Heterogeneous Prediction of Urban Mobility Patterns in Cluj-Napoca Using Lasso and Random Forest Techniques

Amory Yagar
Catherine Hayden
Makai Kost
Noah Meyer
Samuel Tragesser

California Polytechnic State University

June 10, 2022
Spring Quarter

Abstract

Urban sprawl in Romanian cities is a pervasive challenge in Romanian urban development that often finds itself realized through auto traffic congestion. There is motivation to study this issue and find potential solutions to combat pervasive traffic. Through the use of WAZE traffic alert data, we apply random forest and lasso machine learning models to predict the frequency of alerts in localized regions of Cluj-Napoca. We include both spatial and temporal covariates including school and hospital proximity, weather patterns, COVID lockdowns, and sporting events. Our model finds that random forest acts as a more successful predictor of localized traffic alerts. Spatial indicators appear to be better predictors of alerts in more densely populated regions, while temporal features are better predictors outside of the city center.

I - Introduction

Many cities in Romania face challenging traffic dilemmas as their sprawl outpaces the placement of key infrastructure. The imbalance drives residents of the cities' outskirts to more developed areas in the city centers and adds immense pressure to the traffic systems. There is an operational need for the World Bank Romania to understand where traffic bottlenecks are occurring in order to determine the placement of additional key infrastructure.

The World Bank has published many annual reports of traffic impacts and urban planning studies. However, none of these reports have ever utilized data collected from WAZE. The specific question this paper seeks to answer is: what are factors that predict traffic alerts in a given spatial area? In answering this question, this paper utilizes WAZE traffic data collected by the World Bank Romania and Geospatial Operations Support Team (GOST). This project is heavily concentrated on big data cleaning and aggregation using Python. This paper will also introduce a small variety of geospatial analytics methods such as hexagonal hierarchical geospatial indexing system (H3) and nearest-neighbor spatial joining.

By utilizing machine learning techniques, the final model outlined in this paper provides the World Bank Romania with a flurry of public policy implications that aid in urban planning and development in the city of Cluj-Napoca. The analysis performed within this project has been made replicable for other cities and integrated with other standard analytic processes.

This paper additionally seeks to visually display major commuter and transit patterns around high travel times with a focus on 2020 and 2021 data. The following sections will go over former estimations of traffic, and examine the data used in our analysis and issues we found when using this data. Finally this paper will introduce the best prediction model selected and explain conclusions found.

II - Literature Review

In order to understand the impacts and determinants of traffic alerts using WAZE traffic data, we require a comprehensive look at multiple facets of this problem. These issues include the mobility trends in Romania, and Cluj-Napoca specifically; previous studies of the impact of traffic; and past uses of WAZE traffic data.

In the last two decades, Romania has experienced dramatic shifts in mobility practices following immigration and increase to private industry building housing. This rise in car mobility has therefore resulted in increased levels of traffic and road collisions. The city of Cluj-Napoca specifically has faced major concerns with regards to traffic and urban mobility (Sislen et al. 2021). Specifically, a number of residents have migrated out of the city center towards the suburbs, yet continue to work in the city center. This has caused a severe number of bottlenecks entering the city center. Despite these demographic shifts, the local government has been relatively slow to adapt to these changes, leaving a transportation system of roadways unfit for the geometry of the city.

Comparative studies of traffic crashes in the Romanian city of Cluj-Napoca have been conducted to identify relative hotspots of traffic crashes and used the prominence of social causes to explain traffic crash occurrences (Benedek et al. 2016). The data from Benedek's existing work was collected by the Traffic Department of the General Inspectorate of Romanian Police from the years 2010 through 2013. The results were calculated using Kernel Density Estimation to create a hotspot map of traffic crashes in Cluj-Napoca. The study separates the city into four distinct districts in order to parse out any differences in traffic accidents that occur in different areas of the city. This method proved to be a useful tool in that the downtown areas showed the greatest frequency of accidents, where each side of the city had its own distinct characteristics.

The main causal factors of traffic accidents include weather visibility, road characteristics, vehicle quality, human error, and the overall volume of vehicles on the road (Retallack and Ostendorf, 2019). The likelihood an accident may occur increases by the total cars on the road combined with the listed characteristics. It is also likely that different traffic patterns will occur depending on the day of the week and special holidays. In a study of the impacts of tourism on traffic congestion in tourist friendly regions of Spain, Saenz-de-Miera & Rosselló (2012) control for days of the week and holidays a regression model of aggregate traffic volume. They find that these variables do play a significant role in estimating traffic congestion. Similarly, past work looks at the impact of large city events on traffic patterns. These non-recurring events can prove to be significant traffic predictors, as they should increase alert frequency greater than expected patterns. A neural network model of American football games predicting traffic was found to be over 98% effective in prediction (Sun et al, 2017). Transferring this prior work to our predictive model in Cluj-Napoca may not display the same sign or significance, but they should prove to be useful variables in prediction given past results.

The enactment of lockdowns faced by COVID-19 resulted in fewer people in public places and thus fewer people on the roads and highways. In a 2020 summary analysis, the European Transport Safety Council published in a Road Safety Performance Index briefing that Romanian traffic restrictions caused by the COVID-19 pandemic resulted in a 50% reduction in serious road traffic collisions and a reduction of 39% in the number of road deaths. During the period between March 16-April 26, the number of traffic accidents per day in the Spanish Tarragona province fell by approximately 76% (Saladié et al, 2020) when compared to the same

84-day period in 2018 and 2019. The reduction in traffic accidents was measured through figures such as average accidents per day, whether accidents occurred on a weekend/holiday or weekday, and a distribution of the concentration of the accidents using a Kernel Density Estimation function.

In the United Arab Emirates, the hospitalization rate of road traffic collision trauma patients decreased by 33.5% during the height of the COVID-19 pandemic in comparison to the pre-pandemic period. Using data retrieved from the Abu Dhabi Trauma Registry, a comparative analysis was performed between two cohorts of patients seen a year before the pandemic and a year after the first case was reported in the United Arab Emirates in January 2020. The analysis attributed the decrease in road traffic collision patients to the reduction in motor vehicle, pedestrian and bicycle injuries seen during this time period. Our work follows a similar pattern using pre-pandemic data as a baseline for comparison.

Our analysis of Romanian traffic will be observed using data collected through WAZE. The use of WAZE data for analytics is a relatively new concept as the company started their data sharing program with governments in 2014. The identification of traffic crash hot spots is an important parameter for improving road safety. The previous standard for identifying these spatial locations in safety studies was through police crash reports. Li et al (2020) found that WAZE reports are better at predicting crash areas than police crash reports. This is due to reports with low property damage being underreported.

Despite the successes of WAZE data analytics, the use of these crowdsourced datasets have drawbacks as well. WAZE produces self reported data which means there is going to be a variance in the time people report accidents as well as where that person is located when they make a report. The implementation of various classification and regression models have been used as a workaround for this reporting issue. Flynn et al. (2018) used ridge, lasso, and random forests on WAZE data from Maryland to estimate the number of traffic counts for a given day or period.

Urban traffic is greatly impacted by the presence of municipal transit and more specifically the location of curb-side bus stops. In a study of bus route impact on traffic bottlenecks, Jin, Hui, et al. (2019) found that there is sporadic lane blockage at curb-side bus stops which creates aggressive lane changes and traffic delays. The model proposed within that paper is an extended link-node structured two-lane cell transmission that is capable of dynamically capturing the effect of lane blockage at bus stops and modeling mandatory and discretionary lane changes along a bus route under various traffic demand levels. In support of our analysis, this previous work solidifies the understanding that the specific positional coordinates of bus stops along urban streets can work to intensify traffic congestion.

Along with bus stops, other spatial points of interest have been used in previous studies of traffic congestion. Zhang and Xu (2017) analyze the impacts of traffic of proximity to schools and find that a school facility not only traffic with its direct jurisdiction, but also has spillover effects in the surrounding areas. Other work agrees with this finding as there is an estimated 4.5% increase in traffic during school rush hour times (Sun et al, 2021). Community facilities within a city are important predictors and relevant variables that impact traffic congestion due to the fact they are fixed in space and are areas where trasitters do not have much choice in whether they commute to this region of a city.

III - Data

The data used for this analysis came from WAZE and the World Bank. The WAZE data used within this project has been collected by the World Bank Romania through *The WAZE for Cities Data* program. The *WAZE for Cities Data* program is a free two-way data exchange partnership that allows government entities to make smarter decisions in urban planning. This program was launched in October 2014 and has since expanded to include over 1500 partnerships between cities, states, and country governments.

WAZE data includes anonymous user-reported traffic alerts which include road events labeled as accidents, jams, road closures, and weather hazards from 2019 through 2022 (called uuid). Shown by figure 3.1. The WAZE data evolves with every driver and data point added. The original collection of our raw data was scraped from a server every 5 minutes. The raw dataset is at the individual alert level with columns indicating the coordinates and timestamp of each alert.

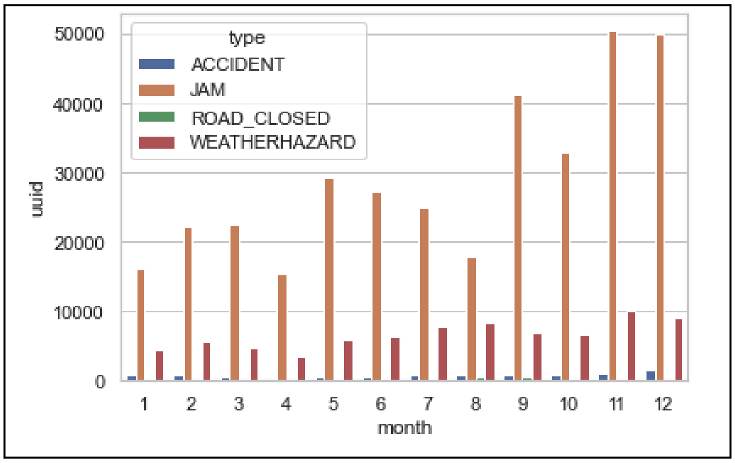


Figure 3.1 Types of WAZE data Alerts

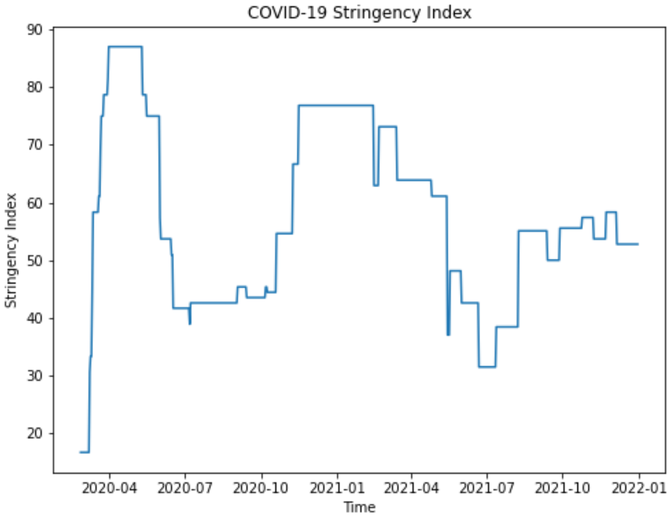


Figure 3.2 Stringency Index

Also contained in the data is a variable that represents the effects of the COVID-19 lockdown mandates. This variable was created from a COVID-19 Stringency Index given by the University of Oxford in *Our World in Data* for the country of Romania. The Stringency Index is a measure based on 9 response indicators including school closure, workplace closure and travel bans. The variable has been rescaled from 0 to 100, where 100 indicates very strict guidelines and maximum closure. Stringency Index can be updated daily and is shown by figure 3.2.

To break down alert frequency into a spatial unit, we utilized Uber's H3 hexagonal index to organize the number of alerts into a hexagon shaped spatial area. These hexagons are fixed in space, but can be adjusted in size. Each observation of our alert data was reformatted into the number of alerts per hexagon per day. An individual alert conveys little meaning, but when they are grouped together within this hexagon shaped spatial unit, we can form a better understanding of alert frequency through space and time. Our observational units become 0.15 km^2 hexagons for a given day. Our final dataset is aggregated from 441,456 individual alerts to 3560 hexagons within the city of Cluj-Napoca to form 199, 283 hexagon-day units of observation.We will refer to these as "child hexagons" from this point on.

In order to account for traffic buildup due to inclement weather, we included in the analysis a measure of precipitation per day given in 0.10mm. This variable was taken from the free source *European Climate Assessment*. To integrate the effects of sporting events, we scraped soccer game data from ESPN and created an indicator for whether there was a soccer game on each specific day . We also web scraped Romanian holiday data to generate an indicator variable for whether a specific day was a holiday. Using the raw data we constructed variables that proved counts of hospitals, schools and bus stops in each child hexagon. These variables are fixed over time but fluctuate by hexagon.

After gathering the initial set of covariates we then added interaction and polynomial terms up to order three to try and capture non-linearity in the data. We only used higher order terms if the covariate was numerical. For example, the measure of precipitation we took higher orders but for day of the week we did not since that is a categorical variable. After computing these interactions we had over 1,300 variables.

IV - Data Limitations

One of the main shortcomings faced in the WAZE data is the lack of observations prior to 2020. Figure 3.3. below shows that there is a large amount of data missing from 2019 and given that some of our data outside of WAZE ended on January 1, 2022 we had to leave out WAZE data from 2022. The WAZE data we received required copious amounts of cleaning and wrangling. We were also required to web scrape data from numerous free access sources in order to create relevant covariates to support our analysis.

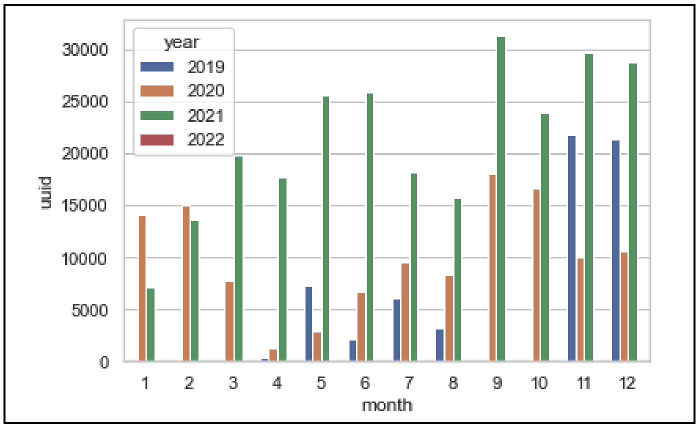


Figure 3.3. Alerts by Month

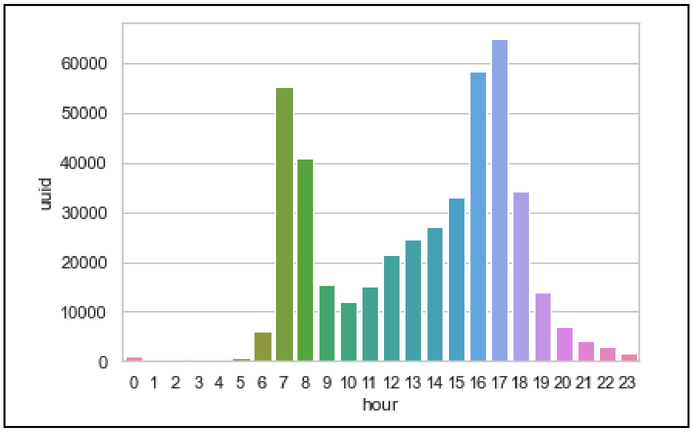


Figure 3.4. Alerts by Hour

Another limitation of our dataset is the lack of availability of temporal covariates subset by hour. Our model was limited to a daily analysis due to the accessibility of complementary data. As shown by Figure 3.4., there are clear trends in alert quantity in morning and afternoon rush hours but we are unable to specify a robust model of hourly traffic predictions.

Data provided by Waze App. Learn more at Waze.com .

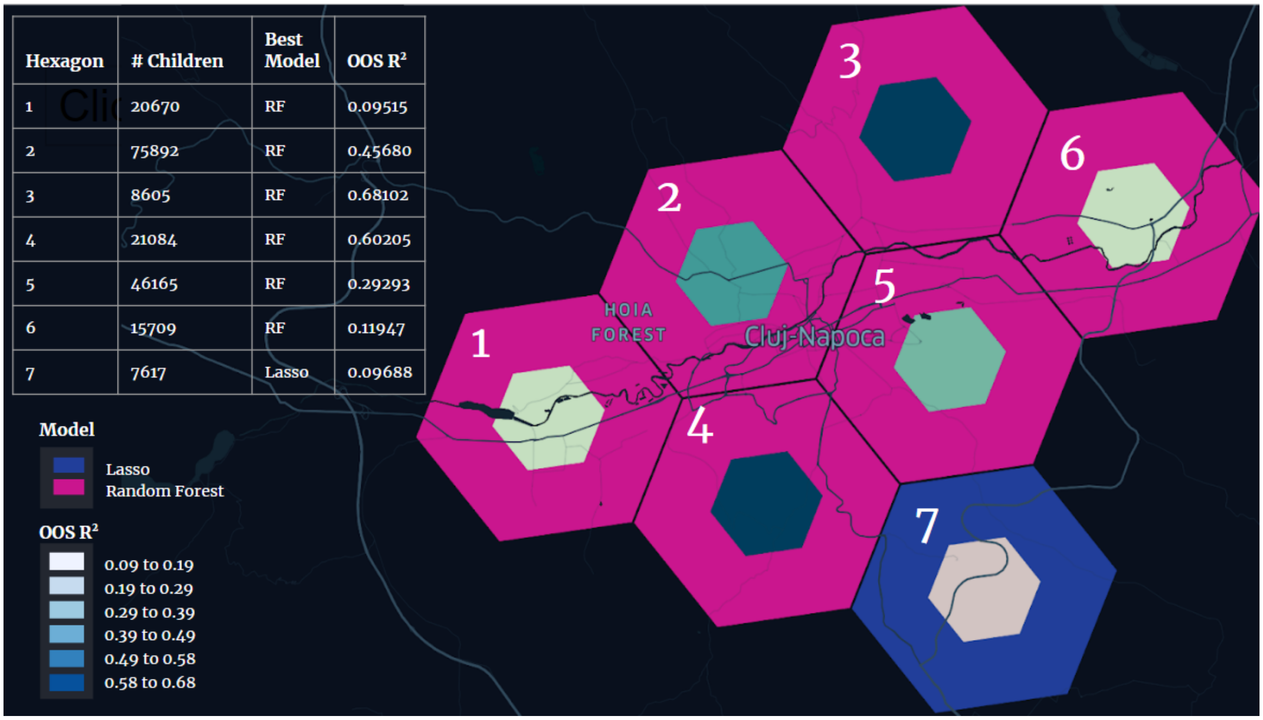
V - The Model

The primary objective of this project is to create a predictive model of traffic alerts given an array of covariates. Our work looks to summarize traffic patterns in a way that will inform city planners and promote infrastructure development to support the expanding urban sprawl. As previously mentioned, the covariates included in our model include, proximity to schools and hospitals, weather, home soccer games, holidays, quantity of bus stops and COVID-19 lockdown stringency.

To account for the various differences across spatial areas within a city, our predictive model segments the city into 7 large H6 *parent* hexagons of size 36.13 km^2. Within each of these parent hexagons, we fit at least 2000 smaller H10 0.15 km^2 *child* hexagons within each parent hexagon. There are differences across the spatial composition of a city, thus estimating separate models within each parent provides us with a more accurate predictive model within a specific district within the city. .

Within each parent hexagon, we run both a lasso and random forest model with daily child hexagon alert counts as the outcome variable. Due to some limitations on our computer's ability to handle the computational intensity of these models, we were unable to do cross-validation for the random forest models. In order to remedy this issue, we based our model selection on the out of sample R-squared. We filtered out any parent hexagons that did not contain at least 2000 observations within them in order to reduce bias induced by a small sample size. This left us with seven parent hexagons that met the criteria. The rest of this analysis will be focused on these seven hexagons.

VI - Empirical Results



Hexagon	# Children	Best Model	OOS R^2
1	20670	RF	0.09515
2	75892	RF	0.45680
3	8605	RF	0.68102
4	21084	RF	0.60205
5	46165	RF	0.29293
6	15709	RF	0.11947
7	7617	Lasso	0.09688

Figure 3.5. Results of Best Fit Models

Of the seven parent hexagons that we tested random forest and lasso models on, six of them had a best out of sample R-squared value from the random forest. The parent hexagon in which lasso generated the most accurate predictions was also the one with the least amount of observations. A preliminary analysis of the model features broken out by hexagon did not provide any information on stark differences that may have led to the anomalous model selection. Our out of sample R-squared values ranged from 0.09 to 0.68. On average, there was a better fit of predictions for parent hexagons near the city center with many observations. Parent hexagon 3 is anomalous in this regard, as it has the highest out of sample R-squared while having the second least number of observations.

Within each parent hexagon we calculated the feature importance. For lasso this is straightforward as the largest coefficients had the most importance. For the random forest we used the Gini Index to get the feature importances. While we have heterogeneous predictions there was some commonality in which features were most important among the models. The stringency index, day of the week, number of hospitals, and number of schools and their polynomial and interaction terms were valuable in our predictions.

.

VII - Project Conclusions

The motivation of this study was generated by the need of the World Bank Romania to understand where traffic bottlenecks are occurring within the city of Cluj-Napoca. The results of our model aim to drive Romanian public policy and the placement of key infrastructure. We found that random forest is generally preferred to lasso for predicting traffic patterns.

It is difficult to draw definitive policy conclusions due to our use of a predictive model, the data timeframe encompassing the majority of the Covid-19 pandemic. Unsurprisingly this meant the covid stringency index was one of the most important predictors. Scaling this model up or implementing it in other areas in the future, hopefully the index will be an irrelevant predictor. This draws major concerns then on the validity of the model and its true predictive abilities. Going forward a more robust set of covariates that can explain the whole city would make this a better model. Our model had decent out of sample r squared values for some of the parent hexagons that contained the center of the city but quite poor otherwise. This is where a researcher that had a more intimate knowledge of the region and its system of infrastructure planning would be better able to formulate a framework for a predictive analysis.

In addition, our model is currently constructed to the predetermined hexagons that h3 has. Where the hexagons fall is arbitrary coming from the researcher's side which has the consequence of potentially partialling out the city into hexagons that don’t make sense. For example, a hexagon could be splitting up a residential neighborhood. In the future we would like to be able to make a model that allows us to move hexagons around as we choose. A benefit of that approach is we could do some sensitivity analysis and see how results would change if we were to shift the hexagons around slightly.

As our model currently stands, we believe the best way to use it is to do static analysis where a policy maker could see what the prediction for traffic could be for a given set of covariates. For instance what would be the prediction for building a new school in a specific area.

Overall, we hope that the World Bank and any other parties interested in understanding Romanian traffic patterns will be able to use our framework and code to help inform future decisions surrounding infrastructure development.

VIII - References

Benedek, József, Silviu Marian Ciobanu, and Titus Cristian Man. “Hotspots and Social Background of Urban Traffic Crashes: A Case Study in Cluj-Napoca (Romania).” *Accident Analysis & Prevention* 87 (2016): 117–26. https://doi.org/10.1016/j.aap.2015.11.026.

Flynn, Dan F.B., Michelle M. Gilmore, J. Patrick Dolan, Paul Teicher, and Erika A. Sudderth. “Estimating Traffic Crash Counts Using Crowdsourced Data: Pilot Analysis of 2017 Waze data and Police Accident Reports in Maryland.” *Transportation Research Record: Journal of the Transportation Research Board*, November 1, 2018. https://doi.org/10.1177/03611981221083305.

Jin, Hui, et al. “Impact of Curbside Bus Stop Locations on Mixed Traffic Dynamics: A Bus Route Perspective.” *Transportmetrica A: Transport Science* (2019): 15- 2. https://doi.org/10.1080/23249935.2019.1601789.

Li, Xiao, Bahar Dadashova, Siyu Yu, and Zhe Zhang. “Rethinking Highway Safety Analysis by Leveraging Crowdsourced Waze Data.” *Sustainability* 12, no. 23 (December 4, 2020): 10127. https://doi.org/10.3390/su122310127.

Retallack A.E., Ostendorf B. Current understanding of the effects of congestion on traffic accidents. *Int. J. Environ. Res. Public Health.* 2019;16(18):3400. doi: 10.3390/ijerph16183400.

Saenz-de-Miera, Oscar, and Jaume Rosselló. "The responsibility of tourism in traffic congestion and hyper-congestion: A case study from Mallorca, Spain." *Tourism Management* 33, no. 2 (2012): 466-479.

Saladié, Òscar, Edgar Bustamante, and Aaron Gutiérrez. “Covid-19 Lockdown and Reduction of Traffic Accidents in Tarragona Province, Spain.” *Transportation Research Interdisciplinary Perspectives* 8 (November 8, 2020): 100218. https://doi.org/10.1016/j.trip.2020.100218.

Sislen,David N.; Proskuryakova,Tatiana A.; Benedek,Jozsef; Moldovan,Sandu Ciprian; Cristea,Marius; Varvari,Stefana Alexandra Diana; Ionescu-Heroiu,Marcel; Zotic,Vasile; Alexandru,Diana Elena; Man,Titus-Cristian; Dolean,Bogdan Eugen; Harangus,Iulia. *Substation Study on Transport and Communications : Cluj County Spatial Plan (English).* Washington, D.C. : World Bank Group. http://documents.worldbank.org/curated/en/615971624852968232/Substation-Study-on-Transport-and-Communications-Cluj-County-Spatial-Plan

Sun, Fangzhou, Abhishek Dubey, and Jules White. "DxNAT—Deep neural networks for explaining non-recurring traffic congestion." In *2017 IEEE International Conference on Big Data (Big Data)*, pp. 2141-2150. IEEE, 2017.

Sun, Weizeng, Dongmei Guo, Qiang Li, and Haidong Fang. "School runs and urban traffic congestion: Evidence from China." *Regional Science and Urban Economics* 86 (2021): 103606.

The Road Safety Performance Index (PIN). “PIN BRIEFING: The Impact of Covid-19 Lockdowns on Road Deaths in April 2020.” Brussels: European Transport Safety Council, July 2020.

Yasin, Yasin J., Michal Grivna, and Fikri M. Abu-Zidan. “Global Impact of Covid-19 Pandemic on Road Traffic Collisions.” *World Journal of Emergency Surgery* 16, no. 1 (November 19, 2021). https://doi.org/10.1186/s13017-021-00395-8.

Zhang, Qianqia, and Huile Xu. "Data driven analytics for school location impacting urban traffic congestion." In *2017 4th International Conference on Transportation Information and Safety (ICTIS)*, pp. 883-889. IEEE, 2017.

IX - Acknowledgements

We would like to express our gratitude and appreciation to the following individuals for their contributions and assistance throughout this project:

Carlos Arturo Flores
Clara Ivanescu
Benjamin P. Stewart
Carli Bunding-Venter
Gabriel Stefanini Vicente
Kurt W. Colvin