Complementary Data Sources for Road Transport Statistics

Use of Machine Learning in Providing Additional Insights into Road Crashes

UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity **Dignified Life**

UNITED NATIONS
الاسكوا
ESCWA

Shared Prosperity **Dignified Life**

VISION

ESCWA, an innovative catalyst for a stable, just and flourishing Arab region

MISSION

Committed to the 2030 Agenda, ESCWA's passionate team produces innovative knowledge, fosters regional consensus and delivers transformational policy advice. Together, we work for a sustainable future for all.

E/ESCWA/CL4.SIT/2020/TP.16

Economic and Social Commission for Western Asia

Complementary Data Sources for Road Transport Statistics

Use of Machine Learning in Providing Additional Insights into Road Crashes

United Nations
Beirut

The findings, interpretations and conclusions expressed in this publication are those of the authors and do not necessarily reflect the views of the United Nations or its officials or Member States.

The designations employed and the presentation of material in this publication do not imply the expression of any opinion whatsoever on the part of the United Nations concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Links contained in this publication are provided for the convenience of the reader and are correct at the time of issue. The United Nations takes no responsibility for the continued accuracy of that information or for the content of any external website.

References have, wherever possible, been verified.

Mention of commercial names and products does not imply the endorsement of the United Nations.

References to dollars ($) are to United States dollars, unless otherwise stated.

Symbols of United Nations documents are composed of capital letters combined with figures. Mention of such a symbol indicates a reference to a United Nations document.

United Nations publication issued by ESCWA, United Nations House,
Riad El Solh Square, P.O. Box: 11-8575, Beirut, Lebanon.

Website: www.unescwa.org

Acknowledgements

The present technical paper was initiated based on a recent transport data questionnaire that was circulated by the Economic and Social Commission for Western Asia (ESCWA) in April 2019 among national statistical offices of the Arab countries. ESCWA received responses from 14 Arab countries, including data relating to road safety. The initial idea was to check whether data provided by the national statistical offices can be combined with other sources of data, including big data, to come up with a better understanding of causes of car crashes and to showcase to member countries readily available solutions at affordable means to improve their data collection and analysis of such data. Several meetings and consultations were held with consultants, big data providers, government research and other government entities and organizations working in this field, as well as colleagues at ESCWA from the Technology Development Division and the Economic Development and Integration Division, and other colleagues from United Nations specialized agencies and entities. They all shared many insights on the tools and the means available. Our thanks and appreciation go to all of them for shedding light on and valuable insights into which technologies are available and which data they provide, in addition to potential cost estimates and means necessary for carrying out more sophisticated and advanced data collection and analysis on road safety. Another target was to employ data analytics in the analysis of the data provided.

The development of the paper was carried out by Mohamad Hossary, consultant, under the continuous direction, guidance, supervision, substantive review, and contribution of Wafa Aboul Hosn, Chief, Economic Statistics Section, and Naela Haddad, Statistician, Economic Statistics Section, who also provided the review and substantive edits. The paper also benefited from insights, discussions, knowledge sharing, and review from Yarob Badr, ESCWA regional advisor on transport and logistics; Fouad Mrad, Technology Development Division; Alexander Blackburn, Transport Division at the United Nations Economic Commission of Europe (UNECE); Thierry Coulet, a key expert at the regional MEDSTAT programme, a European Union cooperation programme with Mediterranean countries in the context of the European Neighbourhood Policy (ENP) aimed at assisting partner countries in the development of their official statistics; Ali Gandour, Lebanese National Centre of Scientific Research; Major Michel Moutran, senior expert, National Road Safety Council, Lebanon; Ziad Abdallah, Central Administration of Statistics, Lebanon; and others. Furthermore, the paper benefitted from the presentations, discussions and recommendations made at the Regional Workshop on the Development of Transport Statistics and Transport-related Sustainable Development Indicators in the Arab Countries, which was held in Beirut in September 2019, where valuable insights were brought to the forefront by speakers and participants. The workshop’s presentations and documentation can be viewed at https://www.unescwa.org/events/regional-workshop-development-transport-statistics-and-transport-related-sustainable.

Executive summary

Road safety is a key issue that is salient in all communities worldwide and the Arab region in particular. As a result, the Statistics Division at the United Nations Economic and Social Commission for Western Asia (ESCWA) conducted a pilot project on the use of complementary data sources on car crashes. Data sources from both the private and the public sector were investigated and several key challenges were faced. These challenges include siloed data, lack of regulatory frameworks for data sharing, lack of transparency between governmental agencies, and general sensitivities towards sharing data. As a result, open data from the United Kingdom on car crashes were obtained, and street data from OpenStreetMap (OSM) were added as a complementary data source. Feature engineering was done on OSM data to extract road curvature, and the results were inputted into decision trees, gradient boosted trees and random forest to predict a crash injury severity. The results show that there seems to be a relationship; however, further work is needed to achieve more reliable results and to bring the experimental nature of machine learning more in line with official statistics. The exercise highlighted the possible benefits of using machine learning algorithms to understand car crashes. These benefits include the abilities to see the logic of the prediction from the decision tree and to see which features the models consider important.

Contents

List of tables

List of figures

Abbreviations and acronyms

API	application programming interface
CART	Classification and Regression Tree
CCTV	closed-circuit television
CNN	convolutional neural network
ESCWA	Economic and Social Commission for Western Asia
GPS	global positioning system
GPU	graphics processing unit
GSM	global system for mobile communication
ISF	Internal Security Forces (Lebanon)
NASA	National Aeronautics and Space Administration
NOAA	National Oceanic and Atmospheric Administration
NSO	national statistical office
OSM	OpenStreetMap
QCRI	Qatar Computing Research Institute
ROC	receiver operating characteristic
SDG	Sustainable Development Goal
WGS	World Geodetic System
WHO	World Health Organization

Introduction

Road safety is a key issue that is salient in all communities worldwide and the Arab region in particular. Research has shown that there is a disproportionate rate of death to the population and the number of vehicles of low- and middle-income countries compared to high-income countries. The risk of road crash death in low-income countries is more than three times higher than in high-income countries, with an average rate of 27.5 deaths per 100,000 population in the former, compared to 8.3 deaths per 100,000 population in the latter.[1]

Regardless of the income level, road safety remains a critical and important issue. Road traffic injuries are one of the leading causes of death of young adults and children.[2] As a result, road safety has garnered international focus to the extent that the United Nations included road safety in the targeted goals of the United Nations 2030 Agenda for Sustainable Development and has specifically assigned some targets addressing this issue. Target 3.6 of the Sustainable Development Goals (SDGs) aims to reduce the number of global deaths and injuries from road traffic accidents by half by the year 2020, and target 11.2 focuses on improving road safety through public transport.[3] The United Nations General Assembly also issued a resolution reaffirming the importance of road safety and urging member States to implement road safety policies that protect vulnerable individuals.[4]

Noting the magnitude of road deaths and injuries in the Arab region, the purpose of this paper is to highlight some of the potential data sources related to road safety, and to present a pilot project conducted by the Statistics Division at the United Nations Economic and Social Commission for Western Asia (ESCWA) on the use of complementary data sources on car crashes. Several studies dive into the detailed use of one or two complementary data sources with a great deal of focus given to the technical aspects. In this paper, a more pragmatic approach will be taken to showcase several complementary data sources used, with comments on their availability, and related costs to be incurred for obtaining them. Moreover, the various methodologies for data analysis will be proposed, highlighting the use of machine learning in providing additional insights into car crashes.

[1] World Health Organization, 2018. Global Status Report on Road Safety.

[2] Ibid.

[3] United Nations Economic and Social Council (ECOSOC), 2019. Special Edition: Progress towards the Sustainable Development Goals. Report of the Secretary-General. Advanced unedited version. New York: United Nations.

[4] United Nations General Assembly, 2020. Improving global road safety. Resolution adopted by the General Assembly on 31 August 2020 (A/RES/74/299). Available at https://undocs.org/en/A/RES/74/299 (accessed on November 2020).

I. Literature Review

Addressing road safety and achieving the stated SDG goals and targets requires greater insight into the issue at hand. To gain this insight, data is needed along with adequate analytics tools and methodologies. One key source are the official data collected by governments. These data are typically sourced from police records that are, in turn, gathered from accident sites.[5] These data consist of officer observations, information gathered from those involved in the crash, and witness statements. Table 1 is an example of the type of information that can be gathered from police records.[6] National statistical offices (NSOs) typically rely on this data for their analysis and reports. Given the importance of official car crash data, several databases exist on the subnational, national and regional levels to store and disseminate official car crash data. Some examples include the French national road traffic accident database, known as BAAC, the United States' Fatality Analysis Reporting System, and the European Union's Community database on Accidents on the Roads in Europe.

Table 1. Dubai crash metadata

Attribute name	Attribute description
psn_id	
record_status	
acd_date	Accident date
acd_time	Accident time
acc_location	The statement describes the location of the accident
Id	System-generated ID for the accident
acc_type	Accident category (minor, major, etc.)
acc_cause	Cause of the accident
Weather	Description of the weather
road_status	Description of the road status
Age	Main actor age

[5] International Transport Forum, 2011. Reporting on serious road traffic casualties: Combining and using different data sources to improve understanding of non-fatal road traffic crashes. Paris. Available at https://www.itf-oecd.org/sites/default/files/docs/road-casualties-web.pdf; and Directorate General Transport and Energy of the European Commission, 2007. Best Practices in Road Safety: Handbook for Measures at the European Level. Available at https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/projects_sources/supreme_d_handbook_for_measures_at_the_european_level.pdf.

[6] Dubai Pulse, 2019. Dubai Traffic Accidents. Available at https://www.dubaipulse.gov.ae/.

Attribute name	Attribute description
Gender	Main actor gender
injury_severities	Driver injury level
driving_license_issue_date	Driving license issue date
Occupation	Driver occupation
Intoxication	Intoxication
seat_belt_status	Flag
year_manufactured	Year manufactured
insurance_company_name	Insurance company name

Source: https://www.dubaipulse.gov.ae/.

There are, however, certain issues facing the collection of official data. One critical issue is that not all accidents are reported to the police.[7] For example, in rural areas, accidents can occur without the knowledge of the police. Additionally, crashes that have material damage only without injuries or deaths are not systematically recorded by the police. That means that car crashes tend to be understated in official datasets. Moreover, fatalities can occur at the hospital days after a car accident, and not all countries account for the 30-day follow-up rule set by the World Health Organization (WHO), which is reflected in the underestimation of the number of fatalities in official data.[8] Nevertheless, the number of fatalities at the time of the accident remains the most reliably measured metric in official data. However, the issue with recognizing and recording fatalities is symptomatic of a larger problem within official data, which is the lack of unified definitions and methodologies for data collection.[9] This makes the data less comparable between different countries. While, on the European front, significant progress has been made to address these issues, they still hold for other regions, especially less developed ones. Finally, data gathered at accident sites provide only part of the full picture;[10] and additional data and metrics are needed to provide better insights.

Several recommendations are currently available to improve the quality and availability of road-safety data. These include the development of national observatories on road safety;[11] the development of common data collection surveys for the police first responders; and the establishment of proper road data systems for reliable and timely data on road safety.[12]

[7] International Transport Forum, 2011.

[8] Ibid.

[9] Directorate General Transport and Energy of the European Commission, 2007.

[10] International Transport Forum, 2011; World Health Organization (WHO), 2010. Data Systems: A Road Safety Manual for Decision-makers and Practitioners; and Sayed, T., M. H. Zaki and J. Autey, 2013. A novel approach for diagnosing road safety issues using automated computer vision techniques. 16th International Conference on Road Safety on Four Continents, Beijing, China, 15-17 May 2013. Swedish National Road and Transport Research Institute.

[11] Morris, Thomas, P. and others, 2005. Designing the European road safety observatory. *International Journal of Injury Control and Safety Promotion*, vol. 12, No. 4, pp. 251-253.

[12] WHO, 2010.

To be effective, data collection surveys need to conform to international definitions and standards on the different aspects of car accidents, such as the severity of non-fatal injuries and the 30-day follow-up rule.[13] Data for road safety needs to be supplemented from other sources for greater insight. It is suggested to link data from transport and health services to supplement police data, thus providing improved insights to support initiatives and legislation, and addressing the issue of underreporting in official data.[14]

Official data capture critical information on car accidents. Police accident datasets have the benefit of granularity compared to those created by NSOs. But the information gathered can still be augmented by additional data to complement it and make it more comprehensive. This can be done by turning to complementary data, which are defined in this paper as data that support or augment official data gathered from sources other than those collected by police officers at the scene of a crash. One key area that falls under complementary data are big data. Commonly defined as data that are too large and complex for traditional computational tools, big data are characterized by volume, velocity, variety, and veracity.[15] These four characteristics describe data sets that are large, expanding very quickly, structured and/or unstructured, and are of a quality level that enables decision making.[16] Examples of big data include sensor data, closed-circuit television (CCTV) footage, and transaction data. The focus on complementary data instead of big data stems from the need to incorporate a variety of data regardless of their volume and velocity. A dataset is considered “right” if it carries valuable information for the intended analysis.[17]

Complementary data can help improve the reliability of official datasets. As previously stated, best practices for improving data quality include linking police data to transport and health-service datasets.[18] This linkage would allow for a better understanding of injuries and fatalities from different incidents on the road and would help address the issue of underreporting. Ghandour, Hammoud, and Telesca used crowdsourced accident reports from social media to conduct spatial analyses and hazard vulnerability analyses.[19] Crowdsourced data were used by the authors to address the underreporting issues in official data sources, and were parsed to include geolocation, a feature that is not available in official police data in Lebanon.[20]

Complementary data can also help shape a more complete picture of road safety. Several studies were conducted on improving insights into road safety. Chong and Sung highlighted different road

[13] International Transport Forum, 2011.

[14] Ibid.

[15] Economic and Social Commission for Western Asia (ESCWA), 2015. Official Statistics and Emerging Sources of Data: Implications for ESCWA Statistical Activities: Big Data (E/ESCWA/SD/2015/IG.1/5). Eleventh session of the ESCWA Statistical Committee, Amman, 4-5 February 2015.

[16] Ibid.

[17] Wessel, M., 2016a. You Don’t Need Big Data — You Need the Right Data. Harvard Business Review, 3.

[18] International Transport Forum, 2011.

[19] Ghandour, A. J., H. Hammoud and L. Telesca, 2019. Transportation hazard spatial analysis using crowd-sourced social network data. Physica A: Statistical Mechanics and Its Applications. Elsevir, vol. 520(C), 309-316.

[20] Crowdsourced data are data obtained through the contribution of a group of Internet users.

management systems employed in Europe, the United States and South Korea.[21] These systems manage data such as road slopes from fiber-optic sensors, weather and traffic volume from CCTVs to evaluate various aspects of road safety. Their goal is to detect bridge collapses, fires and other road-related disasters. Sayed, Zaki and Autey used video data to detect traffic-related incidents and violations and to calculate key metrics such as time to collision.[22] They demonstrated the insights that can be generated by analysing an intersection and providing detailed statistics for bicycle-car conflicts. Their analyses highlighted areas where safety hazards exist and provided suitable recommendations. A paper published by the Bureau of Infrastructure, Transport and Regional Economics in Australia detailed the different sources of data available for traffic analytics.[23] Traditional sources included sensors and traffic cameras, while new sources include the global system for mobile communications (GSM), the global positioning system (GPS), Bluetooth, and roadside sensors.

A lot of the complementary data sources mentioned by different papers and studies have been available and in use for a long time. Yet, they are referenced to now as possible disruptive innovation sources. Maxwell Wessel stated in his article "How big data is changing disruptive innovation" that disruptive innovation, particularly in the field of technology, is built upon the use of existing sources of data in different ways.[24] Some of the sources of new data are existing data sources coupled with new tools to process the data. One example is the use of CCTV footage which Sayed, Zaki and Autey, Mitchell and Chong and Sung refer to.[25] The availability of cameras and their use in generating data is nothing new, but the real innovation is in the new ways to process the captured footage. Computer vision techniques allow for new data to be extracted from videos recording traffic flows, among others.

[21] Chong, K. and H. Sung, 2015. Prediction of road safety using road/traffic big data. International Conference on Semantic Web Business and Innovation (SWBI2015), p. 23, Sierre, Switzerland.

[22] Sayed, Zaki and Autey, 2013.

[23] Mitchell, D., 2014. New traffic data sources: An overview. Bureau of Infrastructure, Transport and Regional Economics (BITRE). Canberra, ACT, Australia.

[24] Wessel, M., 2016b. How big data is changing disruptive innovation. Harvard Business Review, 27.

[25] Sayed, Zaki and Autey, 2013; Mitchell, 2014; and Chong and Sung, 2015.

II. Introduction to ESCWA's attempt at road safety analytics using complementary data

The heads of statistical offices of the Arab countries recommended in their 13th session of the ESCWA Statistical Committee meeting, which was held in Beirut in January 2019, that the ESCWA Secretariat work on a pilot study on the usage of complementary data for a sector of its choosing to showcase the opportunities and challenges involved. The Statistics Division at ESCWA, in cooperation with the Technology Development Division and the Central Administration of Statistics in Lebanon, proposed a pilot study on the use of new technologies as complementary data sources to complement and update data from official sources. One of the identified projects was to investigate the transport sector regarding the use of complementary sources of data in analysing and understanding the different factors related to road safety in Lebanon, Dubai and Jordan.

As an initial step to analyse road safety, the objective of the study needed to be defined and narrowed down. The objective reached was to identify the means to improve the information accompanying fatal car crashes by highlighting the factors that might lead to severe car accidents. By providing such information, policymakers and urban planners may assume the necessary measures to address these factors, thus cutting down on the number of car crashes and fatalities. To come up with insights into the factors that are correlated with severe car accidents, the type of data that can add value to the analysis needed to be identified. Accordingly, the following datasets were deemed to carry valuable information for analysis:

1. Official police data.
2. Weather and solar azimuth data.
3. Road-related data.
4. Traffic-related data.

A. Official police data

Official police data are valuable sources of information. They detail the observations made at an accident site by police officers. The value of this data type increases significantly if the responding officers are also well trained in reporting and analysing accidents. Typically, police data contain information on the type of accident (such as vehicle-vehicle or vehicle-pedestrian), type of vehicle, road condition, condition of the driver, and others. These features differ from one country to another depending on what the officers are trained to collect and depending on the intended use of the data. Often, data is collected primarily for possible litigation. Police data suffer from underreporting, as was addressed in the chapter on literature, and the quality of the data varies from one country to another based on the level of training and data can be recorded by law enforcement. Despite its shortcomings, official police data remain the most valuable source for analysis.

For the project, an attempt to acquire the official police datasets for Lebanon and Dubai was made. In Lebanon, the data is available from the Internal Security Forces (ISF). Since the ISF is part of the Ministry of Interior, a request was made to acquire the data through the ministry, in addition to a request through the focal point of ESCWA's Statistics Division in the Lebanese Central Administration of Statistics. In Dubai, United Arab Emirates, the data are available from the Dubai Police and Dubai Pulse, an online platform for open data created by the Government of Dubai.[26] ESCWA acquired the car accident data for the 2017, 2018 and 2019 from Dubai Pulse. Additional data was requested to ensure that time-dependent patterns are adequately captured. The requests to Lebanon and Dubai are still pending. The ISF had sent out aggregated data on car accidents, while disaggregated data is still needed for the analysis. The request for official police data from Jordan is currently being worked on.

B. Weather and solar azimuth data

Weather conditions can have a significant effect on visibility and grip on the road; and the position of the sun can potentially be blinding to drivers. As a result, incorporating data on hourly weather conditions and solar azimuth can potentially enhance the final analysis. Features typically included in weather data are temperature, humidity, precipitation, wind speed, and visibility. These features are related to the condition of the road (wet or dry) and the condition of the driver (visibility). Weather data are available at online sources such as Weather Underground;[27] Open Weather Map;[28] or government websites, including the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA), or their European equivalents. Solar azimuth data can be found on such websites as Sun Calc.[29] Weather and solar azimuth data can either be obtained free of charge if the website allows for free data download by resorting to web scraping, by using an application programming interface (API), which can be free depending on the service requested, or by using an API that may be free to the public and not tied to a request for service.

C. Road-related data

In many car crashes, the condition of the road itself is a significant factor. Potholes, slippery surfaces and overall unsafe road design can all play a role in road safety incidents. While official police records include a feature column that describes the road condition, often more information

[26] See https://www.dubaipulse.gov.ae.

[27] See https://www.wunderground.com/.

[28] See https://openweathermap.org.

[29] See https://www.suncalc.org.

is required than that collected by law enforcement. In order to find information on road quality, road maintenance records and high-resolution satellite imagery were considered. Data on the shape of the roads and the quality of the highways, noting that it is easier to extract this information given the highway's width, could be extracted through convolutional neural networks (CNN). Additionally, data from OpenStreetMap (OSM) were sought out.[30] Features in this dataset included the shapefiles and speed limits of the roads in the Middle East. The shapefiles could be processed to extract the shape and curvature of different road segments.

These data sets can be acquired from different sources. Satellite imagery can be typically obtained free of charge from NASA and the European Space Agency.[31] However, free availability depends on each agency's policies and the type of imagery requested as some might be available but need approval or payment. Road shapefiles are available free of charge from OSM. Road maintenance records may be available to the public or through requests to ministries of public works (name of ministry changes from country to country), depending on each country's data policies.

D. Traffic-related data

Data on traffic conditions are critical to understand the severity of car accidents. For instance, it can be assumed that crashes involving fast-traveling vehicles cause more severe damage. As a result, investigating this assumption and incorporating it into the analysis can help determine the factors relating to severe car accidents. To gain this type of insight, different data sources were sought to be incorporated into the project. CCTV camera footage of road segments was a key data source. Utilizing different techniques from computer vision, metrics such as the number of vehicles, average speed, and average deviation of vehicles from their lanes can be calculated. Navigation data from third parties can also provide information on average speed across different road segments. Several efforts were made to acquire official traffic counts originating from either manually counting vehicles, which can be extrapolated for the rest of the year, or from sensors installed on the roads. The attempt to acquire this type of data has not been successful as the requests are still being processed. The possibility to use mobile phone data to estimate traffic flows was also investigated into, yet was later abandoned given the challenges that would need to be overcome.

For the project, various sources were sought to acquire the necessary data. CCTV road footage is usually part of databases at the police and/or the ministry of transportation, while official traffic counts are typically kept at the ministries of transportation and infrastructure. Navigation data can be obtained from widespread navigation software providers such as Google, creator of Google Maps, and Here, creator of HERE WeGo. Since Google's historical data was not available through an API, a partnership was required. In order for ESCWA to access the data, a partnership between Google and the United Nation's headquarters was needed. In the case of datasets from HERE,

[30] OpenStreetMap is a collaborative and crowdsourcing mapping project to create an editable map of the world.

[31] See https://earthdata.nasa.gov/earth-observation-data/visualize-data and https://earth.esa.int/web/guest/data-access/browse-data-products, respectively.

which does not offer historical data through its APIs either, a direct purchase was required. The price of navigation data is usually high and can range from tens of thousands to hundreds of thousands of US dollars depending on the features of the data or the service requested.

E. Challenges

Several hurdles were encountered while trying to implement the project. One trait that characterizes many countries in the Middle East is their unwillingness to share data considering them to be sensitive. Similar issues were encountered during the several phases of implementation with delayed clarity on whether such type of information would be shared or not, without mentioning explicitly the underlying reasons for not sharing such information. Lebanon, for instance, had assured to provide the requested police data, to date such data is still lacking. Moreover, CCTV data is unlikely to be shared for national security reasons. Dubai had been requested to provide several different data sets, but no response was received. Some data were acquired from already published online data by Dubai Pulse, but it is worth noting that some of the datasets were no longer available on the websites a few weeks after obtaining them. This raises the issue of availability, reliability and continuity of the data source. Additional key issues are the prevalence of data silos and the lack of transparency between government entities. Upon contacting different NSOs to request CCTV and road sensor data, their responses were that either such data did not exist or they did not have such data. This might indicate that NSOs are not aware of the type of data available and obtainable at other national agencies or entities; they do not have access to data themselves; and/or there is no regulatory framework put in place to organize and institutionalize acquiring such data. Access to this type of data and linking it to other datasets would provide useful insights for line ministries, urban planners, transport engineers, lawmakers, and other stakeholders. In the wake of the data revolution, there is an increasing emphasis on the key role of NSOs in the collection and dissemination of such data.[32]

Once the necessary data are received, several accompanying challenges arise, including analysing mobile phone data for traffic-related purposes and determining whether or not an individual was in a vehicle stuck in traffic. Additionally, mobile phone tracking is lower in accuracy compared to other technologies such as GPS.[33] Hardware requirements were another encountered key challenge. Once CCTV data or historical high-resolution satellite imagery are received, parallel processing is needed to extract the necessary information on time. Parallel processing refers to a form of computation where tasks are distributed to several processors that are typically contained in separate machines and are run simultaneously to reduce the time needed for the task to be completed. This requires hardware and specific software or code libraries. As a result, it is necessary either to acquire high-end computing devices with dedicated graphics processing units (GPUs) to train deep learning models or to rent out virtual machines on the cloud. A possible way to resolve this issue was envisaged by collaborating with the Qatar Computing Research Institute (QCRI) to utilize some of

[32] International Transport Forum, 2011; and ESCWA, 2015.

[33] Mitchell, 2014.

their computational powers and related services when needed. QCRI expressed its readiness to provide such services to ESCWA when needed.

F. Machine learning pipeline and proposed analysis

The main analytical tools used in this paper are based on machine learning. Machine learning algorithms function differently from traditional programmes. They are algorithms that learn patterns by looking at examples of data (trained on data) rather than having the patterns and logic directly programmed into them. To implement a machine learning algorithm, steps are typically followed in a sequence, referred to as a pipeline. A typical machine learning pipeline involves several steps and algorithms. This pipeline also requires that the data be structured in a certain way to allow a machine learning algorithm to function.

1. Preprocessing stage

The predominant work done during the preprocessing stage revolves around preparing the data to be fed into the model. The dataset is typically divided into feature columns and a target column. The feature columns contain information that can help in predicting the target column. In the case of car crashes, for instance, features can include whether or not the driver was intoxicated or the road was illuminated at night. These are just examples of features that can be related to the fatality of a car crash. Features can be numerical, such as the speed of a car, or categoric, such as the make of a car. Categorical data need to be processed before they can be used by a machine learning model since most models require numeric values as inputs. A target column contains the value to be predicted. It can be a numerical (continuous value or integer) or a categorical value.

Several techniques are available to convert categorical into numerical data. One hot encoding converts a column that contains categorical values into a vector representing each categorical value. If an observation contains a specific categorical value, that value is assigned a 1 in the vector and the remaining values are assigned a 0. In the previous example, car crash fatality would be the target column of the dataset. A fatal car crash would be labeled as a 1 (referred to as a positive value) and a non-fatal crash would be labeled as a 0 (referred to as a negative value). The dataset should contain both positive and negative values to enable a machine learning algorithm to learn the patterns in features for fatal versus non-fatal crashes.

The dataset needs to be divided into a training set and a validation set. The training set is used to enable machine learning, while the validation set is used to evaluate the performance of the trained model on data that the model is not familiar with to get an idea of how the model will perform on future unseen data. An alternative method for splitting the dataset is to use an algorithm called k-fold cross-validation. The algorithm works by dividing the data into k pieces (k is a number specified by the user). The model is then trained on k-1 pieces with one piece held out for validation. This process is repeated with a different piece held out until all k-folds have been held out.

The validation score is aggregated for the k-folds which results in a more robust evaluation of the performance of the model.

2. Training and validation stage

Preprocessed data are necessary for the training and validation stage. In this stage, one or several machine learning algorithms are selected to be trained and validated. Supervised machine learning algorithms are trained on datasets that include a target value to enable the models to detect the patterns associated with target values, while unsupervised machine learning models, which try to segment the dataset based on detected patterns, are trained on datasets without a target value.

Performance evaluation is needed as follow-up to identify the quality of the trained models. Metrics are employed in this task to quantify the performance of each model. Unsupervised models are typically coupled with metrics that measure how well the model was able to detect underlying patterns within the dataset, while supervised models are typically coupled with metrics that measure how well the models were able to predict their target. The models are evaluated during the training phase on the training dataset. For supervised learning, training metrics show how well the model was able to learn and predict values from data that it had already seen. Training metrics alone, however, cannot be used to properly evaluate a model.

Validating the results of the trained models is an important step to ensure that the final result is usable and is done by using the same metrics that were used during the training phase on the validation dataset. The model is fed the feature columns of the validation dataset, and predictions are made. The predictions are then compared to actual values in the target column, and the metric is calculated. The validation metric represents how well the model will perform on unseen data. Both the training and validation metrics are important to diagnose any potential issues within the model, such as variance and bias. They also aid in selecting the best model for the problem.

Following the training and validating stage, the resulting metrics are observed to diagnose any possible issues with the models and to identify actions needed to improve the current results.
In real-world applications, the entire process of preprocessing, training and validating is repeated several times in order to maximize the performance of the final model. This is done by testing different models, trying out different processing techniques and changing different options to measure the impact on the selected metrics.

III. Sample analysis of severe car crashes in the Greater London area: An illustration

Due to the inability to acquire the data requested from the selected countries within the time frame and as a demonstration of what can be achieved by applying the proposed analysis explained above, car crash data was acquired for the United Kingdom from 2005 to 2014. The data were originally gathered by police officers under the instruction of the Department for Transport and downloaded from Kaggle, a Google-owned website geared towards data science that includes dataset hosting.[34] The data is split into three datasets: the first dataset contains general information regarding the crash itself; the second dataset contains information regarding the vehicles involved; and the third dataset contains information regarding casualties.

A. Objective and scope

The scope of the analysis was restricted to the Greater London area to accommodate the available hardware. This helped reduce the amount of observation from two or three million to some three hundred thousand. The objective was changed from understanding severe car crashes to understanding severe casualties, which was done to incorporate the data on casualties and vehicles into the analysis. The United Kingdom defines three levels of casualty severity: fatal, serious and slight.[35] Fatal severity includes instantaneous deaths and deaths within 30 days after the crash as a result of injuries. Serious severity includes injuries that lead to hospitalization and injuries such as fractures, burns and severe cuts in addition to injuries that cause death after the internationally recognized 30-day period. Slight severity includes minor injuries that are neither serious nor fatal.

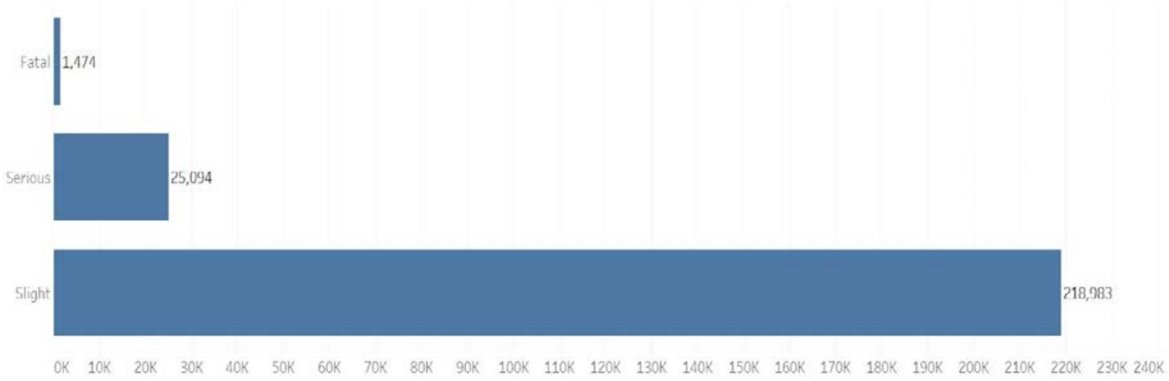
One key issue arises when attempting to predict and explain severe and fatal injuries, namely, class imbalance. Severe and fatal injuries from crashes are typically less frequent compared to slight injuries. It should be mentioned, however, that the number of severe and fatal injuries will be different based on the context of the area under study. Different factors can significantly affect the number of severe and fatal injuries within a dataset for a given area. These factors may include legislation and enforcement, health care and the procedure for recording a car crash. In the Greater London area, severe and fatal injuries are significantly lower than slight injuries, as shown

[34] Department for Transport, 2014. Road Safety Data 2004-2014 [data file]. Available at https://www.kaggle.com/benoit72/uk-accidents-10-years-history-with-many-variables (accessed on 20 May 2019).

[35] Department for Transport, 2004. Stats 20: Instructions for the Completion of Road Accident Reports – With Effect from 1 January 2005. Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/230597/stats20-2005.pdf (accessed on 17 June 2019).

in figure 1. This highlights the class imbalance issue within the dataset where one category, namely, slight injuries, dominates over other categories. The class imbalance makes learning patterns by algorithms difficult. To reduce the impact of class imbalance, data labeled as "fatal" or "serious" were combined into a new category, namely, "severe or fatal". The target of the analysis was changed to predicting and explaining severe or fatal car crash injuries.

Figure 1. Casualty frequency in the Greater London area, 2004-2014



Source: Data compiled by ESCWA.

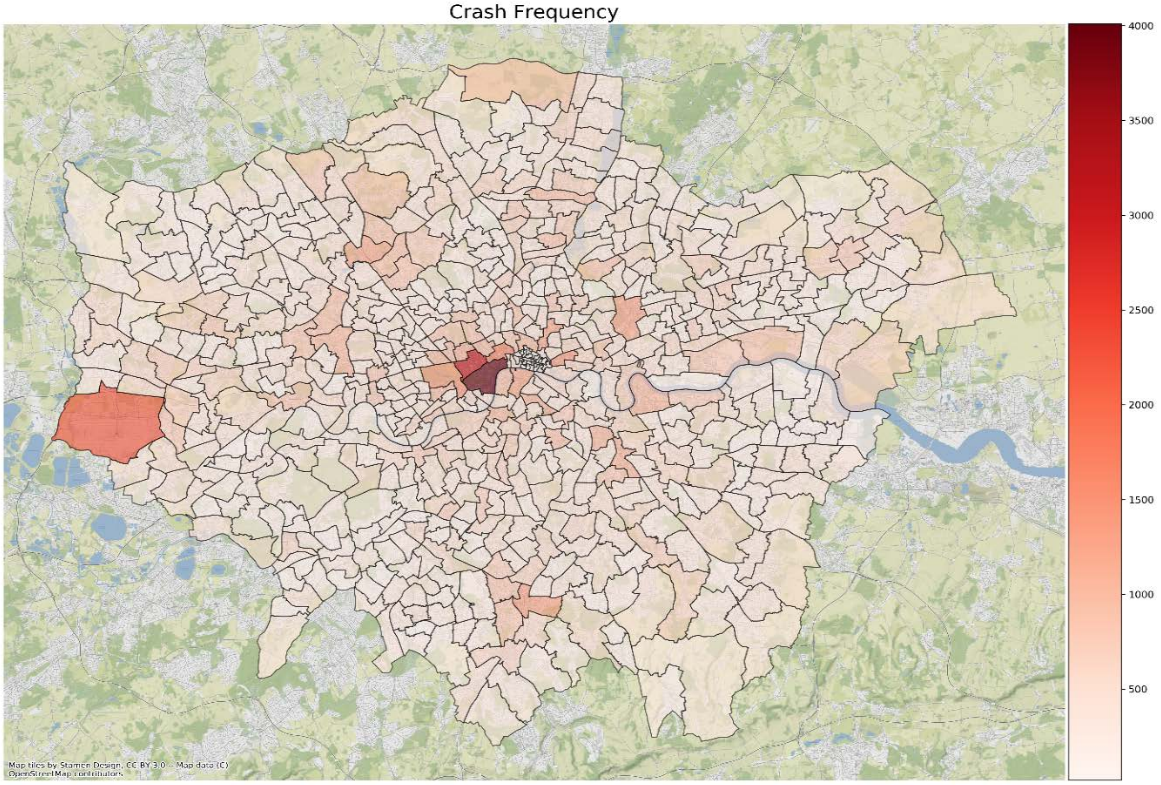
B. Data

The datasets contain several features but only those used in the analysis will be highlighted. By focusing on casualty severity instead of accident severity, more features became available to the model, where accident severity refers to the severity of the most severe injury, while casualty severity indicates the severity of each individual casualty. This was achieved by merging the general data on car crashes with data on vehicle and casualty. Thus, each observation contains information on a single casualty, the vehicle they were in (if they were in a vehicle) and general information about the crash. Table A1 in the annex describes those features in more detail.

In the Greater London area, the heat map generated from the data in figure 2 indicates that the highest occurrences of total car crashes across the years seem to be concentrated in the middle of London. This observation is likely due to the population and traffic density in the area. To get a better idea of at-risk locations, road segments that had at least 100 crashes over the years were highlighted using the ratio of severe or fatal crashes given the total number of crashes in figure 3. The roads were already segmented in OSM data based on the presence of a curve. This is typically done to enable a geometric representation of a road with all its winding features. The 100-crash threshold was selected as the 35 road segments with the highest crash frequencies had a total of 100 crashes or above. Including segments with lower frequencies may have resulted in misleading lethality ratios. Car crashes were assigned to road segments based on proximity as detailed in the section below. The figure paints a different picture compared to the car crash heatmap. Several road segments across Greater London have a high severe or fatal crash ratio given total crashes.

This ratio lies between a maximum of 21 per cent and a minimum of 4 per cent of car crashes for a given segment. However, it is important to note that the highest total crash frequency for a road segment was 194.

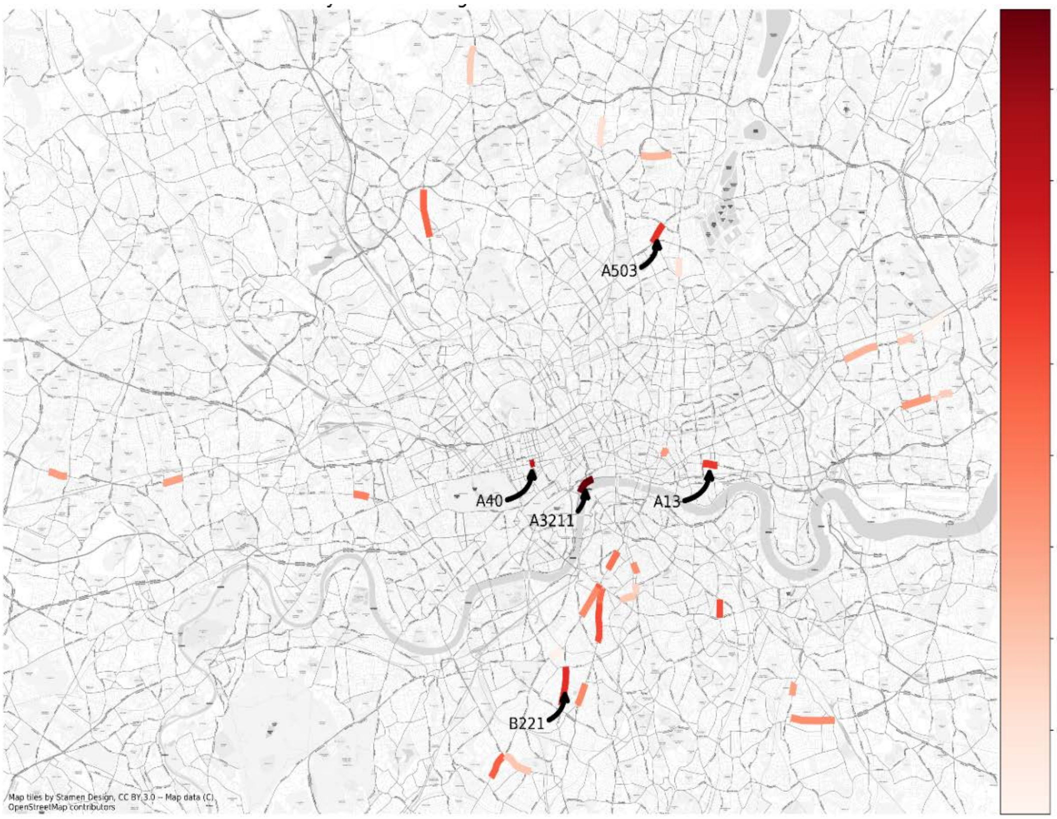
Figure 2. Car crash frequency across the Greater London area


Crash Frequency
4000
3500
3000
2500
2000
1500
1000
500
Map tiles by Stamen Design, CC BY 3.0 — Map data (C) OpenStreetMap contributors

Source: Data compiled by ESCWA.

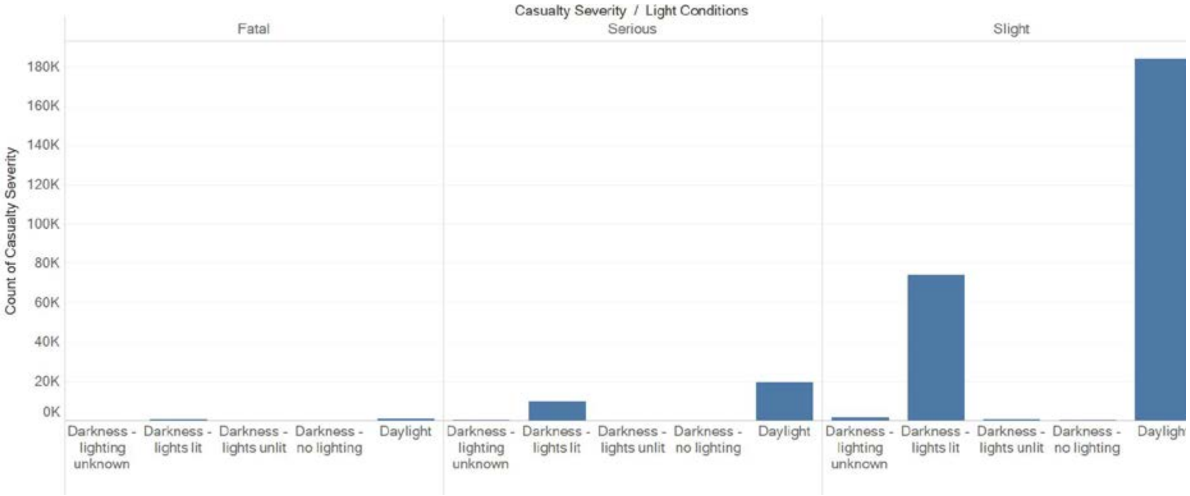
Several features were explored for crash severity. Figures 4, 5 and 6 highlight some challenges with the dataset that may cause problems down the line. Daylight and lights lit in the dark are the two dominating features regardless of casualty severity. "Daylight" has the most frequency of casualty severity followed by "darkness-lights lit", making it difficult to find a pattern. Similarly, comparing casualty severity to the day of the week and hour of the day shows that the frequency of both types of severities increases and decreases in the same periods. The lack of clear relationships between the predictors and the target may result in an underperforming model.

Figure 3. Crash lethality for road segments with more than 100 crashes


A503
A40
A3211
A13
B221
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
Map tiles by Stamen Design, CC BY 3.0 – Map data (C) OpenStreetMap contributors

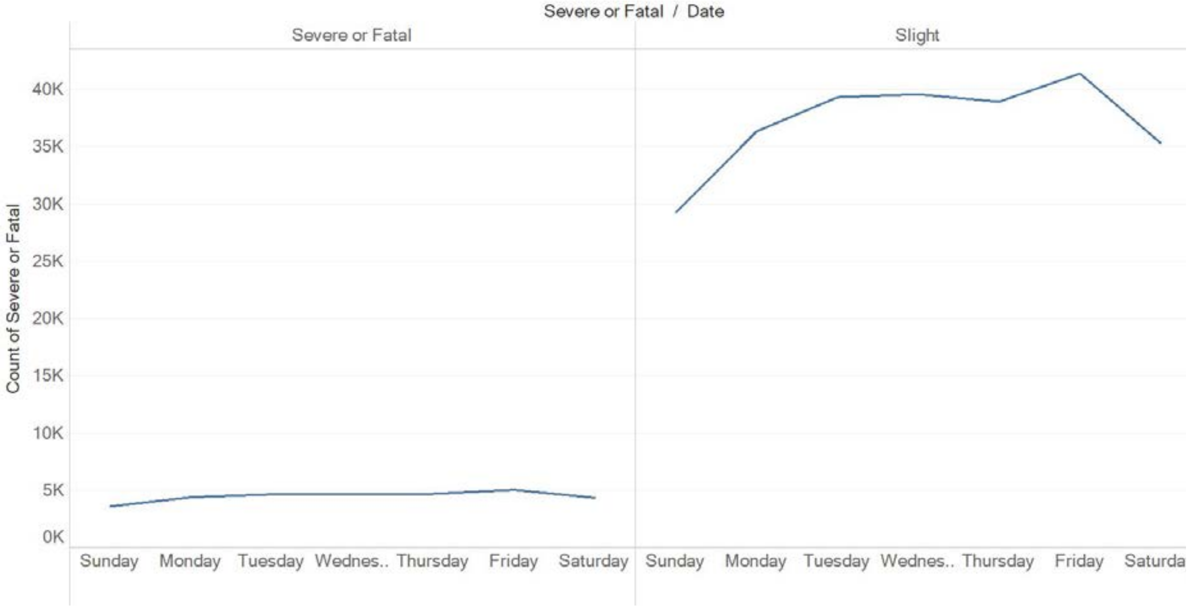
Source: Data compiled by ESCWA.
Note: The five segments with the highest ratio of severe or fatal crashes overall crashes are labeled.

Figure 4. Casualty severity by light conditions



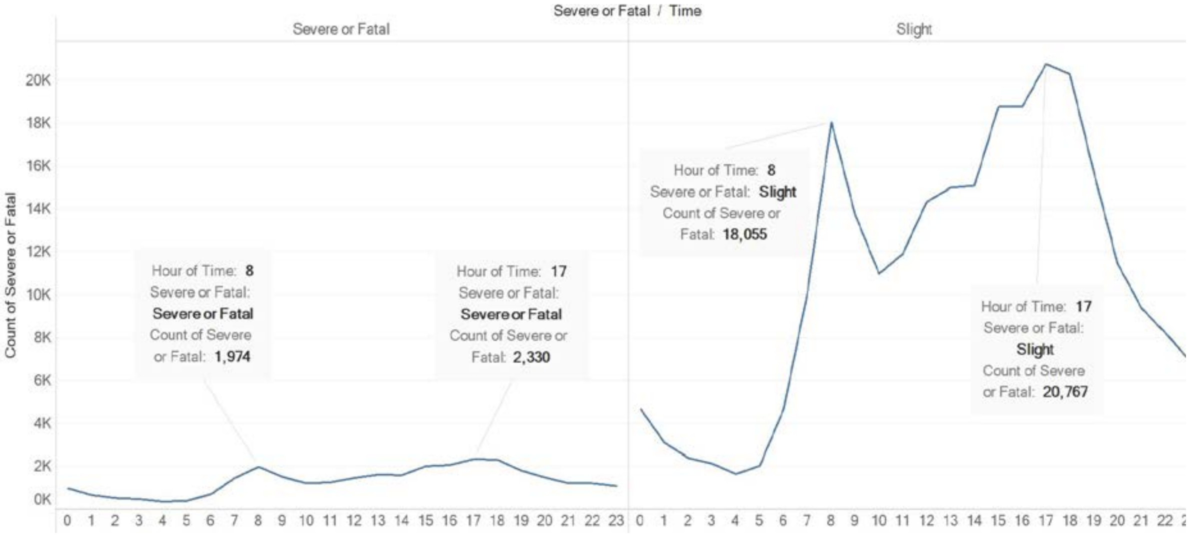
Source: Data compiled by ESCWA.

Figure 5. Frequency of crash severity given the day of the week



Source: Data compiled by ESCWA.

Figure 6. Frequency of crash severity given the hour of the day



Source: Data compiled by ESCWA.

C. Complementary data

Attempting to complement the data provided by the United Kingdom crash dataset, OSM data were acquired.[36] The goal was to incorporate several additional features into the main dataset. Road segment shapefiles for the Greater London area were downloaded from OSM, which include the following data: speed limit, whether there is a bridge or tunnel, whether the road is one way or not, the name of the road, its identification number on OSM, and the multiline string (a series of points connected to each other to create a line that can be plotted) of the road itself. The multiline string data is based on the World Geodetic System (WGS) 84 geographic coordinate system, as is data on the longitude and latitude of the car crash. To assign each car crash to the closest road segment, the multiline strings and the geographic points created by the longitude and latitude of the car crash location were projected from WGS 84 to the EPSG:27700 Ordnance Survey National Grid reference system. This was done to ensure that distances calculated between any two points were relatively accurate. After projecting the data, a k-dimensional tree (kd-tree) was constructed from the multiline strings to speed up the nearest neighbour search. The kd-tree was then queried to retrieve the closest road segment to each crash site.

Road shape data were used to feature engineer new attributes. The main attribute considered was curvature. The decision to extract this particular feature was made based on the literature on the subject. Pande and Abdel-Aty found that the presence of horizontal curvature was not significant for a single vehicle and severe lane-change related crashes but was negatively related to the likelihood of rear-end crashes.[37] To estimate curvature, road sinuosity was used. Sinuosity is typically used in river studies to determine the curvature of rivers and was repurposed to estimate road curvature. Given a certain road segment, sinuosity refers to the ratio of the length of the road divided by the length of the shortest path between the start and endpoint of that road.

$$sinuosity = \frac{road\ length}{shortest\ path}$$

[36] Several factors were considered before deciding to use OSM data and not to confine the analysis to the main road crashes database. In the United Kingdom, the investment in roads was mostly aimed at working on existing roads rather than building new ones. During 2000-2013, 46 motorways were built within the United Kingdom compared to 680 in Germany and 850 in France. Furthermore, the focus of the study is on the Greater London area, which reduces the probability of constructing new roads. The Greater London area is heavily populated, making it harder to develop new roads compared to more rural areas. It is also important to note that the global financial crisis of 2008 happened during the data-collecting period, namely, 2004-2014. This further restricted investments in roads. The focus on road curvature, however, would be less affected by investment on existing roads. There are differences in the data; yet, following an analysis of the roadworks situation, these discrepancies were identified as tolerable especially since this analysis is meant as a simple demonstration rather than a full-blown investigation. Some resources related to this analysis can be found at the following links: https://www.london.gov.uk/what-we-do/transport/improving-londons-roads/road-network and https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/341513/pdfmanforstreets.pdf; https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/212590/action-for-roads.pdf.

[37] Pande, A. and M. Abdel-Aty, 2009. A novel approach for analyzing severe crash patterns on multilane highways. *Accident Analysis and Prevention*, vol. 41, No. 5, pp. 985-994.

Several metrics were extracted based on sinuosity. Overall sinuosity was calculated based on the formula above if a road segment was a polygon, same point for start and end; then a point was interpolated at a distance equal to half of the road segment, effectively cutting the road segment in half; and overall sinuosity was calculated as the average sinuosity of both halves. In addition to overall sinuosity, several other metrics were calculated, namely, maximum (max) change, minimum (min) change, max-min change, and average change. Instead of using the overall distance of the line, these metrics were applied to smaller segments of five and ten meters.

D. Machine learning

The main tool used for analysis is the computer programme Apache Spark. The reason behind this choice was not the size of the data, but rather Spark MLlib's implementation of decision trees. Classification and Regression Trees (CARTs) support categorical features as inputs.[38] Scikit-learn, the *de facto* machine learning package for Python, only supports categorical features that are one-hot encoded. One-hot encoding increases the number of features available and can adversely affect model performance.

The development of the algorithms consisted of several stages. The first stage involved reading the data and converting the encoded data back to their original values. Next, road-based features were added to the main dataset by feature engineering the OSM shapefiles. The features were calculated twice, once based purely on the proximity of a point to a road and once based on the road name similarity, if available. The car crash, vehicle and casualty datasets were merged into casualty, and the features to train the model were selected. The features were divided into subsets to separate related features such as age and age band, and road features calculated using the different methods. Feature sets with the letter "a" have the age of casualty as a numerical feature, while feature sets with the letter "b" have the age of the casualty in ranges rather than numbers. Additionally, feature set 2 includes OSM data assigned based on road names, while feature set 3 includes OSM data assigned based on proximity. More details on the feature sets can be found in table A3 in the annex.

The data was then split into training and test using a stratified sampling approach based on the target (casualty severity). Given the issue of class imbalance, the training set was then undersampled such that the number of negative labels was sampled down to be equal to the number of positive labels. Spark was then used to fit a decision tree model to each feature set. The data were fitted using a ten-fold cross-validation approach. The decision trees were tuned using the grid search algorithm built into Spark's k-fold cross-validation algorithm. The hyperparameters that were tuned were: max depth, which specifies the maximum depth of the tree, and min instance per node, which sets a minimum number of observations required per leaf node. More complex models were also trained to get better results. Random forest algorithm and gradient boosted tree algorithm were selected. Random forest works with several simultaneous deep decision trees.

[38] Breiman, L. and others, 1984. *Classification and Regression Trees*. Chapman and Hall/CRC.

Each decision tree is given a sample of the data and a sample of the features to learn. The outputted prediction is the result of an aggregation of the different predictions using one of many techniques, such as voting where the highest vote determines if the predicted value is 0 or 1. Gradient boosted tree algorithms work by building decision trees sequentially rather than at the same time. Each tree is typically shallow and thus has high bias, and is constructed to better predict the observations that the previous tree got wrong.

The cross-validation algorithm decides on the best model based on a user-selectable metric. The metric chosen for this case was the area under the precision-recall curve. Precision indicates the rate of correctly predicted positive labels given all predicted positive labels. Recall gives the rate of correctly predicted positive labels given all positive labels in the dataset. Improving a model's performance on one can adversely affect the other. The area under the precision-recall curve gives a general value that indicates a model's performance. The higher the area the better the model performs. This metric has the added benefit of not being significantly affected by class imbalance. Class imbalance means that accuracy, the number of correctly classified labels divided by all labels, would have been misleading. If 80 per cent of data was about slight injuries, then a model that simply predicts that all crashes lead to slight injuries would have an 80 per cent accuracy. It is also worth noting that the area under the receiver operating characteristics (ROC) curve and the area under the precision-recall curve are the only metrics supported by the Python implementation of cross-validation in Spark for binary classification. In addition to the area under the precision-recall curve, the F1 score was used to compare the resulting models. The F1 score is a composite metric derived from precision and recall and is the harmonic mean of precision and recall. A higher value indicates a better performing model. An F1 score gives equal weight to, and thus balances, precision and recall. Precision and recall were also calculated along with accuracy. F1 score, precision, recall, area under the ROC curve, and accuracy were calculated on the entire training set, while the area under the precision-recall curve was calculated and averaged across the ten training folds that resulted from the ten-fold cross-validation algorithm. This means that the metrics calculated on the entire training dataset differ compared to those calculated on the folds. However, they are still useful for comparisons between different algorithms.

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \quad recall = \frac{true\ positive}{true\ positive + false\ negative}$$

E. Results

Six different decision trees were constructed from the different feature sets in table A3 of the annex. Table 2 highlights the decision tree metrics for each created model. The results show that the area under the precision-recall curve changes by a negligible amount between each feature set while the F1 score remains unchanged. Decision trees adding or removing the complementary road features extracted from OSM did not affect the model's performance. The slight variation found in other metrics, namely, the area under the precision-recall curve and the area under the ROC curve, can be attributed to the calculation methods rather than the model's performance. More complex models were constructed and tuned to try and achieve better performance and to see if the lack of improvement is due to the fact that the added features do not provide additional useful information for the model, or that the model itself was unable to capture the complexity of the data.

Table 2. Decision tree model details and metrics

Feature Set	Area under ROC curve	Tree depth	Tree nodes	True negatives	True positives	False negatives	False positives	Accuracy	Precision	Recall	F1 score	Area under precision-recall curve
1a	0.6642737	3	15	17 216	8 483	12 124	3 391	0.6235502	0.7144181	0.4116562	0.5223361	0.651076
1b	0.6638445	3	15	17 216	8 483	12 124	3 391	0.6235502	0.7144181	0.4116562	0.5223361	0.651028
2a	0.6642737	3	15	17 216	8 483	12 124	3 391	0.6235502	0.7144181	0.4116562	0.5223361	0.651076
2b	0.6638445	3	15	17 216	8 483	12 124	3 391	0.6235502	0.7144181	0.4116562	0.5223361	0.651028
3a	0.6642737	3	15	17 216	8 483	12 124	3 391	0.6235502	0.7144181	0.4116562	0.5223361	0.651075
3b	0.6638445	3	15	17 216	8 483	12 124	3 391	0.6235502	0.7144181	0.4116562	0.5223361	0.651028

Source: Data compiled by ESCWA.

Gradient boosted trees were trained for all six training sets with ten-fold cross-validation to select the proper hyperparameters. Table 3 details the resulting metrics for each model. With a maximum area under the precision-recall curve of 0.71037, gradient boosted trees outperform even the best performing decision tree model that scored an area of 0.651076. In fact, gradient boosted trees outperformed regular decision trees on almost all the metrics. A deeper dive reveals the reason, namely that decision trees outperformed gradient boosted trees in precision (maximum scores of 0.714418 versus 0.696089). Given all the positive predictions made by decision trees, the rate of correct positive predictions given all positive predictions was better than by gradient boosted trees. When it came to recall, however, and given all positive labels in the dataset, decision trees had a harder time correctly predicting them with a max score of 0.411656 compared to gradient boosted trees with a max score of 0.653322. The area under the precision-recall curve varied slightly across the different feature sets with a minimum of 0.709846 and a maximum of 0.71037. Other metrics, however, showed more pronounced variations. The F1 score increased from 0.661028 for the model trained on feature set 1b to 0.670468 for feature set 2b with a minimum F1 score across all models of 0.649674. Additionally, the area under the ROC curve increased from feature set 1b with a value of 0.740586 to feature set 2b with a value of 0.747752. Both F1 score and area under the ROC curve are at their maximum value when calculated on the model trained on feature set 2b. Feature set 1b does not contain any complementary data on road segments that were extracted from OSM, while feature set 1b contains complementary data of roads assigned through nearest distance with modifications. As a result, gradient boosted trees had a better ability to predict severe or fatal injuries compared to decision trees, but that added performance comes at a tradeoff regarding precision. Feature sets seem to influence, to a certain degree, the performance of the models.

Table 3. Gradient boosted trees model details and metrics

Feature set	Area under ROC curve	True negatives	True positives	False negatives	False positives	Accuracy	Precision	Recall	F1 score	Area under precision-recall curve
1a	0.73961	14 595	13 111	7 496	6 012	0.672247	0.685614	0.63624	0.660005	0.71037
1b	0.740586	14 618	13 130	7 477	5 989	0.673266	0.686751	0.637162	0.661028	0.710245
2a	0.726816	14 381	12 910	7 697	6 226	0.662178	0.674645	0.626486	0.649674	0.710322
2b	0.747752	14 517	13 463	7 144	6 090	0.678896	0.688539	0.653322	0.670468	0.709846
3a	0.747441	14 895	13 083	7 524	5 712	0.678847	0.696089	0.634881	0.664078	0.710289
3b	0.742297	14 585	13 135	7 472	6 022	0.672587	0.68565	0.637405	0.660648	0.710008

Source: Data compiled by ESCWA.

Random forests were also trained for all six training sets with ten-fold cross-validation. Table 4 highlights their key metrics. Their area under the precision-recall curve is lower than that of gradient boosted trees. They scored a maximum of 0.675266 compared to gradient boosted trees that scored 0.71037. However, random forests seem to outperform gradient boosted trees in the F1 score, with a maximum of 0.708988 compared to 0.670468, and in the area under the ROC curve, with a max score of 0.802207 compared to 0.747752. Both metrics were calculated on the training set as a whole rather than averaged across the ten folds. With random forest models, the best performance in F1 and area under the ROC curve metrics were found for feature set 3b (0.708988 and 0.802207, respectively), while the best performance of area under the precision-recall curve was found for feature set 2a. This may indicate that the added complementary data do influence the trained models.

Table 4. Random forest model details and metrics

Feature set	Area under ROC curve	True negatives	True positives	False negatives	False positives	Accuracy	Precision	Recall	F1 score	Area under precision-recall curve
1a	0.78448656	15 646	13 239	7 368	4 961	0.700854079	0.727417582	0.642451594	0.682299585	0.674535107
1b	0.78068293	15 451	13 386	7 221	5 156	0.699689426	0.721928595	0.649585092	0.683848885	0.674331855
2a	0.79984082	15 949	13 376	7 231	4 658	0.711530063	0.741710103	0.64909982	0.692321627	0.675266114
2b	0.801638367	15 791	13 488	7 119	4 816	0.710413937	0.736888112	0.654534867	0.693274395	0.675164656
3a	0.801791074	15 993	13 553	7 054	4 614	0.716892318	0.746023009	0.657689135	0.699076701	0.672177283
3b	0.802207268	15 468	14 139	6 468	5 139	0.718372398	0.733426704	0.686126074	0.708988341	0.672498673

Source: Data compiled by ESCWA.

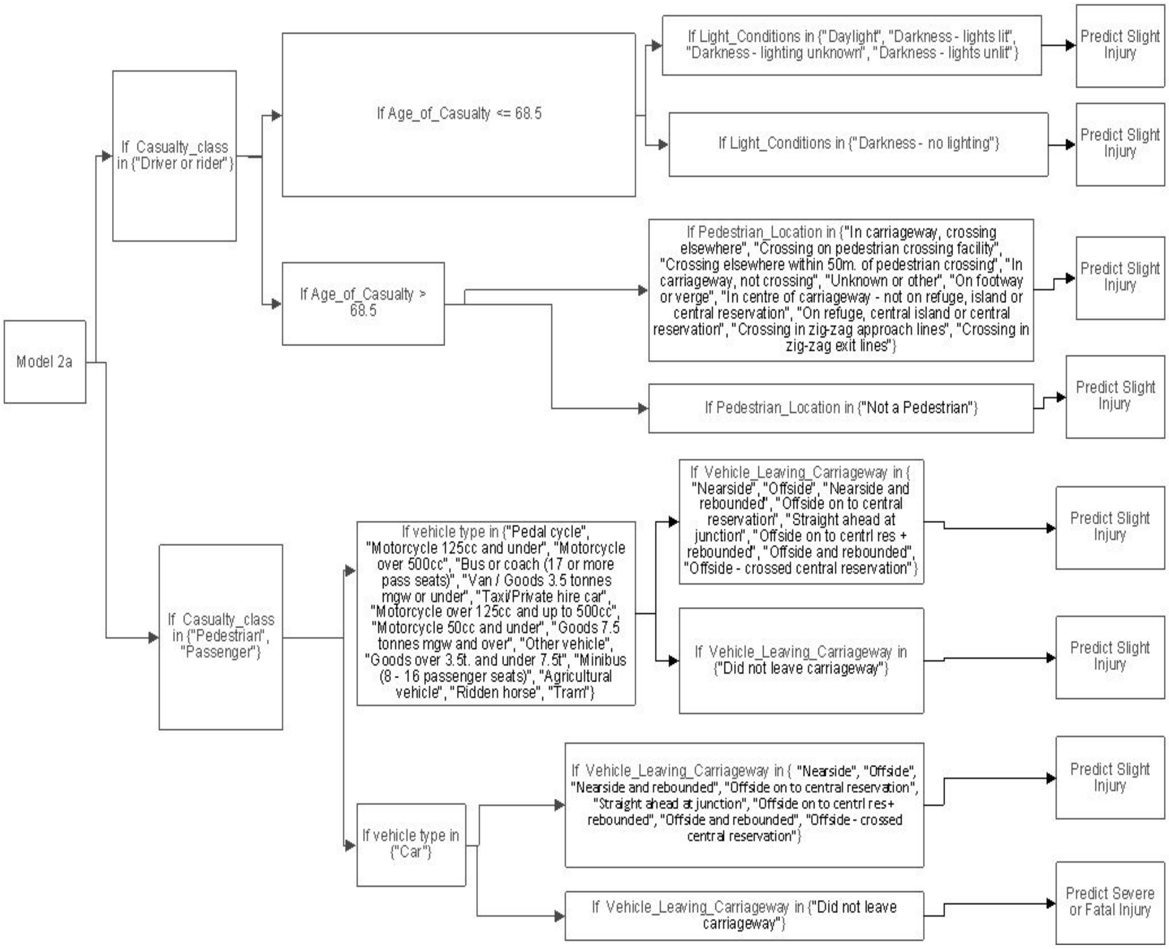
All three models generate feature importance scores. These scores identify which features were more important for predicting the target. For each model type and feature set, the features were sorted from highest to lowest importance. Tables A4, A5, and A6 in the annex highlight the top 15 features (out of 24 for feature sets 1a and 1b, and out of 36 for the rest) for each model. Features that have 0 importance, based on model scores, were dropped from the tables. Decision trees seem to value almost the same features regardless of the feature set inputted. Gradient boosted trees showed slight improvements when shifting to feature sets with the added road features, but the top 15 features do not include any of these features. Random forests, however, seem to value the added road features more as they rank within the top 9 to top 15 features that the models deem important.

F. Discussion

The decision tree models did not benefit from the new features introduced by OSM shapefiles as seen in the results. OSM data was added to feature sets 2 and 3. Models built on feature sets 2a,1a and 3a have the same metrics even though 2a and 3a contain additional OSM features. The lack of variation in the metrics indicates that the features added did not have any impact on the model performance. This is further evident when examining the decision tree (figure 7). No leaf node includes any of the OSM features including the feature engineered curvatures. This, however, does not indicate that the added data were useless. More complex algorithms did show variation between different feature sets. Some metrics varied slightly while others had more significant changes. Random Forest models had the most improvement when the OSM features were added, while gradient boosted trees had less pronounced improvement. This could indicate that road features extracted from OSM can influence the analysis of injury severity. However, the results found here are only the first step, and several methods can be used to further boost performance. Different sampling techniques can be tested and evaluated to see how they perform. Additionally, different implementations of the same model can result in different performance because each implementation uses different underlying algorithms and techniques to create a specific model. For example, scikit-learn implementation of decision trees can be compared to Spark's implementation. Other machine learning models can also be tested. For instance, extreme gradient boosting can be trained and evaluated to compare its performance to that of the above models.

One advantage of decision trees compared to other algorithms is their ability to display trees that denote the step-by-step logic followed to arrive at a certain prediction. These trees can be used to predict the outcome of a certain scenario and can be analysed to better understand the underlying patterns within the dataset itself. Figure 7 highlights the decision tree trained on feature set 2a. The tree indicates that drivers, riders and passengers in vehicles that did not leave the carriageway (perhaps indicating a higher travel speed or no attempt at swerving) are likely to suffer from severe or fatal injuries. This is an example of an insight that can prompt further investigation into this scenario. Better performing trees can provide higher-quality insights. Several steps can be taken to improve the decision tree model's performance. These include adding more features, more aggressive hyperparameter tuning, experimenting with different implementations of the algorithm, and experimenting with other techniques to balance out the classes.

Figure 7. Decision tree trained on feature set 2a


Model 2a
If Casualty_class in {"Driver or rider"}
If Age_of_Casualty <= 68.5
If Light_Conditions in {"Daylight", "Darkness - lights lit", "Darkness - lighting unknown", "Darkness - lights unlit"}
Predict Slight Injury
If Light_Conditions in {"Darkness - no lighting"}
Predict Slight Injury
If Age_of_Casualty > 68.5
If Pedestrian_Location in {"In carriageway, crossing elsewhere", "Crossing on pedestrian crossing facility", "Crossing elsewhere within 50m. of pedestrian crossing", "In carriageway, not crossing", "Unknown or other", "On footway or verge", "In centre of carriageway - not on refuge, island or central reservation", "On refuge, central island or central reservation", "Crossing in zig-zag approach lines", "Crossing in zig-zag exit lines"}
Predict Slight Injury
If Pedestrian_Location in {"Not a Pedestrian"}
Predict Slight Injury
If Casualty_class in {"Pedestrian", "Passenger"}
If vehicle type in {"Pedal cycle", "Motorcycle 125cc and under", "Motorcycle over 500cc", "Bus or coach (17 or more pass seats)", "Van / Goods 3.5 tonnes mgw or under", "Taxi/Private hire car", "Motorcycle over 125cc and up to 500cc", "Motorcycle 50cc and under", "Goods 7.5 tonnes mgw and over", "Other vehicle", "Goods over 3.5t. and under 7.5t", "Minibus (8 - 16 passenger seats)", "Agricultural vehicle", "Ridden horse", "Tram"}
If Vehicle_Leaving_Carriageway in { "Nearside", "Offside", "Nearside and rebounded", "Offside on to central reservation", "Straight ahead at junction", "Offside on to centrl res + rebounded", "Offside and rebounded", "Offside - crossed central reservation"}
Predict Slight Injury
If Vehicle_Leaving_Carriageway in {"Did not leave carriageway"}
Predict Slight Injury
If vehicle type in {"Car"}
If Vehicle_Leaving_Carriageway in { "Nearside", "Offside", "Nearside and rebounded", "Offside on to central reservation", "Straight ahead at junction", "Offside on to centrl res+ rebounded", "Offside and rebounded", "Offside - crossed central reservation"}
Predict Slight Injury
If Vehicle_Leaving_Carriageway in {"Did not leave carriageway"}
Predict Severe or Fatal Injury

Source: Data compiled by ESCWA.

Bridging the gap between data science and official statistics requires merging the experimental nature of data science with the rigor required to create reliable statistics. Additional work is needed to add robustness to the above results. The models need to be tested on several samples from different urban locations within the United Kingdom and compared. Furthermore, the models need to be evaluated using an imbalanced evaluation dataset to get results closer to the real-world performance. Finally, as new data is created, the models need to be evaluated and retrained to see if their previous results still hold. These are some of the steps that can improve the robustness of the results and provide insights that are well tested and more reliable than traditional data science projects.

IV. Conclusion

Several lessons were learned from attempting complementary data analyses in the Arab region and from conducting the related data analytics illustrative exercise applied to data from the United Kingdom. While complementary data can supplement official data, the limitations are usually either prohibitive costs or limited data access. Private companies want to recoup their costs and profit from their data while government agencies are reluctant to share their data. Due to these limitations, data within the region were not accessible. As a result, sample analysis was conducted on the United Kingdom car crash dataset to showcase how data analysis can be improved by complementing official transport data with data extracted from complementary data sources and explain the possibilities which machine learning can provide to support decision making. To that end, geospatial street data were processed to extract curvatures and were incorporated into the machine learning model. The extraction and use of road curvature data for machine learning in the field of road safety are relatively novel, with very few existing cases.

Key lessons and takeaways were demonstrated in the United Kingdom car crash data analysis. First, decision tree outputs can be very useful for decision makers. This paper highlights how decision trees can provide valuable logic behind its predictions, including which features it deems important for its own decision-making and thus providing more useful insights compared to “black box” models. Second, the process of building a good quality model is not straightforward and requires constant tweaking to achieve good results. Third, good results are not always guaranteed, especially on the first iteration of the model building process. Constant iteration and tweaking are required to achieve more usable results. Finally, machine learning is a more experimental approach to dealing with data compared to more established methods employed in official statistics. As a result, procedures and processes must be put into place to ensure that the results from the machine learning model are robust and reliable for official statistics purposes.

The following recommendations could be drawn from the findings of this paper:

1. Establish or strengthen linkages between governmental entities, and maintain them properly to allow data sharing and transparency.
2. Establish a team to experiment with the use of machine learning and unconventional data in different fields to better understand the potential benefits and how to best employ them.
3. Redesign the onsite crash data collection tools and improve police training related to the recording of car crash data and to include road-related data as much as possible, to improve car crash analysis.

Annex

Data features

Table A1. United Kingdom data features and their description

All dataset features	Description
'Casualty_Class',	Casualty is a driver, passenger or pedestrian
'Sex_of_Casualty',	Gender of casualty
'Age_of_Casualty',	Age of casualty
'Age_Band_of_Casualty',	Age of casualty binned into a range
'Casualty_Severity',	Casualty injury severity
'Pedestrian_Location',	If the casualty was a pedestrian where were they located
'Pedestrian_Movement',	If the casualty was a pedestrian how were they moving on the road
'Car_Passenger',	If the casualty was a passenger where were they seated
'Bus_or_Coach_Passenger',	If the casualty was a bus or coach passenger, where were they positioned
'Vehicle_Type',	Type of vehicle
'Vehicle_Manoeuvre',	What maneuver was the vehicle doing before the crash
'Junction_Location',	Identifies the location of a vehicle compared to a junction. A junction is considered when the vehicle is within 20 metres of it.
'Skidding_and_Overturning',	If the vehicle skidded, overturned, or did something similar
'Hit_Object_in_Carriageway',	Identifies what object the vehicle hit on the carriageway, if any
'Vehicle_Leaving_Carriageway',	Identifies the position of the vehicle as it was leaving the carriageway
'Hit_Object_off_Carriageway',	Identifies what object the vehicle hit off the carriageway, if any
'1st_Point_of_Impact',	Identifies the first point of impact
'Sex_of_Driver',	Gender of driver
'Road_Type',	Type of road
'Speed_Limit',	Speed limit of the road
'Junction_Detail',	Describes the type of junction within 20 metres to the vehicle
'Pedestrian_Crossing-Physical_Facilities',	Identifies the type of crossing facilities, if any

All dataset features	Description
‘Light_Conditions’,	Identifies the light conditions during the crash
‘Weather_Conditions’,	Identifies the weather conditions during the crash
‘Road_Surface_Conditions’,	Identifies the condition of the road surface during the crash
‘Did_Police_Officer_Attend_Scene_of_Accident’,	Refers to whether the police attended the scene of the accident

Source: Data compiled by ESCWA.

Table A2. OpenStreetMap features and their description

OpenStreetMap features	Description
‘oneway’,	If the road is one or two ways
‘maxspeed’,	Road speed limit
‘bridge’,	If the road contains a bridge
‘tunnel’,	If the road contains a tunnel
‘oneway_proximity’,	If the road is one way or not assigned based on the closest road to the crash site
‘maxspeed_proximity’,	Road speed limit assigned based on the closest road to the crash site
‘bridge_proximity’,	If the road contains a bridge assigned based on the closest road to the crash site
‘tunnel_proximity’,	If the road contains a tunnel assigned based on the closest road to the crash site
‘max_change_10’,	Maximum sinuosity calculated every ten metres
‘min_change_10’,	Minimum sinuosity calculated every ten metres
‘max_min_range_10’,	Difference between max_change_10 and min_change 10
‘average_change_10’,	Arithmetic mean of the sinuosity calculated every ten metres
‘max_change_5’,	Maximum sinuosity calculated every five metres
‘min_change_5’,	Minimum sinuosity calculated every five metres
‘max_min_range_5’,	Difference between max_change_5 and min_change_5
‘average_change_5’,	Arithmetic mean of the sinuosity calculated every five metres
‘max_change_proximity_10’,	Maximum sinuosity calculated every ten metres assigned based on the closest road to the crash site
‘min_change_proximity_10’,	Minimum sinuosity calculated every ten metres assigned based on the closest road to the crash site
‘max_min_range_proximity_10’,	Difference between max_change_10 and min_change_5 assigned based on the closest road to the crash site

OpenStreetMap features	Description
‘average_change_proximity_10’,	Arithmetic mean of the sinuosity calculated every ten metres assigned based on the closest road to the crash site
‘max_change_proximity_5’,	Maximum sinuosity calculated every five metres assigned based on the closest road to the crash site
‘min_change_proximity_5’,	Minimum sinuosity calculated every five metres assigned based on the closest road to the crash site
‘max_min_range_proximity_5’,	Difference between max_change_5 and min_change_5 assigned based on the closest road to the crash site
‘average_change_proximity_5’,	Arithmetic mean of the sinuosity calculated every five metres assigned based on the closest road to the crash site
‘overall_sinuosity’,	Sinuosity calculated on the entire road segment. If the road was fully connected (oval, circle, etc.), then the road is divided in half and the sinuosity averaged for both halves.
‘overall_sinuosity_proximity’	Sinuosity calculated on the entire road segment. If the road was fully connected (oval, circle, etc.), then the road is divided in half and the sinuosity averaged for both halves. The closest road to the crash site was used.

Source: Data compiled by ESCWA.

Table A3. Set of features selected as input into a machine learning algorithm

Feature set 1a	Feature set 1b	Feature set 2a	Feature set 2b	Feature set 3a	Feature set 3b
‘Casualty_Class’,	‘Casualty_Class’,	‘Casualty_Class’,	‘Casualty_Class’,	‘Casualty_Class’,	‘Casualty_Class’,
‘Sex_of_Casualty’,	‘Sex_of_Casualty’,	‘Sex_of_Casualty’,	‘Sex_of_Casualty’,	‘Sex_of_Casualty’,	‘Sex_of_Casualty’,
‘Age_of_Casualty’,	‘Age_Band_of_Casualty’,	‘Age_of_Casualty’,	‘Age_Band_of_Casualty’,	‘Age_of_Casualty’,	‘Age_Band_of_Casualty’,
‘Pedestrian_Location’,	‘Pedestrian_Location’,	‘Pedestrian_Location’,	‘Pedestrian_Location’,	‘Pedestrian_Location’,	‘Pedestrian_Location’,
‘Pedestrian_Movement’,	‘Pedestrian_Movement’,	‘Pedestrian_Movement’,	‘Pedestrian_Movement’,	‘Pedestrian_Movement’,	‘Pedestrian_Movement’,
‘Car_Passenger’,	‘Car_Passenger’,	‘Car_Passenger’,	‘Car_Passenger’,	‘Car_Passenger’,	‘Car_Passenger’,
‘Bus_or_Coach_Passenger’,	‘Bus_or_Coach_Passenger’,	‘Bus_or_Coach_Passenger’,	‘Bus_or_Coach_Passenger’,	‘Bus_or_Coach_Passenger’,	‘Bus_or_Coach_Passenger’,
‘Vehicle_Type’,	‘Vehicle_Type’,	‘Vehicle_Type’,	‘Vehicle_Type’,	‘Vehicle_Type’,	‘Vehicle_Type’,
‘Vehicle_Manoeuvre’,	‘Vehicle_Manoeuvre’,	‘Vehicle_Manoeuvre’,	‘Vehicle_Manoeuvre’,	‘Vehicle_Manoeuvre’,	‘Vehicle_Manoeuvre’,
‘Junction_Location’,	‘Junction_Location’,	‘Junction_Location’,	‘Junction_Location’,	‘Junction_Location’,	‘Junction_Location’,
‘Skidding_and_Overturning’,	‘Skidding_and_Overturning’,	‘Skidding_and_Overturning’,	‘Skidding_and_Overturning’,	‘Skidding_and_Overturning’,	‘Skidding_and_Overturning’,
‘Hit_Object_in_Carriageway’,	‘Hit_Object_in_Carriageway’,	‘Hit_Object_in_Carriageway’,	‘Hit_Object_in_Carriageway’,	‘Hit_Object_in_Carriageway’,	‘Hit_Object_in_Carriageway’,
‘Vehicle_Leaving_Carriageway’,	‘Vehicle_Leaving_Carriageway’,	‘Vehicle_Leaving_Carriageway’,	‘Vehicle_Leaving_Carriageway’,	‘Vehicle_Leaving_Carriageway’,	‘Vehicle_Leaving_Carriageway’,
‘Hit_Object_off_Carriageway’,	‘Hit_Object_off_Carriageway’,	‘Hit_Object_off_Carriageway’,	‘Hit_Object_off_Carriageway’,	‘Hit_Object_off_Carriageway’,	‘Hit_Object_off_Carriageway’,
‘1st_Point_of_Impact’,	‘1st_Point_of_Impact’,	‘1st_Point_of_Impact’,	‘1st_Point_of_Impact’,	‘1st_Point_of_Impact’,	‘1st_Point_of_Impact’,
‘Sex_of_Driver’,	‘Sex_of_Driver’,	‘Sex_of_Driver’,	‘Sex_of_Driver’,	‘Sex_of_Driver’,	‘Sex_of_Driver’,
‘Road_Type’,	‘Road_Type’,	‘Road_Type’,	‘Road_Type’,	‘Road_Type’,	‘Road_Type’,
‘Speed_limit’,	‘Speed_limit’,	‘Speed_limit’,	‘Speed_limit’,	‘Speed_limit’,	‘Speed_limit’,
‘Junction_Detail’,	‘Junction_Detail’,	‘Junction_Detail’,	‘Junction_Detail’,	‘Junction_Detail’,	‘Junction_Detail’,

Feature set 1a	Feature set 1b	Feature set 2a	Feature set 2b	Feature set 3a	Feature set 3b
‘Pedestrian_Crossing-Physical_Facilities’,	‘Pedestrian_Crossing-Physical_Facilities’,	‘Pedestrian_Crossing-Physical_Facilities’,	‘Pedestrian_Crossing-Physical_Facilities’,	‘Pedestrian_Crossing-Physical_Facilities’,	‘Pedestrian_Crossing-Physical_Facilities’,
‘Light_Conditions’,	‘Light_Conditions’,	‘Light_Conditions’,	‘Light_Conditions’,	‘Light_Conditions’,	‘Light_Conditions’,
‘Weather_Conditions’,	‘Weather_Conditions’,	‘Weather_Conditions’,	‘Weather_Conditions’,	‘Weather_Conditions’,	‘Weather_Conditions’,
‘Road_Surface_Conditions’,	‘Road_Surface_Conditions’,	‘Road_Surface_Conditions’,	‘Road_Surface_Conditions’,	‘Road_Surface_Conditions’,	‘Road_Surface_Conditions’,
‘Did_Police_Officer_Attend_Scene_of_Accident’	‘Did_Police_Officer_Attend_Scene_of_Accident’	‘Did_Police_Officer_Attend_Scene_of_Accident’,	‘Did_Police_Officer_Attend_Scene_of_Accident’,	‘Did_Police_Officer_Attend_Scene_of_Accident’,	‘Did_Police_Officer_Attend_Scene_of_Accident’,
		‘oneway’,	‘oneway’,	‘oneway_proximity’,	‘oneway_proximity’,
		‘bridge’,	‘bridge’,	‘bridge_proximity’,	‘bridge_proximity’,
		‘tunnel’,	‘tunnel’,	‘tunnel_proximity’,	‘tunnel_proximity’,
		‘max_change_10’,	‘max_change_10’,	‘max_change_proximity_10’,	‘max_change_proximity_10’,
		‘min_change_10’,	‘min_change_10’,	‘min_change_proximity_10’,	‘min_change_proximity_10’,
		‘max_min_range_10’,	‘max_min_range_10’,	‘max_min_range_proximity_10’,	‘max_min_range_proximity_10’,
		‘average_change_10’,	‘average_change_10’,	‘average_change_proximity_10’,	‘average_change_proximity_10’,
		‘max_change_5’,	‘max_change_5’,	‘max_change_proximity_5’,	‘max_change_proximity_5’,
		‘min_change_5’,	‘min_change_5’,	‘min_change_proximity_5’,	‘min_change_proximity_5’,
		‘max_min_range_5’,	‘max_min_range_5’,	‘max_min_range_proximity_5’,	‘max_min_range_proximity_5’,
		‘average_change_5’,	‘average_change_5’,	‘average_change_proximity_5’,	‘average_change_proximity_5’,
		‘overall_sinuosity’	‘overall_sinuosity’	‘overall_sinuosity_proximity’	‘overall_sinuosity_proximity’

Source: Data compiled by ESCWA.

Table A4. Top 15 features ranked by decision tree

Rank	Feature 1a	Feature 1b	Feature 2a	Feature 2b	Feature 3a	Feature 3b
1	Casualty_Class Index	Casualty_ClassI Index	Casualty_Class Index	Casualty_ClassI Index	Casualty_Class Index	Casualty_ClassI Index
2	Vehicle_TypeI Index	Vehicle_Type Index	Vehicle_Type Index	Vehicle_Type Index	Vehicle_Type Index	Vehicle_Type Index
3	Vehicle_Leaving_Carriageway Index	Vehicle_Leaving_Carriageway Index	Vehicle_Leaving_Carriageway Index	Vehicle_Leaving_Carriageway Index	Vehicle_Leaving_Carriageway Index	Vehicle_Leaving_Carriageway Index
4	Age_of_Casualty	Light_Conditions Index	Age_of_Casualty	Light_Conditions Index	Age_of_Casualty	Light_Conditions Index
5	Light_Conditions Index	Age_Band_of_Casualty Index	Light_Conditions Index	Age_Band_of_Casualty Index	Light_Conditions Index	Age_Band_of_Casualty Index
6	Pedestrian_Location Index	Pedestrian_Crossing - Physical_Facilities Index	Pedestrian_Location Index	Pedestrian_Crossing - Physical_Facilities Index	Pedestrian_Location Index	Pedestrian_Crossing - Physical_Facilities Index
7						
8						
9						
10						
11						
12						
13						
14						
15						

Source: Data compiled by ESCWA.
Note: Random forest algorithm only selected relatively the same six features regardless of the feature set provided as input.

Table A5. Top 15 features ranked by gradient boosted trees

Rank	Feature 1a	Feature 1b	Feature 2a	Feature 2b	Feature 3a	Feature 3b
1	Vehicle_Type Index	Vehicle_Type Index	Vehicle_Type Index	Vehicle_Type Index	Vehicle_Type Index	Vehicle_Manoeuvre Index
2	Vehicle_Manoeuvre Index	Vehicle_Manoeuvre Index	Vehicle_Manoeuvre Index	Vehicle_Manoeuvre Index	Vehicle_Manoeuvre Index	Vehicle_Type Index
3	Age_of_Casualty	Age_Band_of_Casualty Index	Age_of_Casualty	Age_Band_of_Casualty Index	1st_Point_of_Impact Index	Age_Band_of_Casualty Index
4	Junction_Location Index	1st_Point_of_Impact Index	1st_Point_of_Impact Index	1st_Point_of_Impact Index	Age_of_Casualty	1st_Point_of_Impact Index
5	Junction_Detail Index	Hit_Object_off_Carriageway Index	Skidding_and_Overturning Index	Junction_Location Index	Hit_Object_off_Carriageway Index	Hit_Object_off_Carriageway Index
6	Pedestrian_Location Index	Junction_Detail Index	Did_Police_Officer_Attend_Scene_of_Accident Index	Junction_Detail Index	Junction_Detail Index	Junction_Detail Index
7	Hit_Object_off_Carriageway Index	Junction_Location Index	Hit_Object_off_Carriageway Index	Vehicle_Leaving_Carriageway Index	Junction_Location Index	Skidding_and_Overturning Index
8	1st_Point_of_Impact Index	Skidding_and_Overturning Index	Junction_Detail Index	Skidding_and_Overturning Index	Pedestrian_Location Index	Pedestrian_Location Index
9	Vehicle_Leaving_Carriageway Index	Did_Police_Officer_Attend_Scene_of_Accident Index	Vehicle_Leaving_Carriageway Index	Hit_Object_off_Carriageway Index	Skidding_and_Overturning Index	Did_Police_Officer_Attend_Scene_of_Accident Index
10	Pedestrian_Movement Index	Hit_Object_in_Carriageway Index	Light_Conditions Index	Weather_Conditions Index	Vehicle_Leaving_Carriageway Index	Junction_Location Index
11	Did_Police_Officer_Attend_Scene_of_Accident Index	Pedestrian_Movement Index	Pedestrian_Location Index	Pedestrian_Movement Index	Hit_Object_in_Carriageway Index	Hit_Object_in_Carriageway Index
12	Skidding_and_Overturning Index	Vehicle_Leaving_Carriageway Index	Hit_Object_in_Carriageway Index	Pedestrian_Location Index	Weather_Conditions Index	Pedestrian_Movement Index

Rank	Feature 1a	Feature 1b	Feature 2a	Feature 2b	Feature 3a	Feature 3b
13	Hit_Object_in_Carriageway Index	Pedestrian_Location Index	Junction_Location Index	Hit_Object_in_Carriageway Index	Did_Police_Officer_Attend_Scene_of_Accident Index	Vehicle_Leaving_Carriageway Index
14	Weather_Conditions Index	Weather_Conditions Index	Speed_limit	Did_Police_Officer_Attend_Scene_of_Accident Index	Pedestrian_Movement Index	Weather_Conditions Index
15	Light_Conditions Index	Light_Conditions Index	Casualty_Class Index	Light_Conditions Index	Light_Conditions Index	Road_Type Index

Source: Data compiled by ESCWA.

Table A6. Top 15 features ranked by random forest

Rank	Feature 1a	Feature 1b	Feature 2a	Feature 2b	Feature 3a	Feature 3b
1	Vehicle_Type Index	Vehicle_Type Index	Vehicle_Type ndex	Vehicle_Type Index	Vehicle_Type Index	Vehicle_Type Index
2	Pedestrian_Movement Index	First_Point_of_Impact Index	Pedestrian_Movement Index	Pedestrian_Movement Index	Pedestrian_Movement Index	Pedestrian_Movement Index
3	First_Point_of_Impact Index	Pedestrian_Movement Index	Casualty_Class Index	Casualty_Class Index	Casualty_Class Index	Vehicle_Manoeuvre Index
4	Age_of_Casualty	Casualty_Class Index	Vehicle_Manoeuvre Index	Vehicle_Manoeuvre Index	Vehicle_Manoeuvre Index	Casualty_Class Index
5	Casualty_Class Index	Pedestrian_Location Index	1st_Point_of_Impact Index	Age_Band_of_Casualty Index	1st_Point_of_Impact Index	Age_Band_of_Casualty Index
6	Vehicle_Manoeuvre Index	Age_Band_of_Casualty Index	Pedestrian_Location Index	1st_Point_of_Impact Index	Age_of_Casualty	1st_Point_of_Impact Index
7	Pedestrian_Location Index	Vehicle_Manoeuvre Index	Age_of_Casualty	Pedestrian_Location Index	Pedestrian_Location Index	Pedestrian_Location Index
8	Vehicle_Leaving_Carriageway Index	Vehicle_Leaving_Carriageway Index	Skidding_and_Overturning Index	Skidding_and_Overturning Index	Hit_Object_off_Carriageway Index	Skidding_and_Overturning Index
9	Junction_Location Index	Junction_Location Index	Hit_Object_off_Carriageway Index	Hit_Object_off_Carriageway Index	Skidding_and_Overturning Index	Overall_sinuosity_proximity
10	Skidding_and_Overturning Index	Skidding_and_Overturning Index	Overall_sinuosity	Overall_sinuosity	Hit_Object_in_Carriageway Index	Hit_Object_in_Carriageway Index
11	Hit_Object_in_Carriageway Index	Hit_Object_in_Carriageway Index	Average_change_5	Average_change_5	Overall_sinuosity_proximity	Junction_Location Index
12	Junction_Detail Index	Junction_Detail Index	Hit_Object_in_Carriageway Index	Max_min_range_10	Junction_Location Index	Hit_Object_off_Carriageway Index
13	Hit_Object_off_Carriageway Index	Hit_Object_off_Carriageway Index	Junction_Location Index	Hit_Object_in_Carriageway Index	Vehicle_Leaving_Carriageway Index	Max_min_range_proximity_10

Rank	Feature 1a	Feature 1b	Feature 2a	Feature 2b	Feature 3a	Feature 3b
14	Did_Police_Officer_Attend_Scene_of_Accident Index	Did_Police_Officer_Attend_Scene_of_Accident Index	Average_change_10	Junction_Location Index	Max_min_range_proximity_10	Average_change_proximity_5
15	Light_Conditions Index	Light_Conditions Index	Max_min_range_10	Average_change_10	Average_change_proximity_5	Average_change_proximity_10

Source: Data compiled by ESCWA.

Bibliography

International Transport Forum (2011). Reporting on serious road traffic casualties: Combining and using different data sources to improve understanding of non-fatal road traffic crashes. Paris. Available at https://www.itf-oecd.org/sites/default/files/docs/road-casualties-web.pdf.

Breiman, L. and others (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Chong, K. and H. Sung (2015). Prediction of road safety using road/traffic big data. International Conference on Semantic Web Business and Innovation (SWBI2015), p. 23, Sierre, Switzerland.

Department for Transport (2004). Stats 20: Instructions for the Completion of Road Accident Reports – With Effect from 1 January 2005. Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/230597/stats20-2005.pdf (accessed on 17 June 2019).

Department for Transport (2014). Road Safety Data 2004-2014. [Data file]. Available at https://www.kaggle.com/benoit72/uk-accidents-10-years-history-with-many-variables (accessed on 20 May 2019).

Directorate General Transport and Energy of the European Commission (2007). Best Practices in Road Safety: Handbook for Measures at the European Level. Available at https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/projects_sources/supreme_d_handbook_for_measures_at_the_european_level.pdf.

Dubai Pulse (2019). Dubai Traffic Accidents. Available at https://www.dubaipulse.gov.ae/.

United Nations Economic and Social Council (ECOSOC) (2019). Special Edition: Progress towards the Sustainable Development Goals. Report of the Secretary-General. Advanced unedited version. New York: United Nations.

Economic and Social Commission for Western Asia (ESCWA) (2015). Official Statistics and Emerging Sources of Data: Implications for ESCWA Statistical Activities: Big Data (E/ESCWA/SD/2015/IG.1/5). Eleventh session of the ESCWA Statistical Committee, Amman, 4-5 February 2015.

Ghandour, A. J., H. Hammoud and L. Telesca (2019). Transportation hazard spatial analysis using crowd-sourced social network data. Physica A: Statistical Mechanics and Its Applications. Elsevir, vol. 520(C), 309-316.

Mitchell, D. (2014). New traffic data sources: An overview. Bureau of Infrastructure, Transport and Regional Economics (BITRE). Canberra, ACT, Australia.

Pande, A. and M. Abdel-Aty (2009). A novel approach for analyzing severe crash patterns on multilane highways. *Accident Analysis and Prevention*, vol. 41, No. 5, pp. 985-994.

Sayed, T., M. H. Zaki and J. Autey (2013). A novel approach for diagnosing road safety issues using automated computer vision techniques. 16th International Conference on Road Safety on Four Continents, Beijing, China, 15-17 May 2013. Swedish National Road and Transport Research Institute.

Morris, Thomas, P. and others (2005). Designing the European road safety observatory. *International Journal of Injury Control and Safety Promotion*, vol. 12, No. 4, pp. 251-253.

Traffic Safety Statistics between project partner countries (Egypt, Jordan, Lebanon, Poland, Spain, and Sweden) and internationally (2015). MENASAFE.

United Nations General Assembly (2020). Improving global road safety. Resolution adopted by the General Assembly on 31 August 2020 (A/RES/74/299). Available at https://undocs.org/en/A/RES/74/299 (accessed on November 2020).

Wessel, M. (2016a). You Don’t Need Big Data – You Need the Right Data. Harvard Business Review, 3.

Wessel, M. (2016b). How big data is changing disruptive innovation. Harvard Business Review, 27.

World Health Organization (WHO) (2010). Data Systems: A Road Safety Manual for Decision-makers and Practitioners.

__________ (2018). Global Status Report on Road Safety.

20-00560