

Hesburgh Library Award
Yingxin (Cindy) Liu

Traffic accidents remain one of the leading causes of death worldwide, resulting in approximately 1.19 million fatalities annually and accounting for an estimated 3% loss in global GDP (WPR, 2025). China is among the countries most severely affected by traffic accidents. According to the *China Statistical Yearbook 2021*, there were 244,674 road traffic accidents in China in 2020, resulting in 61,703 fatalities and 250,723 injuries. With the development of information technology, it is now possible to monitor and analyze driver behavior in real time by collecting instantaneous driving data. Using these technologies to manage and improve driving behavior can play a crucial role in reducing accident risks for all vehicles, including commercial trucks. Usage-based insurance (UBI) was developed as a pricing model that uses telematics data to assess driving behavior and environmental risk. While UBI models for private vehicles have been widely studied, their application to vehicle fleets remains limited despite the higher real-world risk associated with fleets.

In my senior thesis, I investigated how driving behavior and road environment factors influence the risk levels of commercial fleets in the Beijing region and proposed a new framework for risk analysis specifically targeting commercial truck fleets. To address this research question, I used GPS trajectory data and road environment datasets from 498 trucks across 36 fleets over 90 days to generate driving behavior and dynamic road condition features at the trip and driver levels. I applied regression models (logistic regression and LASSO) and tree-based machine learning models to identify key risk factors and build predictive models. Model performance was evaluated using testing data and feature significance. The results show that conventional actuarial variables, together with driving behavior and road environment features, are effective in predicting claims for business fleets. Incorporating additional features such as group differences, driver familiarity, and road crash indices further improves predictive performance.

Throughout my research journey, Hesburgh Library provided significant support in helping me successfully complete my senior thesis by offering stable computing resources and platforms for data analysis, workshops on academic writing and locating economic literature, and access to librarians and curators who guided me in identifying reliable databases.

My original telematics data are highly raw and extremely large in scale, consisting of millions of high-frequency (per-second) observations, including vehicle location, speed, and timestamp information over a period of approximately 1.5 months for around 500 vehicles. Successfully downloading and cleaning these data require a stable computing environment with substantial memory capacity, as

transforming raw data into a usable format is computationally intensive. My personal desktop computer was insufficient to meet these demands. Therefore, I relied heavily on the computing resources at Hesburgh Library, frequently working in the library's computing section to carry out my coding tasks. Over time, I came to have a "designated" computer that I regularly used and grew somewhat attached to.

In addition to data cleaning, my research required integrating driving behavior and road environment features into the telematics dataset by appending and joining external data sources. This process involved the use of APIs such as OpenStreetMap (OSM), Google Cloud APIs, and weather APIs, which further increased the computational workload. After I gained access to the computing resources at Hesburgh Library, my workflow improved substantially. These resources allowed me to make repeated API requests without failures, store large-scale datasets, and successfully complete the data analysis, including generating over 100 features from millions of observations and integrating multiple external data sources.

Hesburgh Library librarians also provided valuable guidance and consultation throughout my senior thesis journey. In particular, the availability of economics-specialized librarians was extremely helpful. During the fall semester, Librarian Michael Deike visited our senior thesis class and introduced a range of economic databases for accessing reliable data related to specialized topics such as economic surveys and macroeconomic factors. He also demonstrated effective strategies for using the OneSearch tool to locate relevant academic literature. The filtering functions in OneSearch were especially useful for finding reliable papers. Given that my research topics span multiple disciplines, including economics, statistics, transportation, and decision-making, keywords such as "UBI" often returned many untargeted results. Using filters such as journal title and subject area significantly improved the efficiency of identifying relevant papers focused on risk modeling.

Beyond workshops and general resources, Hesburgh librarians also provided personalized support in identifying specific databases. When I was working on integrating weather data with telematics data, I was trying to find an API capable of handling thousands of requests at once based on the location and timestamp information I had. The initial APIs I explored, such as WorldWeather Online, imposed a limited number of free requests, which were insufficient for my dataset unless I upgraded to a premium version. After consulting with Michael Deike, he connected me with Summer Mengarelli, a Data Services librarian at Hesburgh. She helped me identify alternative data sources and APIs that could better accommodate large-scale requests without requiring paid access. With her guidance, I first used the WorldWeather Online API for initial testing and later identified and implemented the Open-Meteo API,

which supported the full scale of my thesis analysis. I would also like to thank Michael and Summer for their help in identifying the relevant databases that I ultimately used in my research.

Reference

China C.D.C. China Science and Technology Press; Beijing: 2021. China's cause of death surveillance dataset 2021.

World Population Review. (2025, October 28). *Road deaths by country 2025*.

<https://worldpopulationreview.com/country-rankings/road-deaths-by-country>