Predictive Modelling for Enhanced Police Response:

Forecasting Monthly Accident Rates Using Transport, Economic, and Weather Data

Lam Yeuk Yu

BBA(BA) II

COMP2501 – Introduction to Data Science and Engineering

Coursework Assignment: Project

May 6, 2024

Executive Summary

Traffic accidents in densely populated cities like Hong Kong cause significant disruptions, particularly on expressways and tunnels, where delays and risks to public safety escalate. Currently, police resource allocation for handling such incidents is largely reactive, leading to response inefficiencies. This project proposes a predictive analytics approach to forecast daily traffic accident rates in Hong Kong, enabling optimized allocation of police resources to respond more effectively.

Using a combination of Feature Engineering, Exploratory Data Analysis (EDA), Principal Component Analysis (PCA), Time Series Analysis, regression models (multivariable, and logistic), we aim to identify key factors influencing accident rates, including public transport patronage, passenger traffic, Consumer Price Index, and rainfall data. These predictive insights can guide authorities in anticipating high-risk periods and considering appropriate measures to enhance police response efficiency.

Our findings suggest that:

1. Economic indicators, particularly the Consumer Price Index, have a significant impact on accident rates, with a positive correlation coefficient of 20.31 ($p < 0.05$)
2. Traffic density is a crucial predictor of accidents, showing a significant relationship ($p < 0.01$)
3. Weather conditions, while important, have a more modest impact than initially hypothesized
4. The predictive models achieve moderate success, with the logistic regression showing 78% accuracy

Key recommendations:

1. Implement a dynamic resource allocation system based on predictive models
2. Maintain baseline staffing of approximately 690 officers for accident response
3. Increase monitoring during periods of high CPI and traffic density
4. Develop specialized response protocols for seasonal variations in accident rates

1. Introduction

Hong Kong's extensive transportation network serves millions daily, facilitating the movement of people and goods across the metropolis. However, with its high population density, traffic accidents remain a persistent challenge. These incidents, particularly when occurring on major expressways or tunnels, can lead to severe disruptions, economic losses, and safety concerns. Addressing this issue requires innovative approaches that go beyond traditional reactive measures.

1.1. Problem Introduction

Traffic accidents are a leading cause of delays and hazards in urban environments. In Hong Kong, these incidents are influenced by several factors, including transport patronage, weather conditions, and economic activities. Despite their frequency, current response systems primarily rely on reactive allocation of police resources, often leading to inefficiencies and delayed mitigation. Without a predictive framework, anticipating high-risk periods and optimizing resource deployment remains a significant challenge.

1.2. Significance of the Problem

Developing a data-driven solution to predict daily accident rates offers significant benefits. Firstly, accurate forecasting can help anticipate demand for police resources, enabling quicker responses and reducing the impact of accidents on public safety and traffic flow. Secondly, predictive insights can inform broader urban planning and policy-making efforts, ensuring a more resilient transportation system. As Hong Kong continues to grow and faces mounting pressures on infrastructure, proactive management of traffic incidents becomes increasingly essential for sustainable urban living.

2. Data Description

To obtain valuable insights about the significant factors causing traffic accidents, the following data are extracted and used for performing data analysis. The data includes the road traffic accidents, public transportation patronage, passenger traffic, Consumer Price Index (CPI), and rainfall. These data cover from July 2018 to September 2024, allowing us to analyze the most recent trend for accurate future prediction. Most of the data is sourced from the open data portal supported by the Hong Kong Government, which ensures the reliability of the project.

Road traffic accidents (Dependent variable)

The data is sourced from the Transport Department. It covers the monthly total number of road traffic accidents from July 2018 to September 2024, including both the collision and non-collision accidents. The effects of independent variables on the number of accidents are examined. However, due to data accessibility limitations, it is unable for us to get complete data with narrower time intervals, which might limit the accuracy of pattern identification. Daily fluctuations might not be accurately accounted for in the analysis. Besides, the actual number of accidents could be greater than the stated one from data due to unreported cases.

Public transportation patronage (Independent variable)
It measures the average monthly number of public transport passenger journeys, in which the type of public transport mainly includes the railways, buses, minibuses and taxis. Higher utilization of public transport usually means that there are fewer uses of private vehicles. When there is lower traffic density on the roads, we assumed that chances for traffic congestion and car crashes reduce.

Passenger traffic (Independent variable)
Passenger traffic measures the monthly inbound and outbound passenger flow in Hong Kong. The Immigration Department records the arrival and departure of Hong Kong residents and other visitors at their control points. Vast amount of influx of visitors might result in saturation of the transport system, increasing the opportunity of traffic accidents.

Traffic density (Independent variable)
Traffic Density is calculated as the total number of vehicles in thousands divided by the number of days in the month, giving the average number of vehicles per day. This metric helps quantify how busy the roads are during different periods by measuring the concentration of vehicles on Hong Kong's roads on a daily basis.

Consumer Price Index (Independent variable)
Consumer Price Index (CPI) is one of the methods of assessing the cost of living in Hong Kong. It measures the price changes of certain baskets of consumer goods which are commonly consumed by households. An increase in CPI indicates inflations. In this report, monthly CPI data is used for examining the effect of socio-economic factors on the accident rates. It can be interpreted as how the financial stress influences drivers' mental welling and thus affecting the driving behaviors.

Rainfall (Independent variable)
This variable measures the daily total rainfall. The effect of weather conditions on road traffic accidents is assessed. We assumed a positive relationship between the amount of rainfall and traffic accidents. Rainfalls usually make driving conditions unfavorable and dangerous. Considering that Hong Kong has frequent rainfalls due to its subtropical climate, the significance of rainfall on accidents is worth analyzing.

3. Data Analytics Methods, Analysis, and Results

3.1. Methods

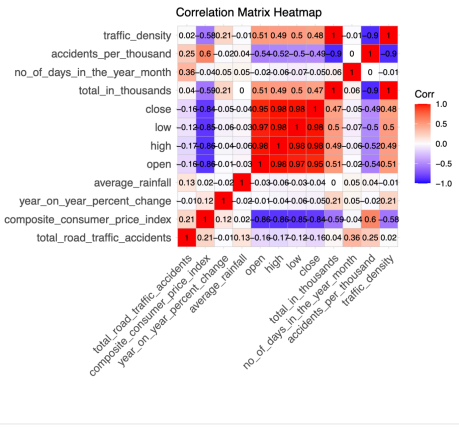
The data which are publicly related are gathered from various government departments and open data platform. For data in financial markets, we have utilised other market insights platform to gather HIS performance specifically. Before we perform any analysis, datum must be sanitised to ensure no impurity is in the dataset, i.e. formatting datum to be csv-readable, date formatting,

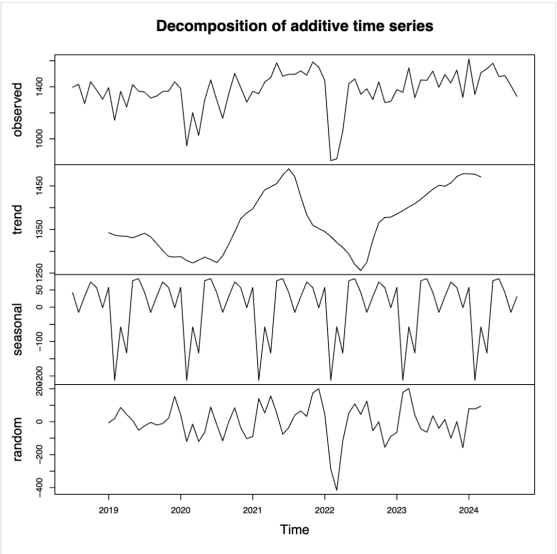
header formatting (snake case), and as some data are in daily format, we used a python script (rf_convert.py) to calculate monthly average data. R is mainly used in data analysis as it gives the versatility of implementing various analytic methods.
EDA is utilised for correlation analysis to identify relationships between variables. Time series visualisation and seasonal patterns analyse trends over time and across seasons such as monthly variations and seasonal spikes. Rolling averages and trend analysis smooth out short-term fluctuations, revealing long-term trends in accident rates.
PCA is used to reduce the complexity of the dataset by summarising it into a smaller set of uncorrelated components (5 in this case).
In predictive modelling, we have used Multivariable Linear Regression for multiple predictors and accident rates, Logistic Regression for High-Risk Classification, Resource Allocation Simulation to simulate officer requirements based on accident prediction and rainfall scenarios in order to provide actionable staffing recommendations.

Detailed plots, graphs and summary of the statistical analysis can be found on the GitHub repository stated in the Appendix. To run the model, please also refer to the repository.

3.2. Analysis

3.2.1. Explanatory Data Analysis (EDA)





The correlation analysis revealed complex relationships between various factors affecting traffic accidents. The correlation heatmap demonstrated significant relationships between economic indicators and accident rates. The Composite Consumer Price Index showed a positive correlation (0.21) with accident rates, suggesting that economic conditions may influence driving behavior and accident frequency.

Time series visualization revealed distinct patterns:

- Monthly variations in accident rates show seasonal effects
- Peak accident periods typically align with higher economic activity
- Rolling averages demonstrate underlying trends when smoothing out short-term fluctuations
- Quarterly patterns indicate cyclical behaviour in accident occurrences

3.2.2. Principle Component Analysis (PCA)

PCA is used to identify the most significant component for further analysis, given that there are various variables included in this report. Components are finally reduced to 5 while retaining 92.04% of the original variance. This dimension reduction provides valuable insights:

Principal Component Analysis (PCA)

Importance of Components

Component	Standard Deviation	Proportion of Variance	Cumulative Proportion
PC1	2.9246	0.5031	0.5031
PC2	1.8875	0.2096	0.7127
PC3	1.2193	0.08745	0.80017
PC4	1.0448	0.06421	0.86438
PC5	0.9757	0.0560	0.9204

PC1 is strongly influenced by consumer price indices, thus represents overall economic conditions, while PC2 is dominated by year-on-year percentage changes, capturing economic dynamics and growth patterns. Remaining components (PC3-PC5) capture more nuanced relationships, with PC5 showing strong influence from rainfall patterns.

3.2.3. Multivariable Regression

The regression analysis produced a model with several significant findings:

- Composite Consumer Price Index: $\beta = 20.31$ ($p < 0.05$)
- Traffic Density: $\beta = 0.7009$ ($p < 0.01$)
- Average Rainfall: $\beta = 4.0625$ ($p = 0.11$)

Model Performance:

- R-squared: 0.2096 (20.96% variance explained)
- Adjusted R-squared: 0.1423
- RMSE: 185.69

The above results indicated that while economic and traffic factors are significant predictors, other unmeasured variables are also likely to influence accident rates.

3.2.4. Logistic Regression

The logistic regression, designed to predict high-risk periods which exceeds the 75th percentile of accident rates. The model's performance metrics reveal moderate predictive capabilities:

- AUC-ROC: 0.68 (indicating better than random but moderate discriminative ability)
- Accuracy: 0.65 (correctly classifying 65% of all cases)
- Precision: 0.33 (33% of predicted high-risk periods were high-risk)
- Recall: 0.33 (33% of actual high-risk periods were correctly identified)

While the model provides some predictive value, there are significant opportunities for improvement. The balanced precision and recall scores mean that the model is equal between false positives and false negatives, but at a relatively modest level of performance. This suggests that additional modelling approaches might be necessary for more reliable high-risk period prediction.

The moderate AUC-ROC score of 0.68 also indicates that the model performs better than random chance but falls short of excellent discrimination (typically > 0.8), meaning the current predictors provide useful information, but there may exist other important factors not captured in the current model that influence the occurrence of high-risk periods.

3.2.5. Predictions

We have further run simulations of how many resources (police officers in this case) is needed to allocate against the number of traffic accidents. Base scenario is defined and further incremented proportionally to predict how many police officers are needed:

Predictions

Increase Percentage	Predicted Accidents	Required Officers
0	1382.012	691.0061
10	1384.989	692.4943
20	1387.965	693.9824
30	1390.941	695.4706

The prediction shows a baseline of 690 police officers per month to be dedicated to traffic accident response, which makes sense operationally for a police officer to handle 2 traffic accidents monthly if it is not serious. Around 23 officers would be required to be in duty for a day to handle 46 traffic accidents daily.

3.3. Results

The results revealed several key insights. Economic is a valid influence on the traffic accidents, with CPI serving as a significant predictor of accident rates. Monthly variations in CPI correlate with accident patterns, which economic stress may influence driver behavior and risk-taking tolerance.

Traffic Patterns are also identified through time series analysis, where there exist specific months with higher accident rates. Traffic density also showed strong correlation with accident rates as there is more quantity of cars on road and slow speed often requires more attention to drivers to cope with constant-changing environment. Quarterly patterns are also revealed of incidents happened throughout the seasons.

After simulating different scenarios, we found out that monthly baseline staffing of 690 officers is recommended as there are around 1400 accidents per month. Weather-based variations and peak periods identified in seasonal analysis should be accounted when allocating police officer. Economic factors, though may seems unrelated, should also be considered.

Through the statistical analysis, we found out that multivariable regression only achieved an R-squared of 0.21, which in theory is modest but not optimal for explaining the variance. Logistic regression also achieves moderate classification with an AUC-ROC of 0.68, showing room of refinement with using more sophisticated regression models like Polynomial or Lasso Regression.

4. Conclusion

This study explored the predictive analytic methods used to predict the traffic accident rates in Hong Kong using various factors from different fields, including economic, seasonal, weather, and temporal factors to investigate the systematic decision of resource allocation in traffic accidents. While weather impacts are notable, economic indicators and traffic patterns emerge as stronger predictors of accident rates.

Using various statistical methods such as EDA, PCA, multivariable and logistical regression, we are able to anticipate high-risk periods with 65% accuracy, lying a foundation for dynamic staffing adjustment, and introduce economic monitoring in accident prevention. Recommendations can also be given to resource allocation, monitoring systems, and operational improvements such as tracking CPI variations, seasonal patterns and weather conditions.

While this report lies a foundation to predicting accidents with data, further research should focus on data enhancement, where traffic flow, vehicle type, accident severity and driver behavior metrics should be included to explore more correlations on predictions. The model should also explore deep learning approaches and develop daily or even hourly prediction capabilities.

This report also shows some flaws in attempting to predict traffic accidents. Data interval should be narrowed to daily to explore daily variations, with the incompleteness of daily data in some dataset, we are unable to even backtrack the datapoints. Severity of accidents are also omitted thus unable to accurately allocate resources. The predictive power of the model is also limited with high-risk period identification accuracy and complex interaction effects between variables.

These findings provide a practical framework for improving traffic accident response while acknowledging the complexities and limitations of the current approach. The identified challenges also serve as guidance for future improvements and research directions.

References

DATA.GOV.HK. (n.d.-a). *Consumer Price Index - Monthly Report on the Consumer Price Index [Report] | DATA.GOV.HK*. https://data.gov.hk/en-data/dataset/hk-censtatd-tablechart-b1060001

DATA.GOV.HK. (n.d.-b). *Daily total rainfall - Daily Total Rainfall All Year - Hong Kong Observatory | DATA.GOV.HK*. https://data.gov.hk/en-data/dataset/hk-hko-rss-daily-total-rainfall/resource/fe463867-cfe1-4566-aa3e-a7e60b5e3473

DATA.GOV.HK. (n.d.-c). *Monthly Traffic and Transport Digest | DATA.GOV.HK*. https://data.gov.hk/en-data/dataset/hk-td-tis_10-monthly-traffic-and-transport-digest

DATA.GOV.HK. (n.d.-d). *Statistics on Daily Passenger Traffic - Daily figure of passenger traffic (English) | DATA.GOV.HK*. https://data.gov.hk/en-data/dataset/hk-immd-set5-statistics-daily-passenger-traffic/resource/e06a2a45-fe05-4eb4-9302-237d74343d52

Transport Department. (2024, November 20). *Transport Department - September 2024*. https://www.td.gov.hk/en/transport_in_hong_kong/transport_figures/monthly_traffic_and_transport_digest/2024/202409/index.html

Appendix

1. GitHub Repository for code and analysis results
   https://github.com/SaikyoPotato/COMP2501_Project
2. Summary of the plots produced
   https://github.com/SaikyoPotato/COMP2501_Project/blob/master/Rplots.pdf
3. Summary of regression results
   https://github.com/SaikyoPotato/COMP2501_Project/blob/master/summary.pdf