Analysing Confidence of GPS Locations to discard Speeding Alerts using Machine Learning

09 Quartix

Charles Marsh, Charles Olive, Mu'izz Asif, Tomas Maranon Finbow

EMAT30005 Mathematical and Data Modelling

Department of Engineering Mathematics, University of Bristol

19th January 2021

Introduction

Vehicle tracking hardware, also known as black boxes, are widely used for commercial fleet tracking and for insurance purposes. Typically, they can monitor aspects of vehicles to ascertain a particular 'driving style' by analysis of real-time location and speed data, and thus can determine whether a driver was speeding at any point in their journey. The confidence of this decision is dependent on the performance of the GPS hardware, which is known to carry inherent inaccuracies due to a multitude of factors: satellite hardware, satellite coverage, weather, etc. It is therefore more cost effective to correct the measurements recorded by the vehicle tracking hardware than to upgrade the hardware both in the vehicles themselves and the satellites in orbit.

Provided vehicle tracking data by Quartix [2] concerning a stretch of road in North Yorkshire, we devise a method to identify whether they are on the A1(M) northbound, A1(M) southbound or A6055. From this classification we can determine if a vehicle is speeding. To start, we create additional features, including a regions feature which is particularly influential, to our data set. We combined two approaches to solve this problem: Principal Component Analysis coupled with K-Means in order to determine inaccurate data points and a Naive Bayes Classifier in order to classify which road each vehicle is on.

GPS Background

GPS consists of 3 main components: satellites, ground stations and receivers. The information is delivered by the satellites to the receivers by a signal, while the ground station is in charge of the corrections on the orbits of the satellites and the errors in measurements caused by external events. Satellites provide the accurate location of the receivers by a process called trilateration. Ideally, for accurate positioning 4 satellites or more should be connected.

The main problems that GPS accuracy may encounter are: signal arrival and time measurement, atmospheric effects, multipath effects, natural and artificial sources of interference, clock errors, geometric dilution of precision computation, relativity, etc [9]. By studying the terrain in which the road lies, we can infer that some of these factors are not causing any disruption in our data. Factors such as the multipath effect, clock errors, the signal arrival and time measurement or relativity are more difficult to trace from our data set. We can assume they add some error in our data. Others such as the road conditions or the the signals ability to reach the black box within the vehicle may affect the signal arrival of the GPS.

Data Preparation

Data Analysis

Data was provided by Quartix for journeys where there is a suspected speeding offence on a particular stretch of land. There are 3 major roads: A1(M) northbound, A1(M) southbound and A6055. The data for each journey is recorded as a set of points at different times on the journey, either side of the suspected speeding incident. For each point in the data set

the vehicle's speed, position, heading, satellite signal etc, was recorded as shown in Figure 1. For each journey there are typically between 2 and 12 points of data recorded, with a modal value of 6 and the time difference between each recorded point of almost always a minute. When the data is visualised a proportion of points are recorded not on any road and are clearly due to an error in the GPS hardware. What adds to the complexity of this problem is that we are unaware of the 'true' locations for any GPS coordinate, i.e. which road the vehicle was actually on.

SpeedAlertsId	AlertDateTime	AlertSpeed	AlertSpeedLimit	DateTime	Speed	Heading	WGS84Lat	WGS84Long	Satellites	SignalStrength
1000000000024610000	18/08/2020 10:09	142	112.65	18/08/2020 10:12	138	125	54.29853	-1.56226	13	44
1000000000022810000	09/07/2020 11:53	93	48.28	09/07/2020 11:48	95	315	54.29718	-1.55502	5	31
1000000000023240000	19/07/2020 13:02	82	64.37	19/07/2020 13:05	109	335	54.19357	-1.46578	13	41
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 1: *A subset of the data provided, which consists of 74261 points of 14726 journeys.*

We identified a subset of the features in the data set for use in classifying tasks to simplify our approaches to the problem. The features we consider are: 'SpeedAlertsId', 'WGS84Lat', 'WGS84Long','Speed', 'Heading', 'Satellites' and 'Signal Strength'. 'SpeedAlertsId' uniquely identifies each journey. 'WGS84Lat' and 'WGS84Long' both refer to the latitude and longitude points for different times recorded for each journey surrounding the alert. 'Speed and 'Heading' are the bearing (in degrees clockwise from North) and the speed (kmh^{-1}) respectively of the vehicle at each point in the journey. 'Satellites' refers to the number of satellites connected to the GPS hardware in each black box at each point on the journey, and 'SignalStrength' is the strength of the signal in decibels. Full list of definitions in Appendix A.

The number of satellites and signal strengths were also expected to be correlated: for more satellites connected, there should be a higher signal strength. The exception here is that data points with 0 satellites connected, where they also have a non-zero signal strength, look to be more consistently on roads than a low non-zero number of satellites connected (as shown in Figure 6 Appendix E).

We recognised that features that more closely related to the context can be inferred from the provided data using information that describes the roads themselves. The current longitude and latitude points have no frame of reference to the surrounding geography, so it would be relevant to include features that acknowledge this. Therefore for each point, the shortest perpendicular distance to each major road and the region that each point is in geographically, is extracted.

Perpendicular Distance Features

We can extract distance measures from every point in the data set to each road. We need to know the positioning and path of the road in reference to the points in the data set. Each road (A1(M) northbound, A1(M) southbound and the A6055) will be described as a set of straight line segments that approximates the centre line of the respective road. Each line segment is encoded as a set of latitude and longitude values that describe the endpoints of each line segment. On this scale, we can assume the longitude and latitude values vary linearly between the line segment endpoints. We manually identified approximations of the

centre line of each road with knowledge of the widths, approximately 14 metres for the northbound and southbound of the A1(M) and 8 metres for the A6055 ([3]).

On this small a scale, we can use the assumption that longitude and latitude are orthogonal axes that increase linearly, which allows us to treat any GPS coordinate as a vector position. We can use this to find the closest point in the line segment to any point in the data set. For every point in the data set, we find the shortest distance from each line segment (whether that is at the endpoints only, or within the line described) and take the minimum of these as the shortest distance from the road. This is then repeated for each of the three roads.

However, for calculating the distances between two points we can use the haversine distance (as implemented in [12]) which accounts for the curvature of the Earth, given that latitude metres per unit angle varies with the longitudinal angle. This is shown in Appendix B. For some endpoint positions, $\boldsymbol{e}_1, \boldsymbol{e}_2$ (without loss of generality), for a given line segment (where $\boldsymbol{e}_i = (e_{i,lat}, e_{i,long})$) and some other point $\boldsymbol{c} = (c_{lat}, c_{long})$, the closest point $\boldsymbol{c^*}$ on the line segment to $\boldsymbol{c}$ is

$$\boldsymbol{c^*} = \boldsymbol{e}_1 + (\boldsymbol{e}_2 - \boldsymbol{e}_1)t^*, \text{ where } t^* = \frac{c_j - e_{1,j}}{e_{2,j} - e_{1,j}}, j \in \{lat, long\}. \tag{1}$$

If $0 < t^* < 1$, then $\boldsymbol{c^*}$ lies within the line segment, else this value should be disregarded as it appears outside the line segment.

Combining these imply that the shortest distance from a point to a line segment is

$$\begin{cases} \min(\mathrm{H}(\boldsymbol{c}, \boldsymbol{c^*}), \mathrm{H}(\boldsymbol{c}, \boldsymbol{e}_1), \mathrm{H}(\boldsymbol{c}, \boldsymbol{e}_2)) & \text{if } 0 < t^* < 1 \\ \min(\mathrm{H}(\boldsymbol{c}, \boldsymbol{e}_1), \mathrm{H}(\boldsymbol{c}, \boldsymbol{e}_2)) & \text{otherwise} \end{cases}.$$

Derivations can be found in Appendix C.

Region Feature

To classify points into regions based on the geography, we divided the site into sections, assuming there is no distance between the A1(M) northbound and southbound roads, as it is comparably small to the distance and number of points located between the A6055 and the A1(M) northbound. In this case any points that are between the A1(M) roads will be classified into the road that they are closest to.

To identify the regions for each point, we use the closest line segment of each road to verify if it is east or not, compared to the closest line segment point. Greater than states the vehicle is to the east of the road, less than to the west. We incorporate the information about the widths of each road to then compare whether the distance from any point is greater than or less than half the width of each road to classify whether it is inside the road.

For some endpoint positions $\boldsymbol{e}_1, \boldsymbol{e}_2$ (without loss of generality) (where $\boldsymbol{e}_i = (e_{i,lat}, e_{i,long})$) for the closest line segment to a point $\boldsymbol{c} = (c_{lat}, c_{long})$ in the data set, then $\boldsymbol{c}$ is east of that line segment and thus the road if

$$e_{1,lat} + \frac{e_{2,lat} - e_{1,lat}}{e_{2,long} - e_{1,long}}(c_{long - e_{1,long}}) < c_{lat}. \tag{2}$$

The regions, from west to east, are defined as follows: West of A6055, inside the A6055, between the A6055 and A1(M) northbound, inside the A1(M) northbound, inside the A1(M) southbound, and east of the A1(M) southbound. These can be seen in Figure 7 Appendix E.

PCA and K-Means Clustering Approach

PCA

Principal component analysis (PCA) is a tool used for dimension reduction and also as an orthogonal transformation of correlated data into uncorrelated variables; this forms the principal components. Using dimension reduction as a tool for extensive data sets could highlight key relationships within our data. Using these methods may lead to some loss of information from the original data set, but can reveal patterns and relations that are not easily perceived. PCA creates components in which the original data is represented as a linear combination of all the features. The principal components are not assigned to any variable: they are a combination of all of them. This can be tracked with the explained variance vector collected for each principal component [5]. In this model we chose to use the calculated perpendicular distances from each road (A1(M) Northbound, A1(M) Southbound and A6055), the speed and the latitude. We only chose the latitude as the road stretches in a straight line from north to south and the longitude therefore gives little extra information. From the use of these features and applying PCA we obtained the following explained variance vector.

P.Components	PCA-1	PCA-2	PCA-3	PCA-4	PCA-5
Explained variance	0.461	0.219	0.164	0.111	0.043

Table 1: *Explained variance to principal components*

The explained variance shows the amount of information that each component has for the principal component analysis. In this model having three components covers over 84% of the features chosen from the data set. This is a reasonable amount to capture the most information without viewing the model in more dimensions than needed. In (see Figure 10 in Appendix E) the number of components has been plotted against the explained variance. This shows how as the number of components we choose for our PCA, the explained variance rises. For this model, using 3 principal components will ensure an accurate representation of the data.

K-Means

The following approach was to include k-means clustering into the principal component analysis. Figure 2 shows the number of clusters that can be used for the PCA against the within-cluster sum of squares (WCSS). WCSS shows the sum of squared distances from each point of the data set to the centroids of each respective cluster. To evaluate how many clusters are necessary for our model, by inspection, the elbow method is used [11]. This method shows how many clusters are necessary for the best representation of our model.

In this case on the second point there is a fold that indicates the elbow point. This shows that including more clusters would not give more information to our model. On other previous trials the elbow on the graph was more pronounced than this final model. Due to the reduction of features and the removal of the satellite and signal features, which were misleading, the elbow method showed a more subtle angle. Further demonstrations will show that having $K = 2$ clusters will provide the best results for this model in the results section.

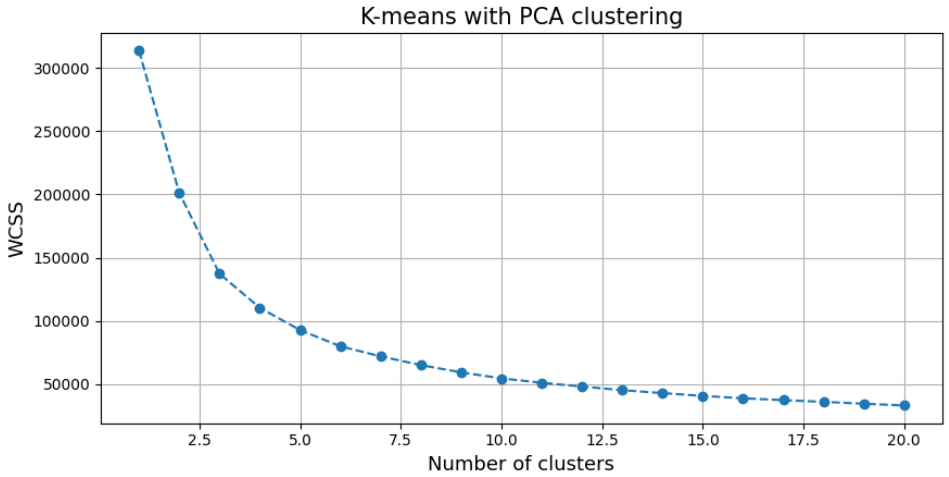

Figure 2: *Line graph showing the WCSS values for different numbers of clusters for the application of elbow method.*

Naive Bayes Approach

Patrick Pantel and Dekang Lin[13] implement a Naive Bayes Classifier to identifying spam emails: by looking at words within the text they could classify the likelihood of a email being spam. This method is analogous, where classifying vehicles to roads using the speeding data set we have is dependant on features from the data. From this method we can derive our likelihood of correct classification.

In machine learning a classifier is a form of supervised learning, using labelled training data to learn how to allocate said labels to new data. Based on a set of learnt features n, Bayesian classifiers assign the most likely class y to each new example x_n from the label with the highest probability [10]. More specifically, a Naive Bayes Gaussian Classifier finds mean μ and variance σ for each feature using the maximum likelihood through assuming each likelihood follows a Gaussian distribution (see Appendix D). Any new testing point has a probability of belonging to each label based on their variation from each mean and the nature of each features Gaussian distribution. We can bring together all the information we now have on the vehicles' speeds and our new distance and regions features to automate the decision on which road the vehicle is most likely driving on.

For the classifier we used 2 different sets of features, considering the speed of the vehicle, its distance from each of the roads with and without the regions feature.

Generated Training Data

In order to train our Naive Bayes classifier, we need a training data set where the classification is already known. As manually classifying data points to use for training would make our model biased, we created a training data set using reasonable assumptions.

Firstly, we needed to know how much traffic flow would be on each road so that we could accurately predict the probability a vehicle would be on a given road. Our first assumption is that all vehicles are cars as the majority of traffic is cars and speeds can vary drastically depending on the type of vehicle. Looking at data from the Department for Transport (DfT) we found that the daily traffic flow across the stretch of A1(M) was 34,859 cars based on an automatic count in 2019 [8]. This count does not account for cars joining and leaving the motorway at junction 50 and 51 so we have assumed that at each junction as many cars join as leave the A1(M). For the A6055 the DfT find the daily traffic flow to be between 3356 and 4737 based on an estimation from traffic in surrounding areas and an estimation of increased traffic since a manual count of 4549 in 2014, respectively [8]. The large difference in the estimation of traffic flow may be due to a roundabout joining the A6055 near the Leeming Bar Service, although it is more likely due to the estimate of 3356 being inaccurate. Therefore, we have taken the traffic flow to be 4737 cars per day on the A6055 and assumed the flow of traffic to be constant down the A1(M) and A6055 for these stretches of road.

To predict speeds, we first assumed that the distribution of speeds would be a Normal distribution for simplicity. To get the mean and standard deviation for the Normal distribution we looked at a statistical release regarding speeds of vehicles traveling on roads with different speed limits. The data in this article assumes all traffic is free flowing: speeds that are observed where there are no external factors that might restrict driver behaviour (e.g. junctions, hills, sharp bends, speed cameras)[7]. From here we find that the Normal Distribution is N(68, 9.785) and N(50, 8.6) for 70mph (A1(M)) and 60mph (A6055) speed limits based off the average speed and percentage of people above the speed limit[7].

For the distance from each road we assumed the distribution to be normal again for simplicity. To calculate the mean of the distribution we found the average distance between each road using our mapped points. This assumes that all the roads are equally distanced for the stretch of road we are classifying. Using some test data on GPS accuracy in mkm^{-1} we find the average error for all GPS from the true value is $42.37 mkm^{-1}$[6]. Assuming our data looks at a point every minute we can calculate the standard deviation by multiplying the average error per km with the distance we expect them to travel in one minute, giving a standard deviation of 77.28m for 70mph (A1(M)) and 56.82m for 60mph (A6055) respectively.

Results

PCA and K-Means Clustering

The graphical combination of both methods gives an accurate visual representation of the PCA and the K-Means clustering results. Figure 3 shows the 3 principal components, the first component as the x axis, the second component as the y axis. The third component does not contribute much for the representation of the PCA, as shown in table 1.

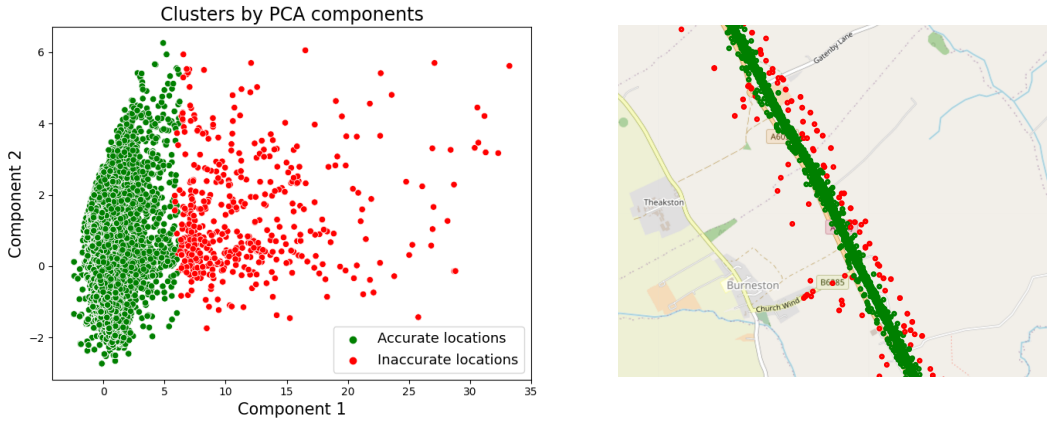

Figure 3: *PCA with K-means clustering in the component space (left), PCA and K-means classified data points on a map of road (right).*

From Figure 3 of the combination of PCA and K-means we can see the green cluster covers the majority of the data points which are concentrated close to the origin. The red points show a sparser cluster in comparison to the green. The first interpretation is that the more separated points are the ones that are far apart from the roads as the feature set is composed mostly by distances and locations. The data and colour are consistent throughout Figure 3, displaying them in the component space and on a map. The map plot confirms that the red cluster is the inaccurate set of data points and the converse with the green data points lie within the roads.

Naive Bayes

Establishing how the model was trained and tested, the two approaches (including and excluding the regions feature) is compared. Following this the speeding data set is classified to both understand how many points were assigned to each road and how confident the model is it chose the correct road per data point.

Classifier Accuracy on Test Data Set

The classifier was tested using a set of 1320 manually labelled data points, from visually checking whole journeys to see what road they should be assigned to. Against this test data set, the classifier's success rate of the model including the regions feature was 73.6%

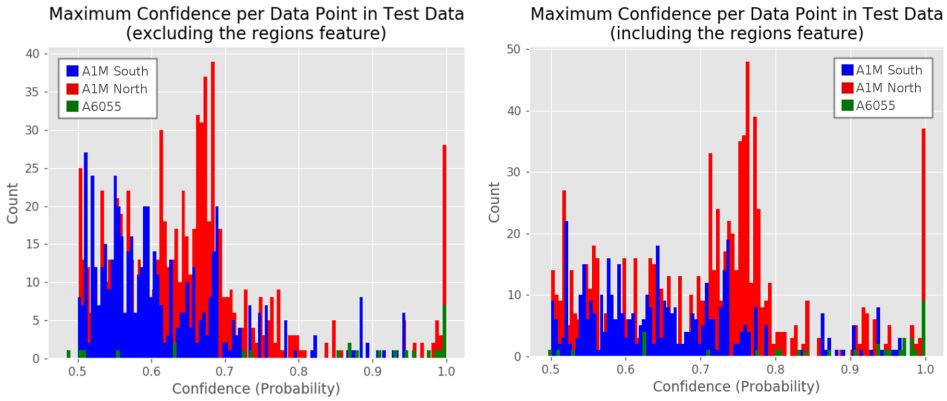

Figure 4: *Histograms showing the count of the number of data points with classification confidence for each data point in the test data set (left with additional regions feature, right without).*

and excluding this feature was 68.2%. This shows the regions feature does improve the accuracy of classification, so should be included. Figure 4 shows a distinct difference in the probabilities for the points between 0.5 and 0.7 across both models. This overall increase in confidence when using regions directly relates to the increase in test data accuracy; this model is correctly identifying which road each data point assigns to. The decrease in blue bars specifically near 0.5 shows the northbound motorway assigned points are more confident. This is key as they will lie close to the A6055. Better classification of points between the motorway and A6055 therefore is being achieved with regions, but to be sure of this both models will be tested against the real speeding data.

Classification of Speeding Data Set

Assessing the classification of the actual speeding data using these models gives us insight into how many speeding incidents are on each road. Figure 5 shows the probability of the chosen classification label for that point in the real speeding data set; one that has low probability has been weakly classified and is not as reliable as one with higher probability.

The peaks close to confidence 1 relate to points that lie directly on any road and have very similar speeds, easily classified as being on said road. The regions feature addition to the model does not affect this, as in this scenario it adds relatively little information. On average, points classified to the A6055 have a higher confidence than that of the other two roads, clear from the green bars near confidence 1. Points that lie on the A1(M) northbound and southbound roads have similar speed and distance measures, giving less confidence in their classification (0.5-0.7). Here, the 'Heading' feature differentiates these to the correct motorway side. In the case between the A1(M) northbound and the A6055, speeds that fit the motorway speed better (110kmh^{-1} for example) are more likely to be classified into the motorway, even if the vehicle is close to the A6055. These points, where simulated without the regions feature are classified into the A6055, are now part of the other two roads.

With the classifier including the regions feature, the confidence overall increases. The modal peak of A1(M) northbound with regions feature are at 0.77 confidence compared to the model without the regions feature where the mode is 0.69 and are approximately 1000

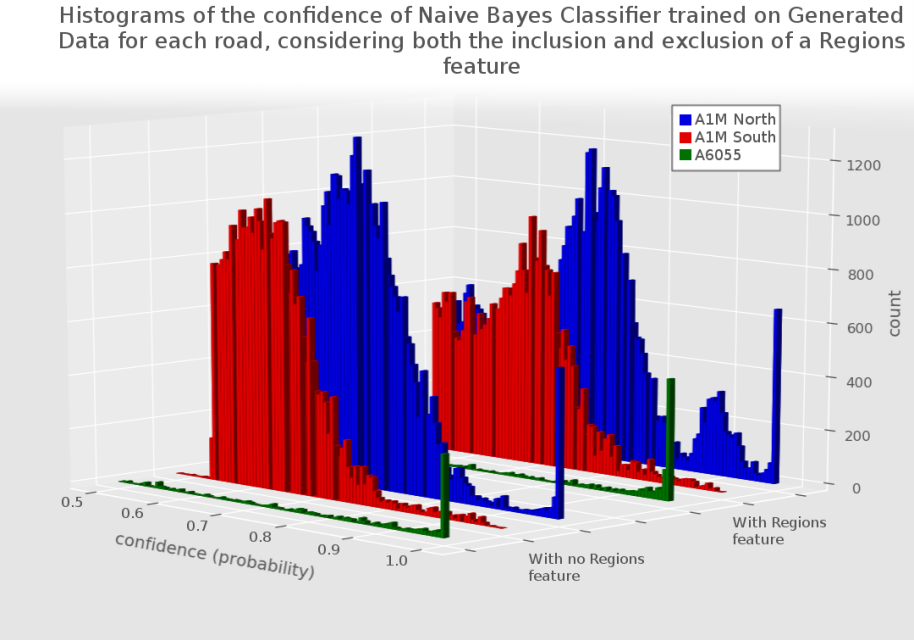

Figure 5: *Histograms showing the for the classified data points confidence. Plotting the 3 roads separately for both models (with and without regions feature).*

points less in height. All confidence values spread almost entirely over the 50% to 100% range, with only around 50 data points being less than this. These points are our very uncertain points, ones that indicate the information we have on that journey may be very inaccurate. Example map plots of this data can be found in Figure 8 Appendix E.

Discussion

The correlation matrix including the added features looks to be more favourable than that of the original features only as seen in Appendix E Figure 11. Between the original features, no significant correlations exist other than those that are contextually related. There exists significantly higher correlation between the inferred features, derived from the original data and geographical information. Between the other original features of 'Speed', 'Satellites', 'Signal Strength', there persists minimal correlation with the inferred features. However due to the inconsistency in data quality of the 'Satellites' and 'Signal Strength', where due to hardware error, zero satellites connected can still have non-zero signal strength (contextually, a contradiction) we can acknowledge it as a source of inaccuracy. 'Heading' appears to have an improved correlation with the inferred features, which was independent to these new features being extracted, and thus can support their validity.

The PCA method implies that some information may be lost when principal components are selected. We used the explained variance and from our results managed to cover 85% of the information from our features. This is favourable, as ideally for PCA more than 60% [1] of the information must be covered.

Other problems that we encountered were the unconsidered smaller roads around the main roads that we studied in our model. These roads have points that lie within them suggesting that we have data where vehicles are on these roads. We don't consider these other roads in our model due to the complexity of mapping all the smaller roads and the small portion of data assign to them. These smaller roads were assumed to be low confidence locations from these main roads.

The assumptions in calculating distances from each road are a significant source of inaccuracy in the feature extraction. However the small scale that these are operating on means that any error can be considered negligible. This seems to bear little significance as the region classification is mostly correct. This inaccuracy affects both approaches.

The Naive Bayes classifier method returns the probability of the allocated label for each data point. This allows us to adjust the confidence threshold of speeding alerts to discard false positives.

A change that may give more insight into this problem would be assessing each journey as a whole. Additional features such as the means, minima and maxima of the speed and distance measures could be considered for each journey. However this would be hard to implement, as the number of points per journey in the speeding alerts data set vary which would have to be accounted for. In addition, modelling training data for a journey would involve making further complex assumptions as data points within each journey are related. Finally, there are other features that we could have considered that may help improve our classification and confidences. The time of day was a feature of the data we did not consider at all and could have included in the classifier. Vehicle speeds vary throughout the day [7]. However, we did not have enough data to model this in our generated data; this could be explored more in the future.

Conclusion

In terms of generalising this feature extraction to any other site, the line segments and region classes would need to be automated (e.g. via image segmentation of the maps) so that only the roads in the local vicinity are considered, given some GPS coordinate.

Ultimately the goal of this project is to correctly identify speeding incidents. The confidences drawn from the classifier directly relate to the probability that the models classification is true. Looking at the generated data confidence values, the majority of points exceed 50% confidence, with 52 points (less than 0.1% of total data points) below this. In contrast, at a threshold of 55%, 14612 points (approximately 20%) are lower (see Figure 9 Appendix A). The significant increase in discarded points for a 5% increase in the threshold may imply it is an unreliable threshold.

The union set of discarded points between the PCA and best classifier is 798 data points (746 from PCA analysis, 52 from low confidence classification). If we look at just the data points of the joint set (where no duplicates across the 2 sets of in 'SpeedAlertIds') we get a set of 746 data points which corresponds to 316 journeys (see Figure 12 in Appendix E). Any of these journeys should not trigger a speeding alert as they contain low confidence data and should be assumed a false negative.

References

[1] Principal component analysis. Available at https://www.researchgate.net/profile/Ehsan_Khedive/post/How_many_components_can_I_retrieve_in_principal_component_analysis/attachment/59d626f2c49f478072e9b1be/AS%3A272185124425729%401441905398541/download/Principal+Component+Analysis+SAS.pdf.

[2] Vehicle tracking - gps fleet tracking for businesses: Quartix vehicle tracking (uk). Available at https://www.quartix.com/en-gb/.

[3] Google earth distance measuring tool. Available at https://earth.google.com/web, 11 2020.

[4] Ethem Alpaydin. *Introduction to Machine Learning*. Lecture Notes in Mathematics. MIT Press, 2010.

[5] N. M. Drawil, H. M. Amar, and O. A. Basir. Gps localization accuracy classification: A context-based approach. *IEEE Transactions on Intelligent Transportation Systems*, 14(1):262–273, 2013.

[6] Fellrnr. Gps accuracy of garmin, polar, and other running watches. Available at https://fellrnr.com/wiki/GPS_Accuracy, 08 2018.

[7] Department for Transport. Statistical release. 06 2019.

[8] Department for Transport. Road traffic statistics. Available at https://roadtraffic.dft.gov.uk/#14/54.2189/-1.4990/basemap-countpoints, 2020.

[9] The free encyclopedia From Wikipedia. Error analysis for the global positioning system. Available at https://en.wikipedia.org/wiki/Error_analysis_for_the_Global_Positioning_System#:~:text=The%20position%20accuracy%20is%20primarily,satellite%20position%20and%20signal%20delay.&text=%2C<%20or%20approximately%2010%20nanoseconds%20for,error%20of%20about%203%20meters.&text=or%20about%2030%20centimeters., 2020.

[10] T.J. Watson Research Cente I. Rish. An empirical study of the naive bayes classifier. Technical report, 2001.

[11] Dmitriy Kavyazin. Principal component analysis and k-means clustering to visualize a high dimensional dataset. Available at https://medium.com/@dmitriy.kavyazin/principal-component-analysis-and-k-means-clustering-to-visualize-a-high-dimensional-dataset-577b2a7a5fe2, 02 2019.

[12] Mangesh Nichat. Landmark based shortest path detection by using a* algorithm and haversine formula. 04 2013.

[13] Patrick Pantel and Dekang Lin. Spamcop: A spam classification organization program. Available at https://www.aaai.

org/Papers/Workshops/1998/WS-98-05/WS98-05-017.pdf?fbclid=IwAR3KTIXyGlzb2SCCmsP11CMpbrGEqieJFlkksktWNfjY9RVmAvoYB4u_iWo, 10 2020.

[14] scikit-learn developers (BSD License). 1.9. naive bayes documentation. Available at https://scikit-learn.org/stable/modules/naive_bayes.html, 02 2019.

EDI Statement

One member (who has now left) had some uncertainty as to whether they would leave for their placement due to the pandemic. At the beginning of the project this was known and we collectively acknowledged that in the case that they would stay in the group till the end of the project, it was important that they had contributed as much as they would if they had stayed – i.e. we would not consider them to able to contribute less in the time they were with us. However, it was also important that we did not ignore the fact that they would potentially leave, so work was allocated that so that they would finish by the time they were due to leave. In the end, the individual did leave for their placement.

As everyone bar one was in Bristol, we attempted to have an in-person meeting with the absent person dialled in through video calling software. This was difficult as often the absent person had connection issues and did not feel they were able to contribute as much as if they were in person. They communicated their concerns and we collectively decided to hold all whole-group meetings online only so that everyone has an equal opportunity to contribute.

To ensure that everyone would have the ability to keep track of all aspects of the project work happening simultaneously, we would always have meeting notes accessible through Teams so that all contribution was transparent to allow anyone to raise concerns if need be. Every member also had equal access to project materials, GitHub was used to let the programming side of the project be transparent through the use of commit descriptions.

We also had multiple ways of communicating so set out some policies about the use of each channel. This was so that the relevant concern, if any, goes in the correct channel and none are missed. Teams was for file access and meetings, as well as making suggestions/raising concerns that were not time-dependant and the Facebook Messenger group chat was used for organising meetings as well as raising more time-dependant queries and concerns. All members had access to all the file storage locations and communication methods.

COVID Mitigation Statement

The effect of the pandemic and the changes in public policy on group members did have an impact on group output. Towards the start of the project, one group member was self-isolating and therefore could not have any in person meetings with group members they were working with on a particular aspect of the project.

Another member needed to travel to their home country due to the change in public policy on national lockdowns which would have affected their output, prior to a meeting where we were would present the latest findings or work we had completed.

A Data Set Feature Definitions

The 'SpeedAlertsId' uniquely identifies each journey. 'AlertDateTime' and 'AlertSpeed' are the date and time values and the speed (in kmh^{-1}) for each journey respectively, when the suspected speeding offence took place. Both relate to the potential speeding offence for the entire journey and are the same for all points on that journey. The 'AlertSpeedLimit' is the speed at which the speeding alert is generated for the road the GPS hardware estimates the vehicle is on (25% above the assumed speed limit for the road) and is also constant for the whole journey. 'WGS84Lat' and 'WGS84Long' both refer to the latitude and longitude points for different times recorded for each journey surrounding the alert. 'Speed and 'Heading' are the bearing (in degrees clockwise from North) and the speed (kmh^{-1}) respectively of the vehicle at each point in the journey. 'Satellites' refers to the number of satellites connected to the GPS hardware in each black box at each point on the journey, and 'SignalStrength' is the strength of the signal in decibels.

B Haversine equation

The haversine definition that we used for some points $\boldsymbol{p}_1, \boldsymbol{p}_2$ (where $\boldsymbol{p}_i = (p_{i,lat}, p_{i,long})$) is:

$$\mathrm{H}(\boldsymbol{p}_1, \boldsymbol{p}_2) = 2r \cdot \arcsin(\sqrt{\sin^2(\frac{p_{1,lat} - p_{2,lat}}{2}) + \cos(p_{1,lat})\cos(p_{2,lat})\sin^2(\frac{p_{1,long} - p_{2,long}}{2})}.$$

C Data feature equation derivation

Derivation from equations from 1 for t^*: From vector parameterisation of a line, $\boldsymbol{l}$, $\boldsymbol{l}(t) = \boldsymbol{e}_1 + (\boldsymbol{e}_2 - \boldsymbol{e}_1)t$. If the line $\boldsymbol{c^*}$ is perpendicular to $\boldsymbol{l}(t)$, it follows that

$$(\boldsymbol{c} - \boldsymbol{l}(t^*)) \cdot (\boldsymbol{e}_2 - \boldsymbol{e}_1) = 0.$$

When expanded,

$$\begin{pmatrix} (c_{lat} - e_{1,lat} - (e_{2,lat} - e_{1,lat})t^*)(e_{2,lat} - e_{1,lat}) \\ (c_long - e_{1,long} - (e_{2,long} - e_{1,long})t^*)(e_{2,long} - e_{1,long}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Solving for t^*, the parameter of that line that is the closest to $\boldsymbol{c}$,

$$t^* = \frac{c_j - e_{1,j}}{e_{2,j} - e_{1,j}}, j \in \{lat, long\}.$$

Derivation for expression (in equation 2):

We convert the parameterised equation of the line segment, $\boldsymbol{l}(t)$ into a cartesian representation, where x and y are the latitude and longitude values respectively.

$$\boldsymbol{l}(t) = \begin{pmatrix} e_{1,lat} + (e_{2,lat} - e_{1,lat})t \\ e_{1,long} + (e_{2,long} - e_{1,long})t \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}.$$

Eliminating t, we can express x as

$$x = e_{1,lat} + (e_{2,lat} - e_{1,lat})\frac{y - e_{1,long}}{e_{2,long} - e_{1,long}}.$$

Therefore, if a point has a higher x value, it is more easterly.

D Naive Bayes equations

For finding the probability of each example given a certain feature

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{3}$$

where each features likelihood is assumed to be Gaussian. For finding the best label's probability/confidence

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y) \tag{4}$$

E Graph Appendix

Figure 6: *Comparison of the heat-map diagrams for data points with zero satellites connected (above), and between one and ten satellites connected inclusive (below).*

Figure 7: *Diagram to show the boundaries for region class to add as a feature for each point.*

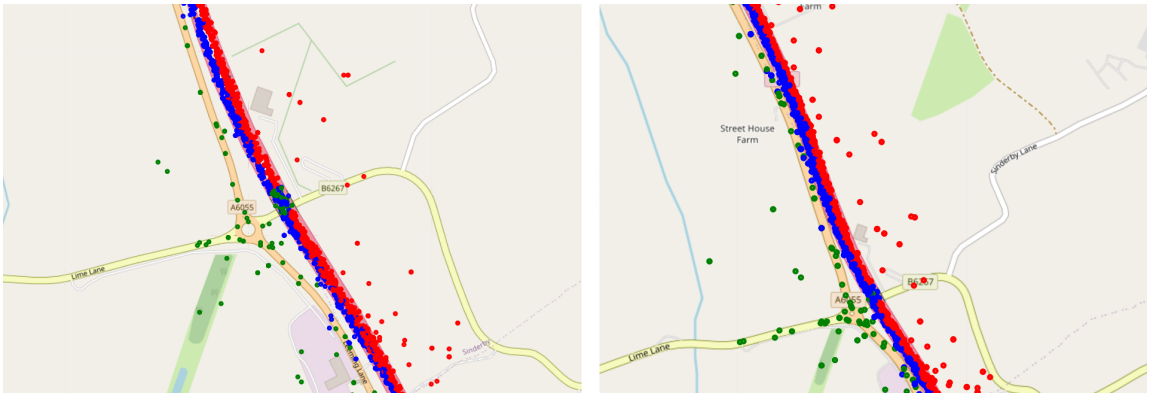

Figure 8: *Classifier trained on generated data, showing points near Lime Lane Roundabout. Shows classifier output both without regions feature (left) and with regions feature (right).*

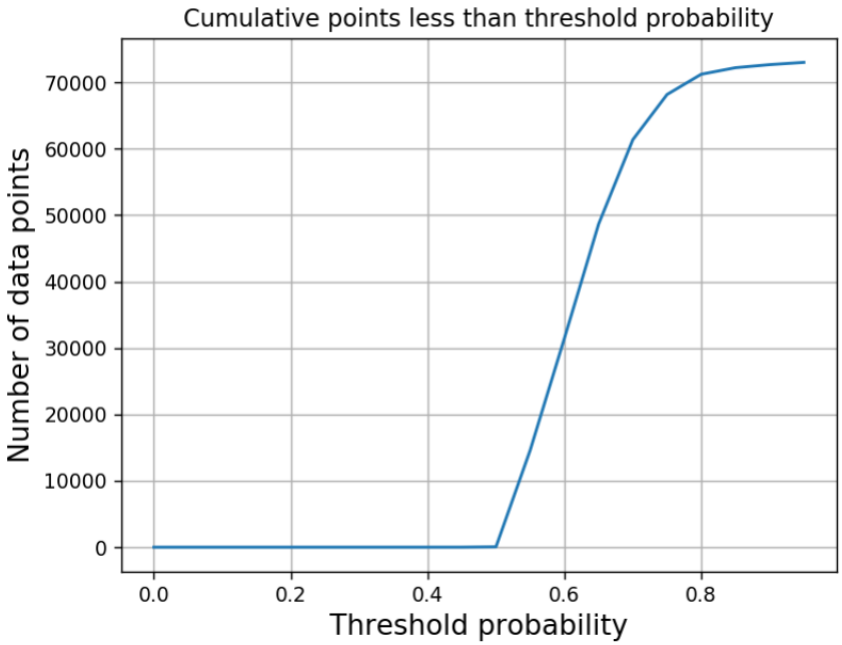

Figure 9: *Data points from the Generated Data with Regions classifier which have greater probability than x for their maximum chosen confidence. Large change (or 'Elbow') point at 0.5 dictated change from guaranteed inaccurate data points to accurate data points.*

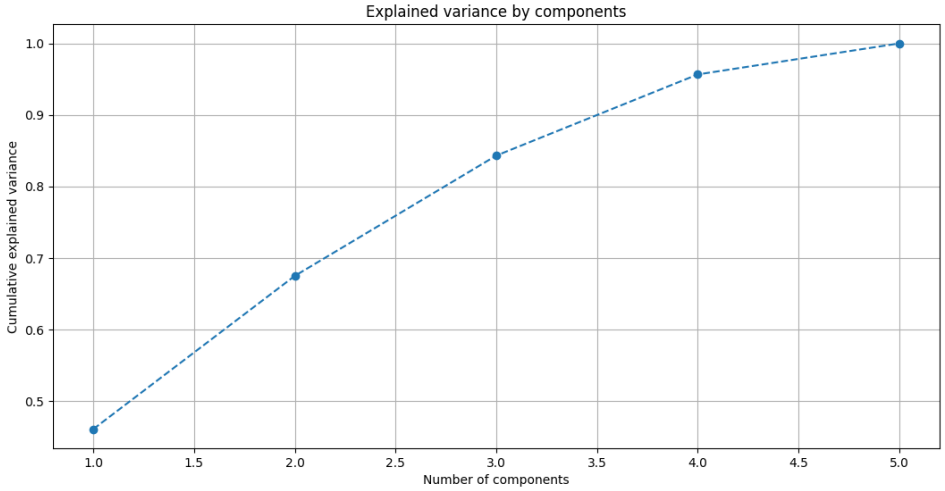

Figure 10: *Line graph to show the explained variance(amount of information) by number of components.*

Figure 11: *Correlation matrix of the features in the original dataset (left), compared to the extracted features and other relevant features (right).*

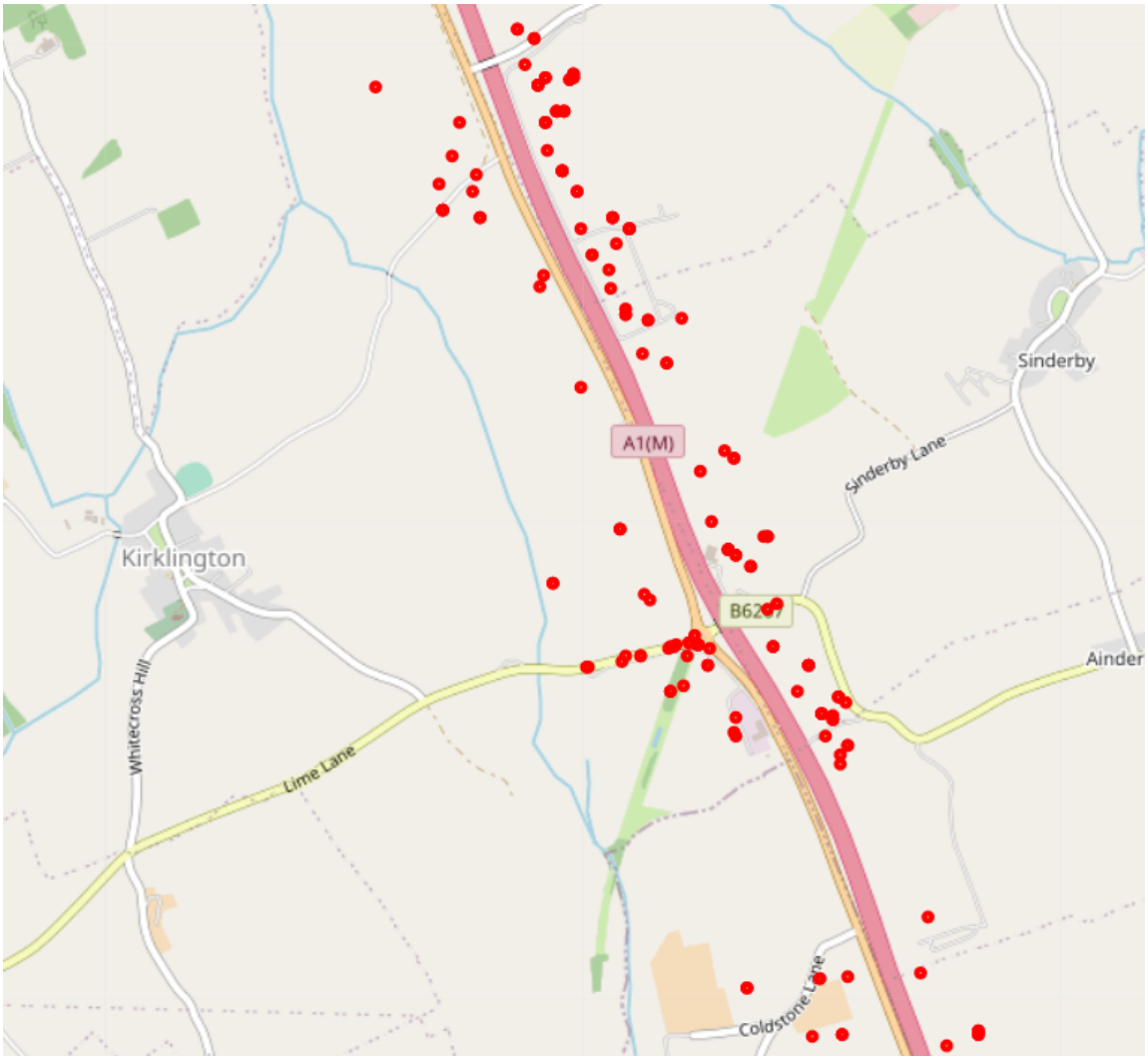

Figure 12: *Data points plotted on a road map(near Lime Lane Roundabout)that should be removed as classified by PCA or Naive Bayes.*