

Figure 2: Discrimination Measure plot over variables in accidents data.

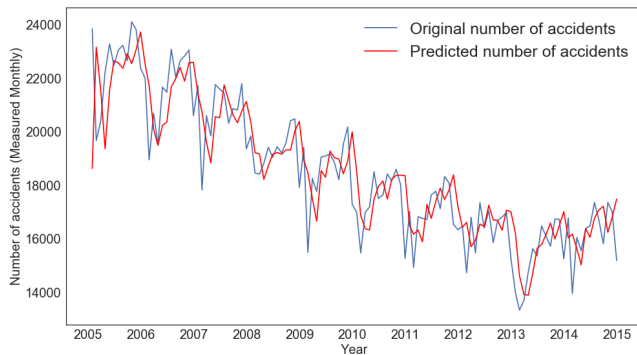


Figure 3: Prediction of monthly accidents using autoregression from 2005 till 2014.

cannot be accurately represented using MCA, we use hypothesis testing and time series analysis to analyze them.

#### 4.2 Multiple Correspondence Analysis variables plot.

We choose three variables which had high variance along Dimensions 1 and 2 from Discrimination Measure Analysis, namely **Location of the accident (Postcode), Day of the week and the Age group of the driver** and project them using MCA with all the features in a single plot to study the correlation. This plot shows how each of these categories is related to others in the form of a scatter plot. Several insights obtained from this analysis are discussed in the results section.

#### 4.3 Analysis over other variables and important events in the UK.

Not all the variables can be efficiently represented in lower dimensions using MCA, hence further techniques to analyze data are required. Hypothesis testing can be used to further understand how each of the features of the variables is related. We used Welch's t-test statistic to study several hypothesis on this dataset. Furthermore, because this dataset is time-bound, we can make some predictions on the data using time series analysis. We applied autoregression on our dataset to analyze and predict the trend in accidents over the years. Results are discussed in the next section.

### 5 RESULTS

#### 5.1 MCA plot with postcode, day of the week and age group of driver (Figure 1)

- The number of accidents on Sundays and Wednesdays is fewer than those on other days in any postcode.

Null Hypothesis	Result
Number of daily accidents in summer and winter are equal	15 to 30 more daily accidents in summer
Number of daily accidents by young drivers (Age 18-25 years) and old drivers (Age 65-85 years) are equal	85 to 89 more accidents by young people
The number of daily accidents before and during the London Summer Olympics (2012) were same	Accept Null Hypothesis. P-value 0.197.
Number of daily accidents in areas close to subway stations is same as other areas	9 to 29 more accidents daily in areas close to subway stations.
Males cause an equal number of daily accidents as females	428 to 439 more accidents by males.

Table 1: Results from hypothesis testing.

- Age groups 11-15 years, 26-35 years and 36-45 years have the similar number of accident records and the major day of accidents for these age groups is Saturday.
- Warrington(WA) and Guildford(GU) have more accidents on Tuesdays and the most common age group of people causing accidents is 46 to 55 years.
- Age group 6-10 years is responsible for a lesser number of accidents compared to other age groups.

#### 5.2 Hypothesis testing

We found out that the number of accidents before and during the London Summer Olympics remained same. Similarly, other interesting hypothesis were tested which are discussed in Table 1.

#### 5.3 Time Series Analysis

Figure 3 shows the prediction of the number of monthly accidents over the years. We see that the number of accidents has decreased over the years. The prediction accuracy can be measured by the root mean square error value, which was 699.84.

### 6 CONCLUSION

In this paper, we combined visualization and data analysis techniques for the effective study of a dataset. We visualized the correlation between the location of the accident, day of the week and age of the driver using MCA. Further, we studied other important features using hypothesis testing and predicted the trend in accidents using time series analysis. Future work will include more detailed analysis of the data using Machine Learning and other advanced visualization techniques.

### 7 ACKNOWLEDGMENTS

This research was partially supported by NSF grant IIS 1527200 & MSIT, Korea under the ICTCC Program (IITP-2017-R0346-16-1007).

### REFERENCES

- [1] <https://www.kaggle.com/silicon99/dft-accident-data/data>.
- [2] H. Abdi and D. Valentin. Multiple correspondence analysis. *Encyclopedia of measurement and statistics*, pp. 651–657, 2007.
- [3] P. Ljubič, L. Todorovski, N. Lavrač, and J. C. Bullas. Time-series analysis of uk traffic accident data. In *Proceedings of the Fifth International Multi-conference Information Society*, pp. 131–134, 2002.
- [4] P. Sikdar, A. Rabbani, N. Dhapekar, and D. G. Bhatt. Hypothesis testing of road traffic accident data in india. *International Journal of Civil Engineering and Technology (IJCIET)*, 8(6):430–435, 2017.