REPORT OF RESEARCH RESULTS

A. Title

Exposure-based Insurance System based on Accident-prone Map and User Travel Trajectories

B. Primary Researcher

Chia-Jung Yeh (chia-jung.yeh@sydney.edu.au)
Master Student, National Yang Ming Chiao Tung University, 2021-2023
PhD Candidate, University of Sydney, 2023-

C. Summary

Usage-based insurance (UBI), which takes driving distance and behaviour into consideration, has raised attention and has been applied in recent years. However, some drawbacks of this method should be addressed. First, the data is not easy to collect. Second, the link between driving behaviour and crashes may be weak. Last, since the geometry design, traffic flow and the building factors of each road network are not identical, there is a huge difference in crash frequency and severity. Therefore, the "exposure" of the drivers on distinct crash frequencies of the roads would have different accident risks. Based on the aforementioned, unlike the traditional UBI, the research will develop a framework of exposure-based insurance systems based on the crash likelihood map and user travel trajectories. First, the crash likelihood map is built through the joint model architecture to further estimate the crash frequency and severity of each network link. Then, map matching via the Hidden Markov model (HMM) and mode classification through the machine learning process are applied to retrieve the trajectories of private vehicles. Last, the relative risk of 60 surveyees is evaluated by the trajectory and risk matrix according to the crash likelihood map. This risk evaluation is then transformed into the premium adjustment factor to serve as the discount rate for different types of drivers. The research is dedicated to integrating the techniques of data collecting and processing to break through the constraints of UBI.

D. Aim of Research

The main drawback of the current UBI mechanism lies in the privacy and information security issues of the collection of positioning data. This makes UBI less attractive for the insured, resulting in a less successful promotion of UBI. Therefore, this study intends to construct an exposure-based insurance mechanism based on travel exposure and crash likelihood maps. In addition, to consider the privacy issue, cellular data should be authorised by the insured and processed by a third party (i.e., service provider) to avoid data breaches. The insurance company would only obtain the risk score of each insured and take it as the discount rate of the premiums. The proposed mechanism is expected to accurately reflect the driving risk and serve as the basis for calculating insurance premiums objectively.

E. Method of Research & Progression

1. **Crash Likelihood Map**
   The joint model is applied to predict the crash frequency and severity of each network link by developing the econometric model based on the historical crash data. The independent variable is collected to explain the causality effect. Note that since the availability of external variables depend on the road hierarchy, the whole model is separated into four sub-models, including freeway, expressway, provincial highway, and city road. Afterwards, the risk matrix is developed based on frequency and severity to determine the relative risks for each link.
2. **Map Matching and Mode Detection**
   Map matching and mode detection are the processes to retrieve the trajectory and distinguish the mode through cellular data. The map matching process utilises Hidden Markov Model and Viterbi Algorithm, while the mode detection process applies machine learning techniques. These two procedures are used to filter the routes by private vehicles, which are associated with the driving risks of the insured. Afterwards, intersect the trajectories with the crash likelihood map to calculate the total risk result from passing through the links.

F. Results of Research

1. **Crash Likelihood Map**
   (1) For the freeway crash model, road length, traffic flow, length of climbing lane, total lane width, whether service area exists, and whether lies in a metropolitan are the variables positively correlated to the crash frequency. In terms of the severity, the percentage of heavy cars, total lane width, and whether exists speed camera are the significant variables that influence the severity levels. As for the model performance, the MAE is about 4.55, and the NMSE is about 0.795, which is acceptable. The accuracy of the prediction of the percentage of death-involved crashes is nearly 82.2%,

which points out a good severity model.

(2) For the city road crash model, the spatial stratified sampling technique is applied to deal with the huge number of samples. The crash model for city roads is tested 300 times, and then select the best model with the highest SMAPE (0.395). The final model shows that road length and width, whether exists bike lanes or sidewalks, commercial density, and the distance to the nearest intersection are the significant variables for the crash frequency. In addition, increasing the distance to the nearest intersection causes a negative effect on the crash frequency, while it leads to a more severe crash.

(3) By inferring the frequency and severity based on the joint model, a 5*5 risk matrix can then be formed to evaluate the road risk. Since there is a significant long-tail distribution in the crash frequency and severity, the research applies the head/tails breaks algorithm to classify all the values. The risk matrix and the crash likelihood map are shown in Figure 1 and Figure 2, respectively.

Crash Severity \ Crash Frequency	0	[1~3)	[3~14)	[14~52)	≥52
≥44.55%	Medium N=2,282 5	High N=0 10	Very High N=2 15	Extreme N=1 20	Extreme N=1 25
[12.72%~44.55%)	Medium N=4,620 4	Medium N=4 8	High N=2 12	Very High N=1 16	Extreme N=11 20
[3.38%~12.72%)	Low N=30,734 3	Medium N=231 6	Medium N=163 9	High N=43 12	Very High N=145 15
[1.03%~3.38%)	Very Low N=179,215 2	Low N=18,880 4	Medium N=6,362 6	Medium N=1,938 8	High N=455 10
[0%~1.03%)	Very Low N=870,553 1	Very Low N=239,713 2	Low N=62,350 3	Medium N=4,947 4	Medium N=1,149 5

Figure 1. Crash Risk Matrix of Taiwan Road Network

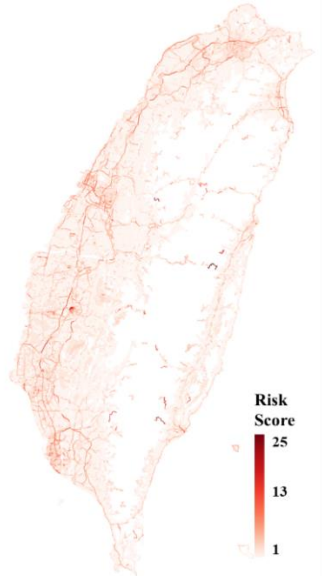

Figure 2. Crask Likelihood Map (Risk Score)

2. Map Matching and Mode Detection

(1) HMM is applied to retrieve the original trajectories of each traveller. Note that parameters σ_N and β in HMM should be calibrated in the emission and transition probability. Multiple numerical experiments are conducted to find a set of parameters that can maximise the accuracy of the map matching process. It is found that the parameters σ_N=90, β=90 are the optimal set, in which the accuracy would be up to 60%.

(2) The study also finds that the accuracy of map matching would be substantially lower with a shorter travel distance or with a slower mode such

as walking, bike, or motorcycle.

(3) Machine learning models are applied to test the appropriateness to detect the mode used for each trip. It is found that Random Forest and XGBoost are the two methods with the highest F1 score, and hence, this study adopts the assembly model of both approaches to enhance the accuracy up to 78.2%.

(4) In terms of the overall accuracy of each mode, bus and motorcycle have the worst performance, and this may be due to strong feature similarity between these modes.

3. Risk Evaluation of Phone Users

This research finally integrates the result of the previous two models, intersecting the trajectories of private vehicles retrieved from the cellular data and the crash likelihood map to further calculate the total risks of each trajectory and phone user. Total risk is defined as follows: $\sum_i \mathrm{Risk}_i * \mathrm{Path\ Length}_i$. The maximum value of the total risk among all users is about 2600, and the maximum of risks per kilometre is about 4.5. Total risks can be used to serve as the discount rate of insurance premiums in the future.

G. Future Areas to Take Note of, and Going Forward

This study intends to integrate the techniques of data collecting and processing to break through the constraints of UBI, hoping that the car insurance system can be fairer and more reasonable. Some limitations of this research are addressed in the following. Crash analysis models are not as accurate as expected, which may lead to the misestimation of the risks. Hence, different econometrics models are suggested to develop to find out the appropriate one. As for the mode detection model, it is not easy to distinguish between car and motorcycle due to the high similarity of the features. It is important to distinguish two private vehicles to design a more customised insurance system in the future. Last, all the models have the error, according to the error propagation, the risk evaluation in this proposed framework might have a huge gap between the actual risk. Hence, it is vital to improve the accuracy and efficiency of all the models and algorithms to facilitate the real application of this study.

H. Means of Official Announcement of Research Results

Proposed as a dissertation for the degree of master in Traffic and Transportation in Department of Transportation and Logistics Management, College of Management, National Yang Ming Chiao Tung University.