Optimal placement of ambulances to best serve emergency calls

MSc Research Project
Data Analytics

Apurv Dubey
Student ID: x21141495

School of Computing
National College of Ireland

Supervisor: Dr Giovani Estrada

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Apurv Dubey
Student ID:	x21141495
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr Giovani Estrada
Submission Due Date:	15/12/2022
Project Title:	Optimal placement of ambulances to best serve emergency calls
Word Count:	XXX
Page Count:	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Apurv Dubey
Date:	1st February 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Optimal placement of ambulances to best serve emergency calls

Apurv Dubey
x21141495

Abstract

Every day, the number of traffic accidents rises as the automobile population increases. According to a survey by the World Health Organization (WHO), 1.3 million people die and 50 million are wounded annually around the globe [1]. Most people die because they don't get medical help at the scene of an accident or because it takes too long for rescuers to get there. This study suggests a specific way to shorten the time it takes for an ambulance to arrive at the scene of a traffic accident in the UK, which is the installation of different ambulance stations or parking areas where ambulances can be found in case of an emergency. Two geocoding techniques and two clustering techniques are used to answer the research question. We discovered that, for geocoding and clustering, one technique performed significantly better than the other.

1 Introduction

1.1 Rationale

One billion and four hundred forty six million automobiles are now navigating the globe's highways, roads, and streets [2]. When one of these large, powerful devices is not under careful human supervision, tragic consequences may ensue. The rise of human civilization to its current pinnacle may be directly attributed to the advent of modern transportation systems, which set humans far apart from any other species on Earth. We can't imagine our lives without the convenience of cars. We use it to go to work, stay in contact with loved ones, and transport things. However, it also has the potential to cause catastrophic events, including accidental death. One of the most fundamental and crucial dangers on the road is speed. It increases the likelihood of being in an accident and also makes any collisions that do occur more severe. There are still occasional accidents despite the fact that governments and non-governmental groups from throughout the globe have implemented several campaigns to educate against irresponsible driving. However, if the rescue workers had been able to get to the victims in time, many lives may have been spared. Therefore, it is crucial to have an efficient recovery system in order to save lives. The public and private harm caused by automobiles rubbernecking other users after an accident leads to increased traffic congestion, which in turn delays the arrival of the ambulance, which may have prevented additional injury to the sufferer.

[1] https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[2] https://www.pd.com.au/blogs/how-many-cars-in-the-world/

This research provides an addition to the existing recovery system with the use of clustering algorithms. The methodology followed in this research is KDD (Knowledge Discovery in Databases). The dataset used in the research is publicly available on the UK.gov website [3] and consists of accident data from 2011 to 2019. The clustering algorithms used in the research are K-means and DBSCAN.

1.2 Process Flow of Existing System

Emergency calls are the foundation of the current system. In the event of an accident, someone must step forward and request medical assistance by dialing 911. It's possible that the ambulance won't be nearby, in which case the emergency services telephone operator will direct it there. Unfortunately, this often results in a wait of more than 15 minutes before help can reach the injured. According to Breen et al. (2000) wait for longer than 15 minutes in severe accident leads to catastrophe. This research will tend to introduce a new approach to recovery as the population increases through the use of clustering algorithms.

1.3 The research question

What is the optimal geographical distribution of ambulances to minimize the response time to emergency calls?

1.3.1 Research Objectives

- Find the quickest algorithm for geocoding.
- Find the most optimal algorithm for clustering spatial data.
- Find information about accidents in the United Kingdom from 2011 to 2019.

2 Related Work

As the number of people living in cities continues to rise, public officials are increasingly worried about meeting residents' expectations for access to pre-hospital emergency medical treatment. There are estimates indicating that by 2050, urban areas should have a contingent of 6.29 billion people, equivalent to 69 percent of the world's total population McDonald (2015). The frequency of traffic accidents and other major incidents that call for assistance on the road is bound to rise significantly. The primary intent of ambulances is to provide the medical care on the go. Unfortunately, ambulances may only be kept in certain areas, such as hospitals. Since this is the case, it takes longer for emergency calls sent beyond the first eight minutes to reach their destination. Multiple ambulances are sent from various UK hospitals in the event of an emergency, with a goal of getting to the patient within 8 minutes. This study aims to determine "the optimal geographical distribution of ambulances needed that minimizes the response time to emergency calls". This study will also show various clustering algorithms and their performance on location data.

[3] https://www.gov.uk/government/statistics/reported%2Droad%2Dcasualties%2Dgreat%2Dbritain%2Dannual%2Dreport%2D2019.

2.1 Response time

According to the World Health Organization, the primary purpose of Emergency Medical Services (EMS) is to "provide basic life support in all risk circumstances involving people and products" Nogueira et al. (2016). Acute pre-hospital treatment for patients with diseases and injuries is provided by emergency medical services (EMS), which is discussed here because of its importance in delivering excellent care to patients, reducing the severity of their injuries, and saving lives Aringhieri et al. (2016). For pre-hospital treatment to be effective, the EMS response time is a crucial aspect that must be managed to improve patients' chances of survival. Lawner et al. (2016).

Response time is the main indicator of this service. It is defined as the time between notification of an occurrence and the ambulance's arrival at the scene Cabral et al. (2018). According to the WHO, an ideal response time is less than 8 minutes. Regarding the definition of response time, Lawner et al. (2016) consider secondary outcomes, which include changes in other main ambulance time metrics such as average ambulance response interval (time from ambulance dispatch to arrival at the scene) and overall out-of-service interval (the amount of time that an ambulance is not available to respond to another incident). Another authors (Vile et al. (2016)) have remarked that response times are one of Welsh Ambulance Service Trust's (WAST) Key Performance Indicators (KPIs) since they are believed to provide a good indication of the quality and timeliness of care provided by the service. Worldwide, this parameter is quantified because of its relevance in the evaluation of quality of service. Examples identified in cities with more than one million inhabitants in Brazil and in the world explain the average response time for urgent care: in the UK, it should be a maximum of 8 minutes in 75 percent of calls and 19 minutes in serious, but not urgent, calls.

2.2 Geocoding and Reverse Geocoding

The term "Geocoding" refers to the process of georeferencing data representing location information, and is most often used as an address matching technology. Since this address matching method can transform string-based addresses into geographic information system (GIS) data, it has broad applicability. The best-known geocoding methodology for road-based addresses was created by the US Census Bureau, and it involves linear interpolation based on the address range attribute supplied to each road segment's data (Drummond (1995) ; Lee and Kim (2006)). Road segment data may be used as one of the most lightweight geocoding reference datasets due to the fact that a single data record can include a great deal of address information and has the benefit of being highly usable as basic data that can be applied to a variety of fields. Parcel-based geocoding is a method for locating a location by utilizing the unique identifier of a single parcel, such as a lot number, as a point of reference. This is especially helpful in areas where street networks are sparse or if addresses are erratic. In cases when a single parcel contains several addresses, however, the geocoding match rate approaches a maximum. To overcome these constraints, Lee (2009) presented a BlockObject model, where a closed block is defined by starting at an intersection, as a means of doing parcel-based geocoding using linear interpolation. To compensate for the constraints of geocoding systems based on parcel identifier information, the UK Ordnance Survey (2004) presented a geocoding methodology employing address-point data with location information on specific building addresses as an attribute. This method has the advantage of being adaptable to any address system and providing pinpoint location precision, but it suffers from the drawback of being very sensitive to the quality of the input data, which in turn has a significant impact on the geocoding outcome. In addition, it lacks location information for places for which addresses are not issued; hence, it cannot be utilized as reference data for reverse geocoding Lee (2009). Such a thing does exist. As a result, geocoding quality varies depending on the quality of the reference data used by the geocoding method for address matching. Studies have shown that road-based geocoding, followed by address-point geocoding, and finally parcel-based geocoding, achieve the highest geocoding matching rates Zandbergen (2008).

Due to the dataset containing only coordinate data from the accident, this research uses a reverse geocoding technique for the extraction of an address from the location data.

2.3 Existing Algorithms

2.3.1 Daily Ambulance Demand Prediction

Lin et al. (2020) analyzed data on ambulance calls made to the Singapore Civil Defense Force (SCDF) during a ten-year period. In place of the previously used clustering approach, this research used a massive dataset comprised of geographical characteristics, past ambulance records, and event records. The purpose of these projections is to facilitate easier staffing choices and other pre-shift preparations by providing an idea of what tomorrow's demand will be like in ALL departments.

Population and age distribution maps, as well as district classifications, were used to describe regional attributes such as climate and economy (financial, residential, etc.). Time dimensions such as weekday and month were also taken into account. Furthermore, historical criteria were employed to account for pandemics and/or major sporting events.

There are additional databases that record details about each ambulance call, such as the time it was placed, the nature of the incident, the patient's condition, the age of the patient, the hospitals from which the ambulance originated and to which it was transported, the gender of the patient, and their exact location.

Multiple machine learning methods were applied to these massive datasets for this research, and their relative merits were weighed.

These results suggest that linear regression and the light gradient boosting machine are the most effective models (LightGBM). LightGBM is a powerful and efficient gradient-boosting decision tree method. This method, which is comprised of many separate regression trees, is more suited to capturing non-linear interactions, such as the integration of socioeconomic factors with other factors in a specific region (although socioeconomic features did not improve prediction performance).

2.3.2 Dispatch of ambulance by using decision tree

The study relied on the experience, insight, and dispatch method of the Dutch EMS service in the region of Brabant-Zuidoost (BZO). The primary goal was to improve EMS resource allocation so that they could meet the national goal of responding to extremely urgent ambulance calls 95% of the time within 15 minutes Theeuwes et al. (2021) In contrast to prior strategies, which focused on forecasting the need for ambulances, this research aimed to enhance dispatch judgments. Instead of trying to predict demand, the computer may use the knowledge of the dispatchers to improve its response times. For example, if a transport is in route to a "moderately urgent" request but a "extremely urgent" one comes in, the dispatcher should immediately re-dispatch the transport to the latter. The decision to re-dispatch, however, is not always clear cut. This element of human judgment and experience in re-dispatch is a part of the "dispatch decision tree" employed by machine learning to enhance dispatch decisions. To train the computer to make dispatch decisions, first a decision tree is built using dispatch policies that human dispatchers were instructed to follow in writing. The dispatcher's prior judgments (experiences) in a similar situation are then included in the decision tree to improve its accuracy. All available information is factored into the dispatcher's "experiences," such as whether ambulances are idling, in transit, waiting, unloading at the hospital, or were assigned to a less urgent request but did not arrive. The model was then refined by simulating events in a realistic spatial setting, which allowed us to account for the dynamic elements of real-world events. Findings of great importance were uncovered. Response times for critical inquiries improved by 0.77 percentage points. According to the results of the study, this efficiency gain is equivalent to adding nearly seven more ambulance shifts per week.

3 Methodology

The technique used in this research to extract information is KDD (knowledge discovery in databases). KDD is defined as the planned, exploratory investigation and modeling of significant data sources. Data pattern discovery and interpretation often entails the iterative use of the following procedures:

3.1 Data Selection

This step is the initial and most crucial step, as it is about selecting the raw data that is used to extract the information. This research required a dataset with location points as well as date of the accident. The intent was to find the accident data closest to Ireland, which turned out to be UK. The dataset was located on the gov.uk website for public use without any requirement of a consent form.

Selecting the raw data that is utilized to extract the information is the first and most important stage. This research required a dataset that included both location points and the date of the incident. The dataset is available for public access on the gov.uk website without the need for a consent form. It was intended to locate the Irish accident dataset, however, this was not possible due to lack of publicly available information.

3.2 Preprosessing

The are 67 columns and 459272 indexes. Data preprossessing involves:

- Removal of noise or outliers
- Extraction of necessary information to model.
- Stratergies for handling missing data fields.

You will of course want to discuss your research as well as evaluation methodology – otherwise how will your examiners know that you have followed an appropriate scientific process and rationalised your choice of evaluation. Note that it may also be useful to base decisions in this section off your discussion of related work in Section 2. You should also include cite any previous work used in defining your methodology.

3.3 Transformation

Data transformation involves exploratory analysis and model and hypothesis selection, which are used for searching for data patterns. This process is about finding useful features to represent the data based on the goal of the task. Multiple methods are used to reduce the effective number of variables under consideration in order to find invariant representations for the data.

3.4 Data Mining

In this step, we select the most appropriate data mining technique, such as classification, regression, or clustering, to achieve our goals.

3.4.1 choice of algorithm

This step involves a decision on which algorithm, technique, or combination of both should be used for the research to search for a pattern or obtain knowledge. For each approach, there is a variety of ways to pick them, and meta-learning is concerned with conveying the reasons behind why certain algorithms perform better than others. Each algorithm has its own essence, its own method of operation, and its own style of producing results; thus, it is important to be familiar with the characteristics of the possibilities and to choose which one best suits the data.

3.4.2 Application of algorithm

Finally, after the techniques have been chosen, they may be applied to the cleaned and processed data. The algorithms may be run several times with different parameter settings in an attempt to get optimal results. These values change depending on the technique used.

3.5 Interpretation/Evaluation

3.5.1 Evaluation

This step involves evaluation of the patterns that were generated and the performance that was obtained through the application of algorithms to the dataset to ensure that it satisfies the goals stated in the earlier phases. To carry out this evaluation, there is a technique called cross-validation that accomplishes data partition, dividing it into training data (which is used to develop the model) and testing data (which is used to measure the accuracy of the algorithm).

3.5.2 Interpretation

If all procedures are followed correctly and the assessment findings are satisfactory, the final step is to apply the acquired information to the context and begin to address its issues. In the event that the findings are unsatisfactory, it is required to return to the previous steps to make modifications, reviewing the data selection and evaluation stages.

Results must be provided in a way that is easily understood. Since mathematical models or descriptions in text format might be difficult for end-users to comprehend, data visualization techniques are essential for producing effective results.

You need to give a completely accurate description of the research procedure you followed, equipment used, the technique(s), applied, set-up of scenarios/case studies run. You must provide an explanation of how the raw data gathered/compiled and analyzed. Describe the statistical techniques used upon the data. Detail all the steps from data collection to final results.

4 Design Specification

The architecture of the research is shown in the figure below.

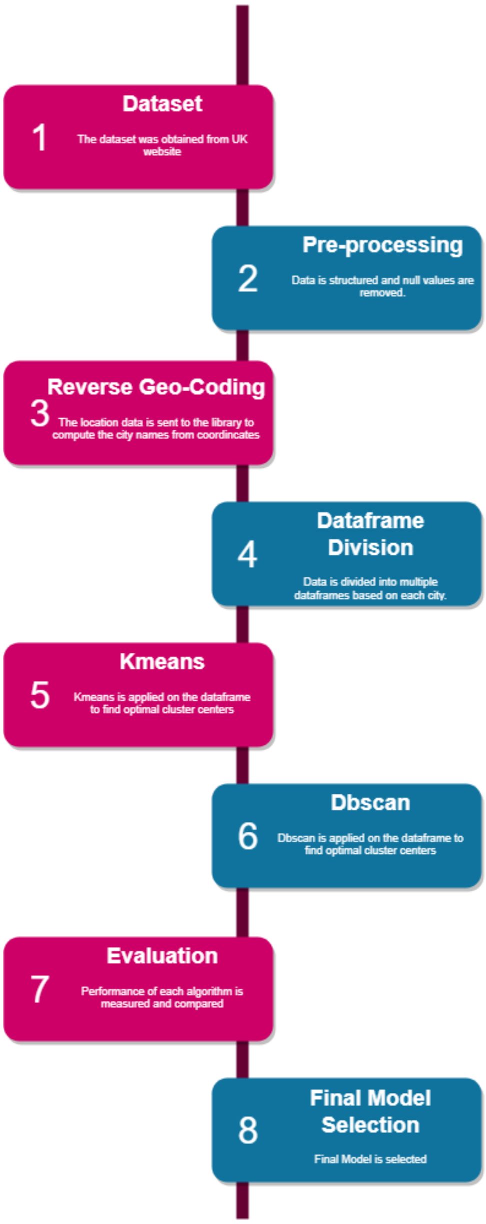

1
Dataset
The dataset was obtained from UK website
2
Pre-processing
Data is structured and null values are removed.
3
Reverse Geo-Coding
The location data is sent to the library to compute the city names from coordincates
4
Dataframe Division
Data is divided into multiple dataframes based on each city.
5
Kmeans
Kmeans is applied on the dataframe to find optimal cluster centers
6
Dbscan
Dbscan is applied on the dataframe to find optimal cluster centers
7
Evaluation
Performance of each algorithm is measured and compared
8
Final Model Selection
Final Model is selected

Figure 1: Design Architecture

5 Implementation

You will of course want to discuss the implementation of the proposed solution. Only the core part of the implementation should be described.

It should describe the outputs produced, e.g. transformed data, code written, models developed, questionnaires administered. The description should also include what tools and languages you used to produce the outputs. This section must not contain code listing or user manual description.

The dataset used for this research was acquired from the UK.gov website. It contains 455781 rows × 68 columns. The most important columns of this dataset required for the research are the latitude and longitude of the accident, the age of the casualty, the age of the vehicle, and the severity of the casualty. The data is structured and required very little pre-processing, which was just the removal of null values.

Table 1: DataColumn

Source Name	pedestrian_location	longitude	local_authority_district_
did_police_officer_attend_scene_of_accident	vehicle_leaving_carriageway	propulsion_code	light_conditions
accident_index	pedestrian_movement	latitude	local_authority_highway_
pedestrian_crossing_human_control	lsoa_of_accident_location	hit_object_off_carriageway	age_of_vehicle
vehicle_reference	car_passenger	police_force	1st_road_class
pedestrian_crossing_physical_facilities	vehicle_type	1st_point_of_impact	driver_imd_decile
casualty_reference	bus_or_coach_passenger	accident_severity	1st_road_number
towing_and_articulation	was_vehicle_left_hand_drive_	driver_home_area_type	2nd_road_number
casualty_class	pedestrian_road_maintenance_worker	number_of_vehicles	road_type
weather_conditions	vehicle_manoeuvre	journey_purpose_of_driver	engine_capacity_cc_
sex_of_casualty	casualty_type	number_of_casualties	speed_limit
road_surface_conditions	vehicle_location_restricted_lane	sex_of_driver	hit_object_in_carriageway
age_of_casualty	casualty_home_area_type	date	junction_detail
special_conditions_at_site	junction_location	age_of_driver	urban_or_rural_area
age_band_of_casualty	location_easting_osgr	day_of_week	junction_control
carriageway_hazards	skidding_and_overturning	age_band_of_driver	2nd_road_class
casualty_severity	location_northing_osgr	time	

5.1 Reverse-Geocoding Algorithms

An address parsing algorithm, or reverse geocoding algorithm, is a software tool used to convert target coordinates in the form of latitude and longitude into a human-readable street address. Its primary use is in the military, but it also has a lot of practical uses in disaster aid and other related industries, as well as for finding out where exactly an address is located using GPS coordinates. Currently, the most common practice for incorporating a reverse geocoding algorithm into a system or project is to make use of an existing online service interface made available by Internet-based product makers. Companies like Baidu Map and Amap Map, who have a sizable portion of the domestic market, have released map API SDKs built on top of map services.

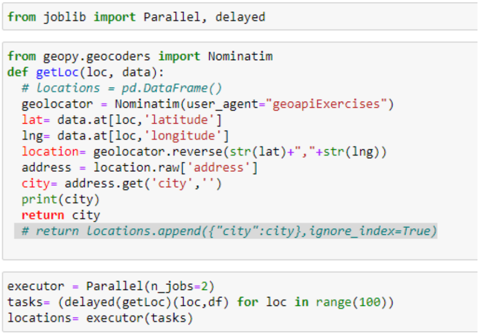
Post-removal of the null values, there arose a need to acquire city names since the data was of accidents in entire UK and just had co-ordinates. There were two algorithms tried for this task:

5.1.1 Server-Based

To begin, the first technique tried to be used in this study was server-based reverse-geocoding, which involved sending the coordinates to the server, which then returned the city name. The implementation of this algorithm was performed in Python. The library used for this was geopy [4]. There are three servers to choose from in this library:

[4] `https://geopy.readthedocs.io/en/stable/`

Bing Server, Google Server, and Nominatim Server. This technique was discarded due to its speed. The data was being processed at the rate of 1 second per process, and the dataset included 455781 entries. To overcome this hurdle, the author even tried to use multi-processing with the help of joblib [5] library, but due to server limitations, it was still not fast enough to process all the requests.



```python
from joblib import Parallel, delayed
```

```python
from geopy.geocoders import Nominatim
def getLoc(loc, data):
  # locations = pd.DataFrame()
  geolocator = Nominatim(user_agent="geoapiExercises")
  lat= data.at[loc,'latitude']
  lng= data.at[loc,'longitude']
  location= geolocator.reverse(str(lat)+","+str(lng))
  address = location.raw['address']
  city= address.get('city','')
  print(city)
  return city
  # return locations.append({"city":city},ignore_index=True)
```

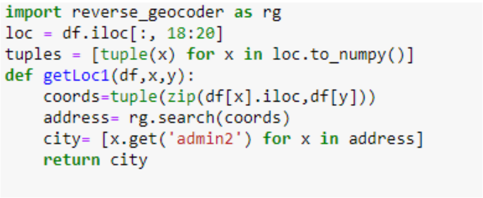
```python
executor = Parallel(n_jobs=2)
tasks= (delayed(getLoc)(loc,df) for loc in range(100))
locations= executor(tasks)
```

Figure 2: Sever based approach for reverse geocoding example

5.1.2 Offline Reverse Geocoder

The library used for offline reverse geocoding was reverse_geocoder [6] This library is an improved version of an existing library called reverse_geocode[7] which only uses Single-threaded K-D tree to find the name of the city Yu et al. (2021).

Reverse_geocoder uses a parallel implementation of K-D trees which are efficient for large inputs. K-D tree is populated with cities that have population larger than 1000 inhabitants. The source of the data used for this library is GeoNames [8]. This library also accepts custom data source for geocoding.



```python
import reverse_geocoder as rg
loc = df.iloc[:, 18:20]
tuples = [tuple(x) for x in loc.to_numpy()]
def getLoc1(df,x,y):
    coords=tuple(zip(df[x].iloc,df[y]))
    address= rg.search(coords)
    city= [x.get('admin2') for x in address]
    return city
```

Figure 3: Offline geocoder example

5.2 Spatial Clustering Algorithms

Clustering has long been a core topic in data mining and machine learning, providing a vast number of techniques. The term ”cluster” is used to describe a collection of similar data points.

[5] https://joblib.readthedocs.io/en/latest/
[6] https://github.com/thampiman/reverse-geocoder
[7] https://pypi.org/project/reverse-geocode/1.0/
[8] http://download.geonames.org/export/dump/

5.2.1 K-means

In data mining, the K-means algorithm begins with an initial set of randomly chosen centroids that serve as initial centroids for each cluster, and then uses iterative (repetitive) computations to fine-tune the centroids' final positions. It halts creating and optimizing clusters when either:

- The clustering has been effective, and as a result, there has been no change to the values of the centroids.
- The defined number of iterations has been achieved.

In other words, the K-means algorithm identifies k number of centroids and then assigns each data point to the closest cluster while keeping the centroids as small as possible.

The "means" in K-means refers to the arithmetic mean of the data, or locating the centroid.

5.2.2 DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm proposed by Martin Ester, et al. in 1996 Ester et al. (1996). It is a non-parametric clustering approach that uses density to group together points that are close together (points with numerous nearby neighbors), while labeling as outliers points that locate alone in low-density regions (whose nearest neighbors are too far away). When it comes to clustering algorithms, DBSCAN is among the most popular and often referenced.

6 Evaluation

EDA

This section is about exploratory data analysis which is used to summarize and investigate dataset using visualization methods.

In the following experiments, Accident data from London is used. Furthermore, performance of three clustering algorithms is measured through silhouette score and inertia (Lower the better).

6.1 Experiment 1 (K-means k =20)

For the K-means the although k = 10 got best silhouette score but in inertia 10 was on a higher scale which was not good for clusters. The next best value of k we got from silhouette and inertia was 20. Following figure shows the clusters which were obtained from using the value of K = 20. The value of K resembles the ambulance stations required and since it has to be cost effective the value of k should be as low as possible.

6.2 Experiment 2 (K-means k=40)

In the second experiment, the next possible good value for k was 40 which could be seen on inertia and well as silhouette graphs. Following figure shows the clusters which were obtained using the value of k= 40.

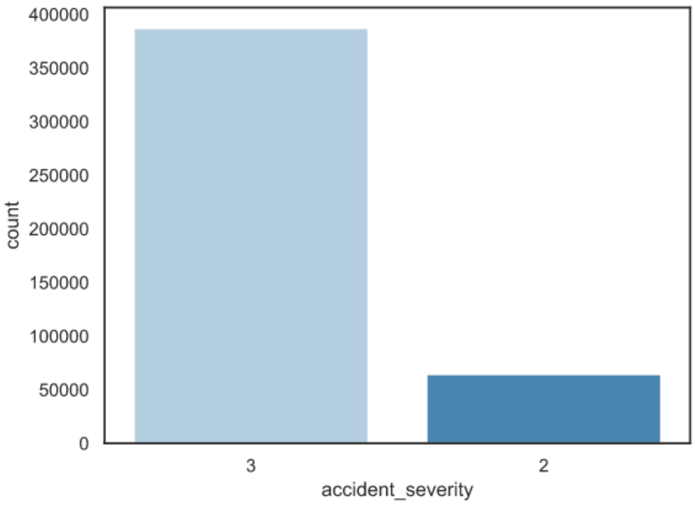


Figure 4: Above figure shows there were way more severe accidents recorded than less severe accidents

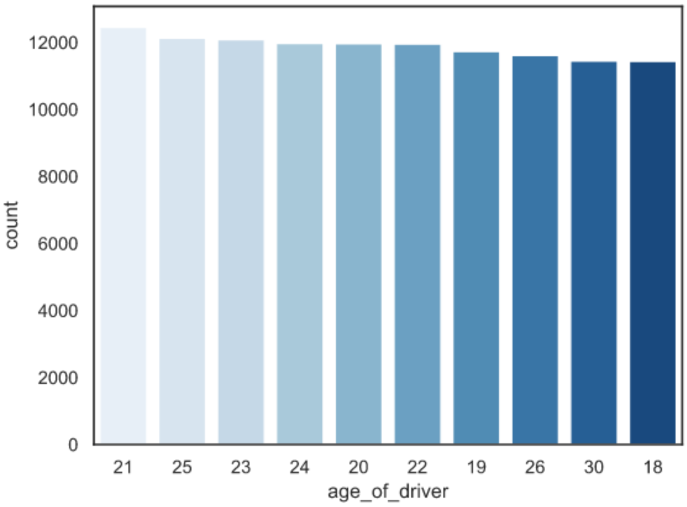


Figure 5: Above figure shows Drivers within the age of 18 to 26 were in the most accidents.

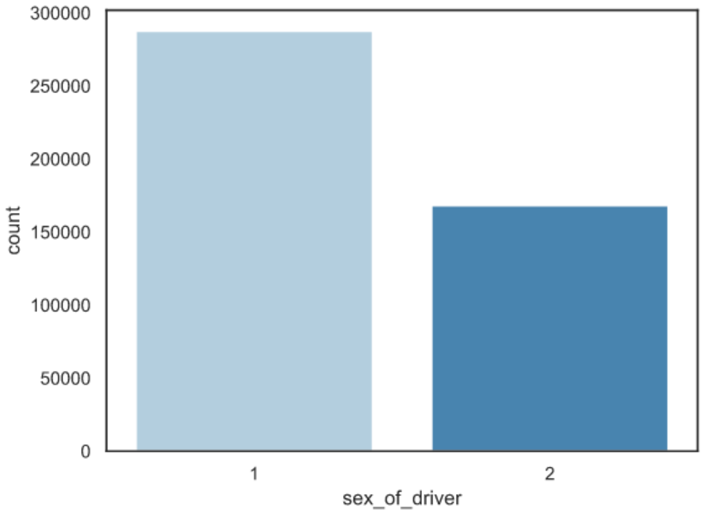


Figure 6: The number 1 represents men while the number 2 represents women. Statistically, more males than women were engaged in traffic crashes.

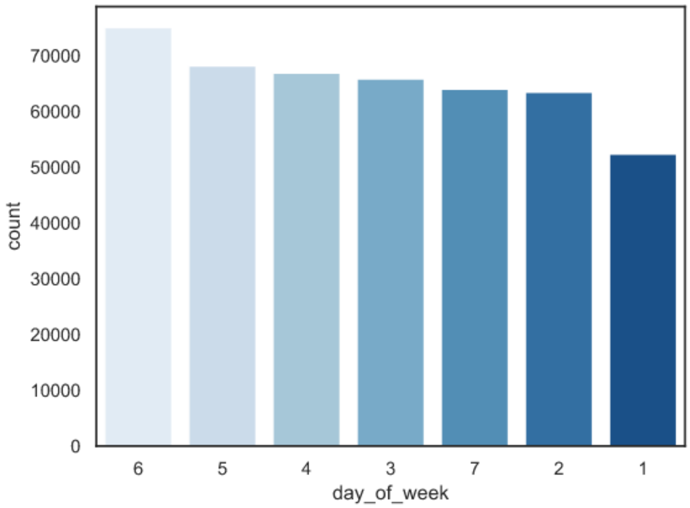


Figure 7: Based on the data shown above, it appears that Saturday is the day of the week with the highest accident rate.

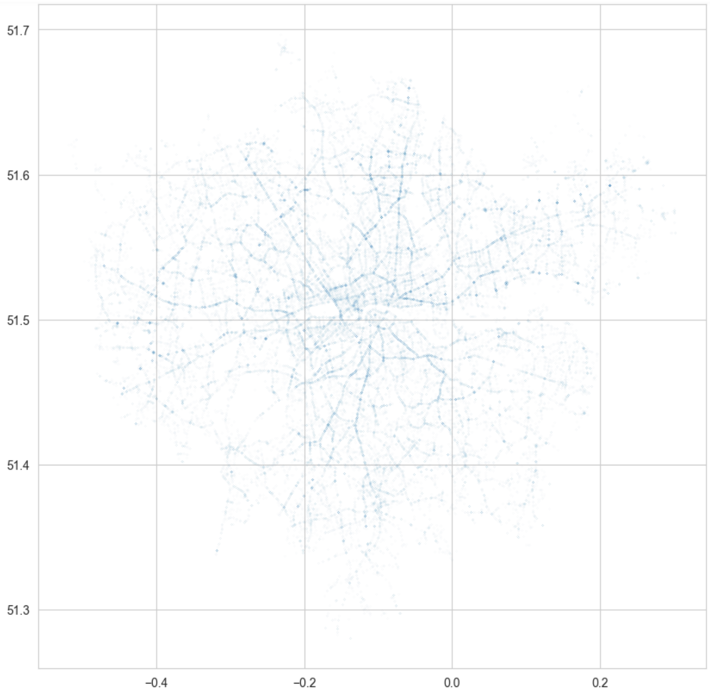


Figure 8: Data points of accidents in London for 2011-2019.

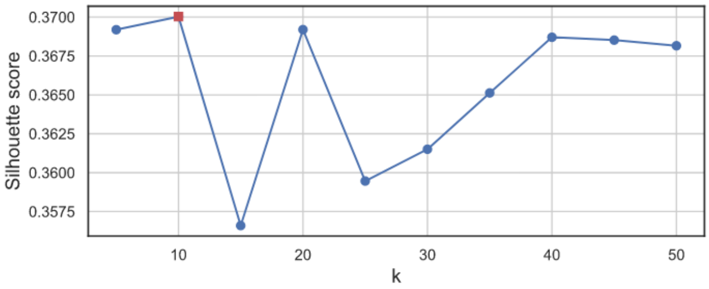


Figure 9: Silhouette score closer to one and greater than 0 indicates accurate clustering and sufficient distance between other clusters and itself. The K value which got the best silhouette score was 10.

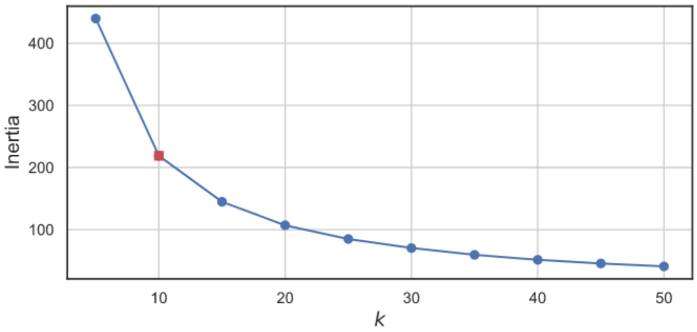


Figure 10: Inertia is the sum of the squared distances between each training instance and its closest centroid i.e. lower is better.

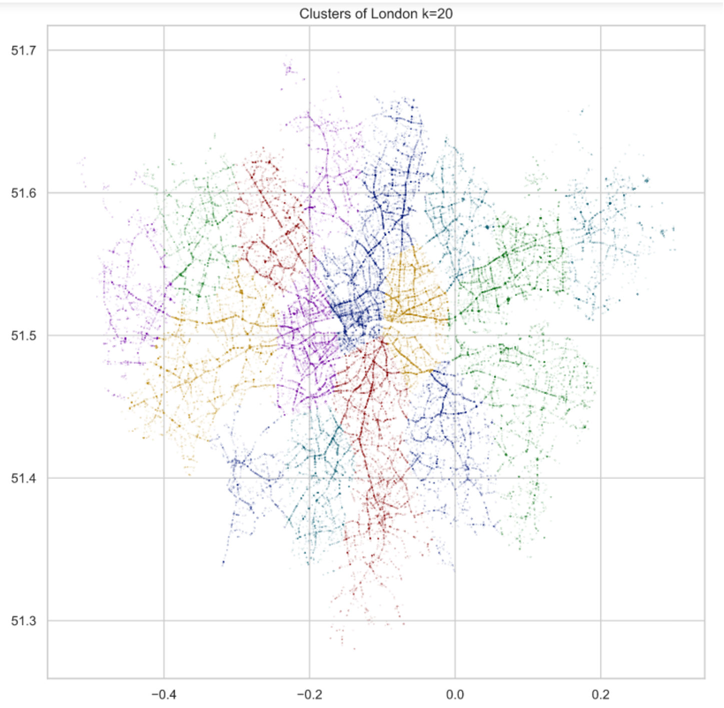


Figure 11: Clusters with the value of k= 20

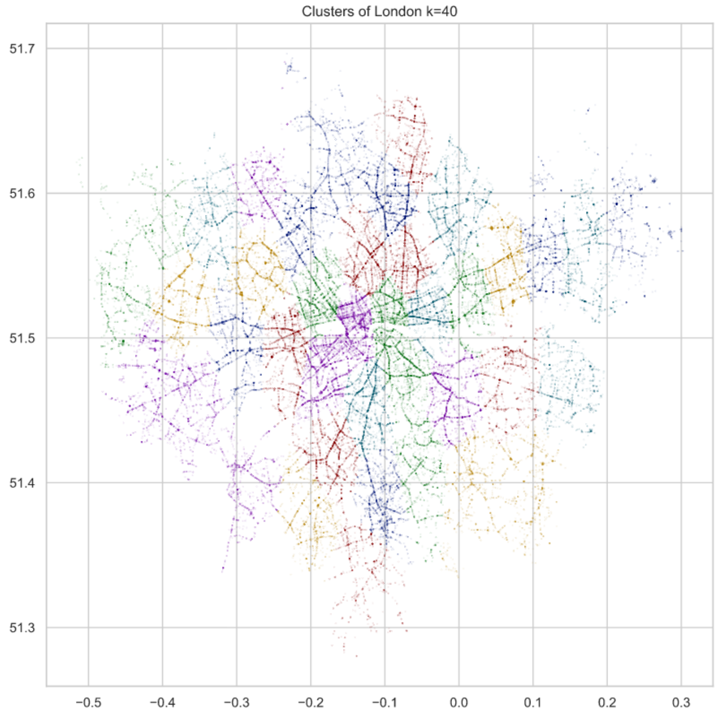
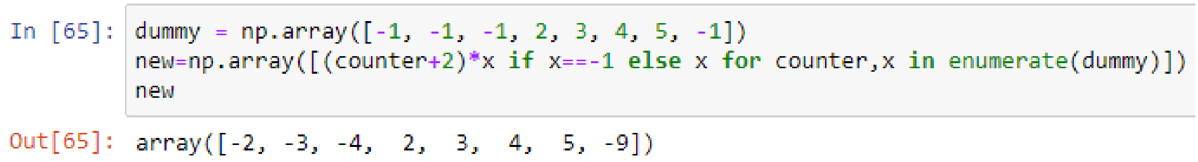


Figure 12: Clusters with the value of k= 40

6.3 Experiment 3 (DBSCAN)

Density-Based Spatial Clustering of Applications with Noise. The following figure shows the output from the implementation of DBSCAN on the London accidents dataframe.



```
In [65]: dummy = np.array([-1, -1, -1, 2, 3, 4, 5, -1])
         new=np.array([(counter+2)*x if x==-1 else x for counter,x in enumerate(dummy)])
         new

Out[65]: array([-2, -3, -4,  2,  3,  4,  5, -9])
```

Figure 13: For outliers, DBSCAN uses a value of -1. The author modified values to obtain silhouette scores by replacing the value of -1 with a decrement in value by -1. The example is shown in the figure above.

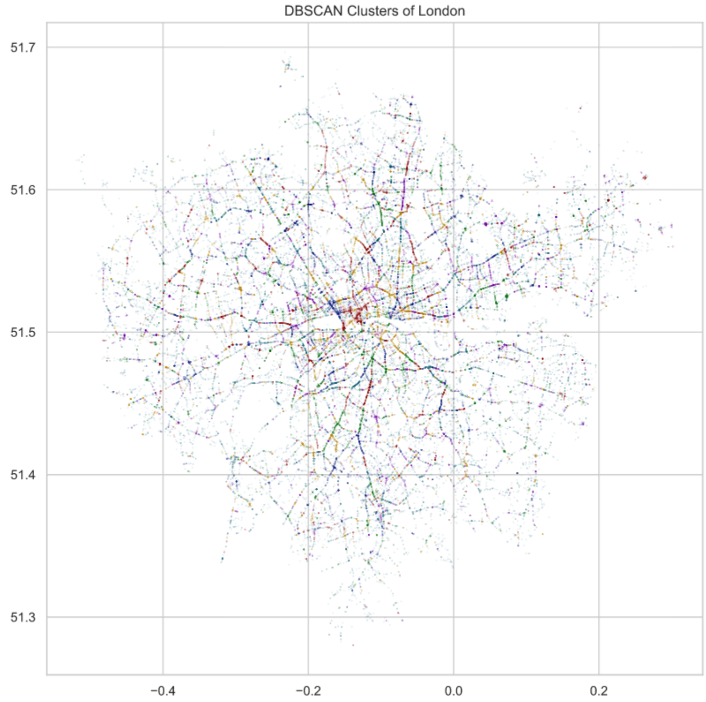


Figure 14: Clusters formed by DBSCAN

6.4 Discussion

6.4.1 Result K-means

The data in this section strongly suggests that k-means succeeded in this application of solving the research question. The locations of the ambulance stations are suggested by the cluster centers. This approach has a minor flaw in that there are two cluster centers on the map that are quite near to one another; however, this doesn't effect performance and can be easily fixed by combining the two centers into a single one and adding more ambulances than the other stations.

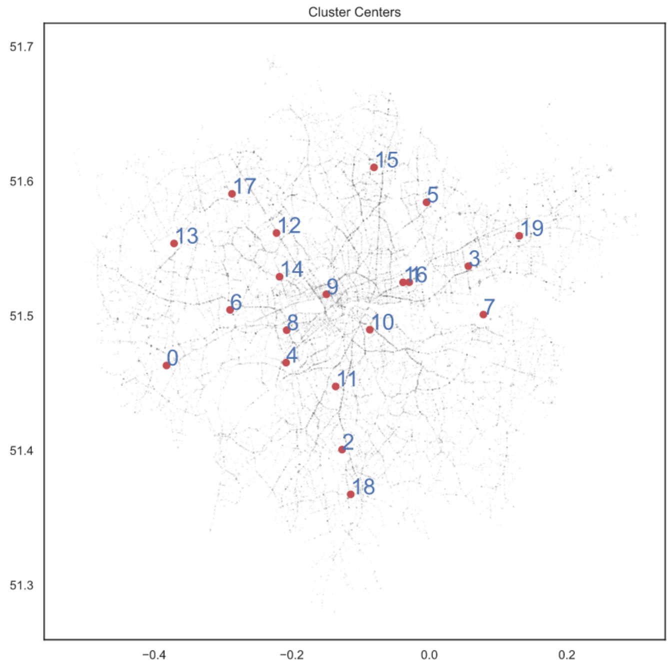


Figure 15: Cluster Centers Of K-means k=20

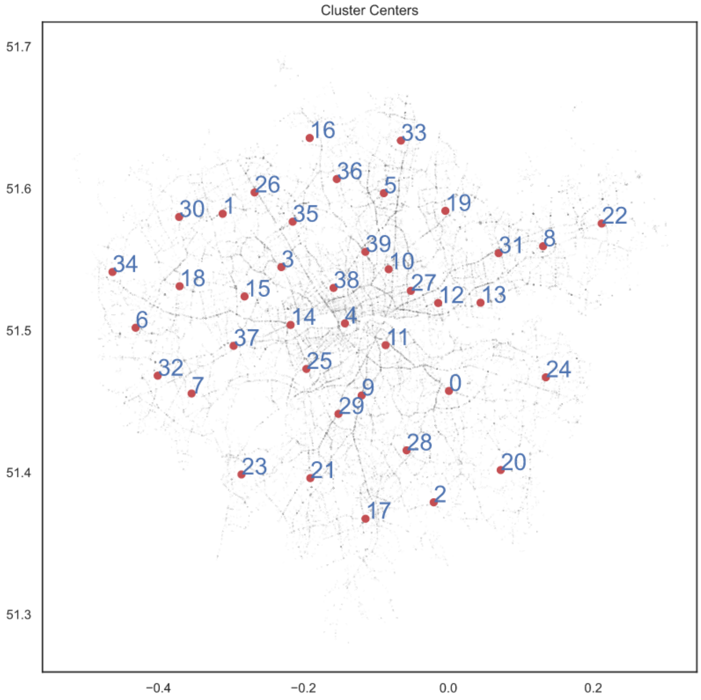
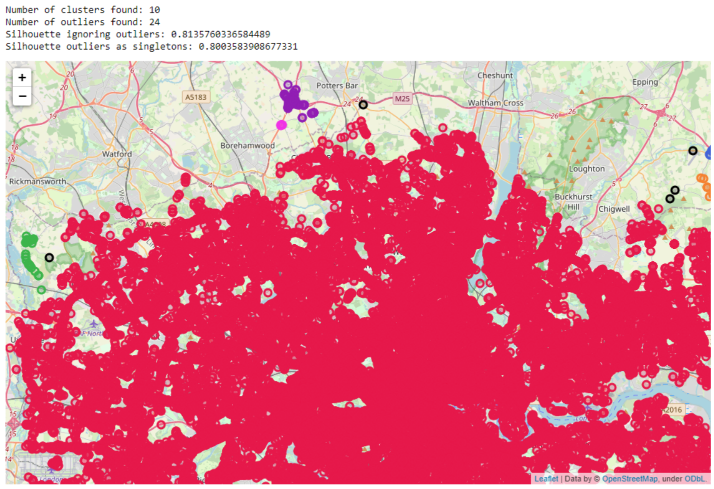


Figure 16: Cluster Centers Of K-means k=40

Centers in a cluster with k = 40 are evenly distributed, however it would be too expensive to set up 40 stations across London.

6.4.2 Result DBSCAN

DBSCAN produced unreadable results. The clusters were unbalanced even though the best value for the clusters in DBSCAN was 10 and the silhouette score was 0.8.



```
Number of clusters found: 10
Number of outliers found: 24
Silhouette ignoring outliers: 0.8135760336584489
Silhouette outliers as singletons: 0.8003583908677331
```

Figure 17: DBSCAN Cluster output

7 Conclusion and Future Work

With the increase in population, it is essential to improve the current accident recovery system. The objective of the research was to analyze accident data and find a way to optimize the current emergency recovery system by using clustering techniques. The research revealed offline reverse geocoding performed faster in comparison with server-based geocoding, and the best clustering technique for answering the research question was K-means and establishing 20 ambulance stations could be beneficial in improving the accident recovery rate in London. DBSCAN was very prone to noise, and surprisingly, it did not perform as expected. Most accidents occurred on Saturday, and mostly men were involved in road accidents in comparison to women. All the experiments were performed using only K-means and DBSCAN. Due to time constraints, the author was not able to test more clustering algorithms.

In future, more techniques could be applied to get a better solution to the problem consisting of HDBSCAN and fuzzy c-means available in QGIS. Constraints could be included, such as removing areas that are already within 10 minutes' driving distance of a hospital. This was part of the original research plan but was not implemented due to time constraints.

References

Aringhieri, R., Carello, G. and Morale, D. (2016). Supporting decision making to improve the performance of an italian emergency medical service, *Annals of operations research* **236**(1): 131–148.

Breen, N., Woods, J., Bury, G., Murphy, A. W. and Brazier, H. (2000). A national census of ambulance response times to emergency calls in ireland, *Emergency Medicine Journal* **17**(6): 392–395.

Cabral, E. L. d. S., Castro, W. R. S., Florentino, D. R. d. M., Viana, D. d. A., Costa Junior, J. F. d., Souza, R. P. d., Rêgo, A. C. M., Araújo-Filho, I. and Medeiros, A. C.

(2018). Response time in the emergency services. systematic review, *Acta cirurgica brasileira* **33**: 1110–1121.

Drummond, W. J. (1995). Address matching: Gis technology for mapping human activity patterns, *Journal of the American Planning Association* **61**(2): 240–251.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise., *kdd*, Vol. 96, pp. 226–231.

Lawner, B. J., Hirshon, J. M., Comer, A. C., Nable, J. V., Kelly, J., Alcorta, R. L., Pimentel, L., Tupe, C. L., Vanhoy, M. A. and Browne, B. J. (2016). The impact of a freestanding ed on a regional emergency medical services system, *The American journal of emergency medicine* **34**(8): 1342–1346.

Lee, J. (2009). Gis-based geocoding methods for area-based addresses and 3d addresses in urban areas, *Environment and Planning B: Planning and Design* **36**(1): 86–106.

Lee, J.-Y. and Kim, H.-Y. (2006). A geocoding method implemented for hierarchical areal addressing system in korea, *Spatial Information Research* **14**(4): 403–419.

Lin, A. X., Ho, A. F. W., Cheong, K. H., Li, Z., Cai, W., Chee, M. L., Ng, Y. Y., Xiao, X. and Ong, M. E. H. (2020). Leveraging machine learning techniques and engineering of multi-nature features for national daily regional ambulance demand prediction, *International journal of environmental research and public health* **17**(11): 4179.

McDonald, R. I. (2015). The effectiveness of conservation interventions to overcome the urban–environmental paradox, *Annals of the New York Academy of Sciences* **1355**(1): 1–14.

Nogueira, L., Pinto, L. and Silva, P. (2016). Reducing emergency medical service response time via the reallocation of ambulance bases, *Health care management science* **19**(1): 31–42.

Theeuwes, N., Houtum, G.-J. v. and Zhang, Y. (2021). Improving ambulance dispatching with machine learning and simulation, *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, pp. 302–318.

Vile, J. L., Gillard, J. W., Harper, P. R. and Knight, V. A. (2016). Time-dependent stochastic methods for managing and scheduling emergency medical services, *Operations Research for health care* **8**: 42–52.

Yu, J., Zhang, W. and Chen, R. (2021). Research on offline reverse geocoding algorithm based on k-d tree, *2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI)*, pp. 548–552.

Zandbergen, P. A. (2008). A comparison of address point, parcel and street geocoding techniques, *Computers, Environment and Urban Systems* **32**(3): 214–232.