latitude and longitude due to the fact that data points contain the accumulated driven distance estimate from the disregarded high resolution measurements.

Trip Level Aggregation

Spatio-temporal trajectories represented by the raw data have to be further aggregated to a level where patterns useful for the risk profiling problem can be extracted. In the case of vehicles, one natural aggregation level that is common in the literature stems from the notion of a vehicle *trip*, i.e. a time-ordered sequence of position measurements whose spatio-temporal distances lie within a lower bound threshold defined by some maximally feasible vehicle velocity v_{max}, so that $\Delta d / \Delta t < v_{max}$ where Δd and Δt correspond to distance and time differences of the data points, and furthermore within an upper bound threshold ΔT that temporally separates trips based on a significant sojourn time at trip destination. Intuitively, ΔT should differentiate brief vehicle stops that belong to the same trip, e.g. at a red traffic light, from longer vehicle stops that mark the end of a trip, i.e. parking of the vehicle.

Column Parameters	Unit
Vehicle ID	-
Timestamp Trip Start / End	Matlab datenum
Trip Driven Distance	meters
Trip Duration	seconds
Latitude and Longitude of Trip Start / End	radians
Number of Data Points for each Road Type	-

Table 1. Trip matrix column structure

In our dataset, trip aggregation is supported by the ignition status indicator variable that augments the position measurement in the database. We thus do not explicitly employ the thresholds defined above, but separate trips based on the ignition status, where an ignition-on event denotes the beginning of a trip and an ignition-off event denotes its termination. After this separation, we employ a range of consistency checks to refine the quality of our trip aggregation algorithm. A special case arises when trip start and termination points are missing, e.g.. due to GPS signal obstruction at vehicle stopping point, which typically causes two subsequent trips to be misinterpreted as one uninterrupted sequence. This can be circumvented by considering some spatio-temporal threshold (Δd, Δt) where we chose (10 meters, 60 seconds) as a lower bound criterion for trip separation. Simultaneously, this choice yields an upper bound criterion for the merger of trips with close proximity due to mid-trip ignition status resets. We omit a detailed algorithmic discussion of further common consistency checks that were employed. As a result of the trip aggregation phase, approximately $n = 2.4 \times 10^6$ separate trips where identified and stored in a n-by-m trip matrix, where the number of columns m corresponds to the number of trip description parameters such as given in Table 1. Again, it should be noted that the aggregated driven distance is not equal to the distance between trip start and end points.

There is strong evidence in the literature that as a consequence of an accident, drivers reduce vehicle usage and generally change their driving habits, confer to [10], for instance. This justifies the assumption that accident risk potentially decreases after an accident has occurred. It is therefore reasonable to remove trips from the aggregated matrix that were recorded after an accident has been recorded for a specific vehicle ID. To this end, all entries with a trip start timestamp corresponding to the accident month or any month thereafter were not considered in our analysis. However, it is an interesting sub-problem to quantify post-accident driving behavior change based on

the sample data set. Once a better understanding of these dynamics has been achieved, the corresponding samples should remain in the dataset, possibly marked as accident-free drivers.

Feature Extraction

Based on the trip matrix, we develop a set of algorithms to extract features for risk profiling in accordance with existing literature on antecedents of traffic accidents. These algorithms operate on subsets of the overall trip set that consist of n' trips with common vehicle ID and month as given by the trip start timestamp. For every n'-by-m subset, the expression of a feature i is captured by a scalar, normalized value f_i in {0;1} that is computed based on some properties of its contained trip parameters. In the most straightforward case, features can be obtained through simple arithmetic operations. In the more complex cases, namely for features constructed from spatial patterns in the trip subset, intermediate data structures have to be generated from the trip parameters. In the following, we briefly discuss the features considered in our analysis so far.

Trip Frequency and Accumulated Exposure Measures

For every vehicle-month subset, we trivially count n' and subsequently divide all n' by a reasonably chosen maximum upper bound for trip frequency, e.g. 10^3 / month, so that the resulting feature value does not exceed 1 in any subset. This feature follows the notion that some measures of risk exposure such as driver and vehicle state, environmental conditions etc. only change in between trips and are fixed for an individual trip. Each trip therefore represents a random draw, where an increased number of trips leads to a higher probability of high-risk realizations. In an alternate approach, we accumulate the driven distance over all trips in a subset and again normalize against an upper bound such as 10^4 m / month. As opposed to trip frequency, the distance feature then captures the *ceterim paribus* probability of an accident per driven meter. As a higher number of trips typically increases the accumulated distance, a high correlation between these two features has to be expected. Analogously, we accumulate the overall time in a month during which the vehicle was operated. Again, a high correlation with the previous two features has to be expected. Nevertheless, we propose that all three features are sufficiently distinct to justify a simultaneous consideration, ideally by a risk profiling method that can exploit prevalent explanatory differences.

Temporal and Spatial Patterns

As stated above, previous work indicates that accident risk follows periodic temporal patterns. As we limit our analysis to monthly data subsets, the two reasonable periods to be considered are weeks – i.e. the distribution of vehicle operation over weekdays and weekends – and days – i.e. the distribution over daily hours. The latter contains driving situations during night, early morning and rush hours as special cases. For temporal patterns, the determination of a normalized scalar feature value becomes more complex. In our approach, we count vehicle operating time derived from trip start and duration parameters over 24 and 7 bins, respectively. This procedure is again prone to produce a range of highly correlated sub-features, that have to be weighted in the risk profiling method. An additional feature in the temporal pattern category that may be derived from the given data is driver fatigue – which would accumulate over long trips or trips with short breaks in between, and return to a minimal value over longer breaks. One basic spatial feature is the proportion of trips spend on various road types. For this purpose, the trip aggregation algorithm accounts for the distribution of road type indicators over data points of each trip. So far, we did not yet include more complex spatial pattern features in our analysis. Promising approaches include the distribution of trip destinations, which would allow for a discrimination of frequent an infrequent routes taken by a driver. Ultimately, high risk regions maybe identified from the dataset, though this approach appears somewhat hindered by the low resolution of the aggregated position data.

PRELIMINARY RESULTS

For two selected features, trip frequency and accumulated distance, we computational results for our dataset. To make the visualization more tangible, normalizations have been reversed so that x-axis labels show actual units of the distributions.

In Figure 1, the distribution of the trip frequency feature overall vehicle-month subsets is depicted for both trip groups, accident and accident-free drivers. By inspection, it follows a log-normal density function. The right-shift in the distribution for accident-free drivers is clearly visible. From a simple linear discriminant analysis, a decision boundary around 110 trips per month yields a confusion matrix with 65.4% of accident-free drivers and 55.7% of accident drivers classified correctly. In Figure 2, the distribution of accumulated distance, i.e., the driven distance in a vehicle-month is depicted for both groups. Here, the log-normal distribution quality is even stronger pronounced. For a decision boundary around $2x10^6$ meters, 66.5% of accident-free drivers and 67.0% of accident drivers are classified correctly. One interesting observation in this figure is that beyond a certain threshold of around $9x10^6$ meters, accident-free drivers are again more frequent than accident drivers. This may be attributed to the fact that very high monthly mileage values – in fact more than 9000 kilometers – are typically found amongst professional drivers that, due to their driving experience and skill, fall in a low-risk category. The correlation coefficient between the two features is 0.46 over the entire dataset. The generally observable right-shift in the exposure curves is explained by the fact that driving style only contributes to a limited extend to accident risk. As accidents are caused by a variety of highly random events or situations that occur with some probability, e.g., per intersection or per mile, the more frequent a vehicle confronts these events and situations the more likely it is to be involved in an accident.

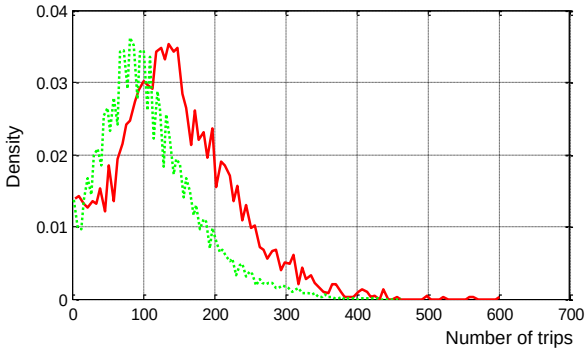

Figure 1. Trip frequency distribution over vehicle-month subsets. Dotted line: Accident-free drivers

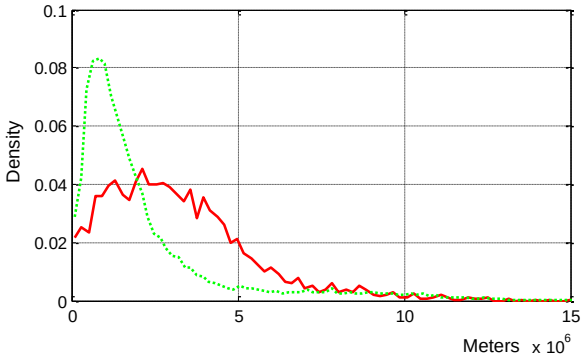

Figure 2. Accumulated distance distribution over vehicle-month subsets. Dotted line: Accident-free drivers

CONCLUSION

In this paper, we have proposed a method and constructs for the extraction of features from GPS trajectories to address the driver risk profiling problem based on a large-scale set of empirical data. We would like to point out that our analysis so far focused on features of an unidirectional, linear structure that are suitable for fundamental methods of statistical analysis and correspond to an intuitive understanding of accident risk exposure. These features are bound to be highly correlated in several cases, and of limited explanatory power, e.g., due to the separation of temporal and spatial patterns. Multi-dimensional, non-linear features may offer more structure to capture the complexity inherent in the problem. One could, for instance, establish more lower-level features on trip level to augment the trip matrix. With respect to risk profiling, initial results are promising, yet demonstrate the limited explicatory power of individual features. The depicted distributions display a significant overlap, which confirms that in spite of a range of justified predictors, accidents are determined by complex and highly random interactions in the real world, and consequently a binary classification approach is unreasonable. Further challenges in this direction are the elimination or fusion of highly correlated features and the development of representations that allow for a higher degree of complexity. Particularly, this could mean to develop more sophisticated trip-level features to be aggregated in the monthly domain.

Finally, we remark that at the present state our evaluation lacks external validity for a variety of reasons. The selection of vehicles in the dataset was biased towards an overrepresentation of accident drivers. Furthermore, vehicle-month subsets where not equally distributed over the year. Both these issues can be addressed by a stratified sampling of the dataset before the feature extraction procedure. Another limitation is a lack of representativeness towards driver demographics. While all driver data was anonymized and inaccessible to the authors of this paper, the fact that the vehicles where equipped with GPS devices may impose a certain bias on our sample. Also, a vehicle may have been operated by several drivers. However, for an application in insurance this is acceptable, as the premium calculation should take into account the actual risk exposure of the vehicle and not some specific attributes of its owner.

REFERENCES

[1] T. Litman, “Distance-Based Vehicle Insurance As A TDM Strategy,” *Transportation Quarterly*, vol. 51, 1997, pp. 119-138.

[2] I.W.H. Parry, “Is Pay-as-You-Drive Insurance a Better Way to Reduce Gasoline Consumption than Gasoline Taxes?,” *American Economic Review*, vol. 95, May. 2005, pp. 288-293.

[3] J. Yuan, Y. Zheng, C. Zhang, and W. Xie, “T-drive: driving directions based on taxi trajectories,” *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, San Jose, CA: ACM, 2010, pp. 99-108.

[4] W. Liu, Y. Zheng, S. Chawla, and J. Yuan, “Discovering Spatio-Temporal Causal Interactions in Traffic Data Streams,” *KDD□11: 17th ACM SIGKOD Conference on Knowledge Discovery and Data Mining*, San Diego, CA: ACM, 2011.

[5] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma, “Understanding mobility based on GPS data,” *Proceedings of the 10th international conference on Ubiquitous computing - UbiComp ’08*, 2008, p. 312.

[6] J.H. Ogle, “Quantitative assessment of driver speeding behavior using instrumented vehicles,” Georgia Institute of Technology, 2005.

[7] M. Porter, M. Whitton, and D. Kriellaars, “Assessing Driving with the Global Positioning System: Effect of Differential Correction,” *Transportation Research Record*, vol. 1899, Jan. 2004, pp. 19-26.

[8] J. Jun, J. Ogle, and R. Guensler, “Relationships Between Crash Involvement and Temporal-Spatial Driving Behavior Activity Patterns: Use of Data for Vehicles with Global Positioning Systems,” *Transportation Research Record*, vol. 2019, Dec. 2007, pp. 246-255.

[9] J. Jun, R. Guensler, and J. Ogle, “Differences in observed speed patterns between crash-involved and crash-not-involved drivers: Application of in-vehicle monitoring technology,” *Transportation Research Part C: Emerging Technologies*, vol. 19, Oct. 2010, pp. 569-578.

[10] R. Mayou, B. Bryant, and R. Duthie, “Psychiatric consequences of road traffic accidents.,” *British Medical Journal*, vol. 307, 1993, p. 647.