Impact of Data Quality and Target Representation on Predictions for Urban Bus Networks

Thilo Reich, Marcin Budka
Department of Computing and Informatics
Bournemouth University, Poole, UK
Email: treich@bournemouth.ac.uk

David Hulbert
Passenger Technology Group Ltd.
Bournemouth, UK

***Abstract*—Passengers of urban bus networks often rely on forecasts of Estimated Times of Arrival (ETA) and live-vehicle movements to plan their journeys. ETA predictions are unreliable due to the lack of good quality historical data, while 'live' positions in mobile apps suffer from delays in data transmission. This study uses deep neural networks to predict the next position of a bus under various vehicle-location data-quality regimes. Additionally, we assess the effect of the target representation in the prediction problem by encoding it either as unconstrained geographical coordinates, progress along known trajectory or ETA at the next two stops. We demonstrate that without data cleaning, model predictions give false confidence if mean errors are used, highlighting the importance of a holistic assessment of the results. We show that target representation affects the prediction accuracy, by constraining the prediction space. The literature is vague about quality issues in public transport data. Here we show that noisy data is a problem and discuss simple but effective approaches to address these issues. Research generally only focuses on a single method of target representation. Therefore, comparing several methods is a useful addition to the literature. This gives insight into the value of addressing data quality issues in urban transport data to enable better predictions and improve the passenger experience. We show that 'rephrasing' the prediction problem by changing the target representation can yield massively improved predictions. Our findings enable researchers using deep learning approaches in public transport to make more informed decisions about essential data cleaning steps and problem representation for improved results.**

***Index Terms*—Public transport, ETA prediction, Traffic analysis, Modeling and prediction, Machine learning, Deep learning**

I. Introduction

Bus passengers increasingly rely on Real-Time Passenger Information (RTPI) systems at bus-stops, online and in mobile apps. Current RTPI systems attempt to account for deviations from the timetable but are often unreliable [1]. This affects the convenience of bus passengers and is reflected in customer surveys as the most frequently requested area of improvement [2]. This highlights the importance of accurate Estimated Time of Arrival (ETA) predictions to improve the customer experience [3] and increase public transport usage.

Many cities suffer from severe congestion by an increasing number of cars [4], making travelling a challenge. In a recent report, it was estimated that in the UK travellers spent 10% of their driving time in gridlock costing the economy £38 billion [5]. The same report ranked Bournemouth as the 8th most congested city in the UK. Prospective studies suggest that the biggest environmental and societal impact can be achieved if the public is encouraged to change from private cars to public transport and thus reducing air pollution and congestion [6]. This was illustrated by a study suggesting that cancelling just 1% of daily commutes from specific neighbourhoods in the Boston area, can reduce the delays for all road users by as much as 18% [7].

To encourage such a shift it is important to address the passengers' desire for reliability. As delays in bus services are inevitable, it is crucial to keep the passengers informed. As many public transport apps give 'live' positions of vehicles, these are often used by passengers to decide when to leave to catch their bus without having to wait too long at a bus stop. However, due to latency of this information caused by delays in wireless-network infrastructure and passing through a number of 3rd party systems, this data is delayed, suggesting the vehicle is further away than it is in reality. In the Bournemouth area, for example, the latency of the internet-based 'live position' is approximately 30s and could be the difference between a passenger catching a bus or missing it. Therefore, a reliable short-horizon prediction to tackle this delay would undoubtedly be useful.

The infrastructure required to allow such predictions is already in place in the form of Automatic Vehicle Location (AVL) systems [8]. As AVL systems stream data continuously, in theory, they could be readily leveraged to develop better data-driven solutions. However, the AVL data suffers from serious quality issues. These include the lack of clear journey identification linkable to the timetable, artefacts such as gaps in recordings, falsely reported line numbers and directions. The biggest positive impact for passengers can be achieved by improving not only the delay seen in 'live' locations but also the ETA predictions at bus stops. To this end, this study uses one bus line from the city of Bournemouth (UK) as an example and addresses: (1) the data quality issues encountered, (2) their impact on prediction using Recurrent Neural Networks (RNN) and measures to overcome the identified issues, and (3) the impact of target representation on ETA prediction accuracy where we compare the accuracy of two types of output – the position in the next 40s, which is the equivalent to a next-step

978-1-7281-2547-3/20/$31.00 ©2020 IEEE

prediction based on the sample rate of our dataset and the arrival time at the next two bus stops.

II. Related work

Urban bus networks generate highly multidimensional data. This not only includes the geographic and temporal aspects but also data generated by several vehicles serving the same line on different timetables or directions. This large data source can easily be affected by quality issues. The importance of data quality has been highlighted in the literature in the context of bus travel for example for pattern analysis [9] but also to allow general improvements of public transport services [10]. Other authors have proposed methods to tackle these issues [11]. However, it is notable that few literature examples are directly addressing data quality. This problem should be much more prevalent considering the strive to include big data into urban-transport predictions from novel data sources especially via crowdsourcing [12; 13; 14]. The assumption that cleaner data will allow for better predictions is examined in this paper.

The second question is how to best represent prediction problems. Reducing the complexity of input data can have beneficial effects on prediction tasks [15]. This is generally applied to the input. Furthermore, the technique of representation learning suggests that there is a right way of posing a question to a machine learning algorithm [16]. More well-known examples highlighting the importance of target representation come from medical image classification, where algorithms have been found to use confounding clues such as cables visible in an image to make a prediction [17]. Therefore, an empirical approach will be used to compare the quality difference of three target-representations of similar prediction problems specific to public transport.

III. Methods

A. Data collection

The data used was collected from one of two bus-operators in the city of Bournemouth (UK). The vehicles transmit their position approximately every 40 s which is collected by the company providing the Electronic Ticketing Machines (ETMs) with the integrated AVL-system. Due to the involvement of several companies handling the data, only a limited amount of information is transmitted. The available data are:

- Timestamp
- Position (latitude and longitude)
- Line number
- Direction (outbound or inbound)

It became apparent that neither the direction nor the line numbers are reliable. The transmitted direction is often incorrect and so are line numbers when a vehicle changes its line during an operational run. This becomes evident when observing data identifying as one line but serving another as well as vehicles travelling in the opposite direction from their transmitted data. This suggests that although coordinates are updated continuously, the additional information is not always updated after a vehicle starts its journey. Based on this limited information it is typically not possible to match a vehicle to a timetable corresponding to the journey it is currently serving. A journey is a specific trip found in the timetable of a bus line e.g. the outbound 9 AM service 1. In contrast, a route pattern is the route as travelled on the road which can vary slightly for each journey for the same bus service. In the example of line 1 in Bournemouth, each line has several patterns which can include different start-points along the route, resulting in shorter overall journeys or slightly different routes – see Fig. 1 for examples. Matching a vehicle directly to a specific route pattern is not possible as no specific identifiers are transmitted.

Fig. (1) Map showing different route patterns associated with line 1 (Yellow Buses) in Bournemouth (UK). Overall this line has 12 more or less distinct patterns (4 inbound and 8 in the outbound direction). For clarity each shape was offset by 0.0005° northwards to prevent overlapping.

Therefore, a specific route-pattern was selected for proof of concept. This route-pattern is line 1 in Bournemouth in the outbound direction from the city-centre (Triangle) to the final destination (Christchurch). The reason for this choice is that the start-point for all inbound journeys is the same, making these journeys indistinguishable. The inbound journeys, however, do not always have the same destination, so the outbound direction was chosen to allow better identification of journeys.

B. Identification of individual journeys

As the data lacks an explicit indication about the progress of the journey (e.g. bus stops already visited) it is not self-evident when a journey ended and a subsequent journey started. An observation made was that between two timetabled journeys, the vehicle generally goes offline briefly. Thus, once it comes online again a gap in the recordings can be detected. A new journey was defined as a time gap of more than 15 min. If such a gap is detected it is assumed a new journey has started.

C. Representation of a journey as trajectory

All buses should follow a predefined route which can be represented as a trajectory. The trajectory is the distance a vehicle has travelled along a route over time. This means the trajectory will always be different for each journey. The transmitted coordinates are simply projected onto a route pattern by assuming that the closest point on the route to the current coordinates represents the position of the vehicle (see Section III-D for filtering approaches). These trajectories are used as one target representation as well as for benchmarking as described below. As the route is known, positions along the trajectories can be converted back to coordinates along the route (Fig. 2).

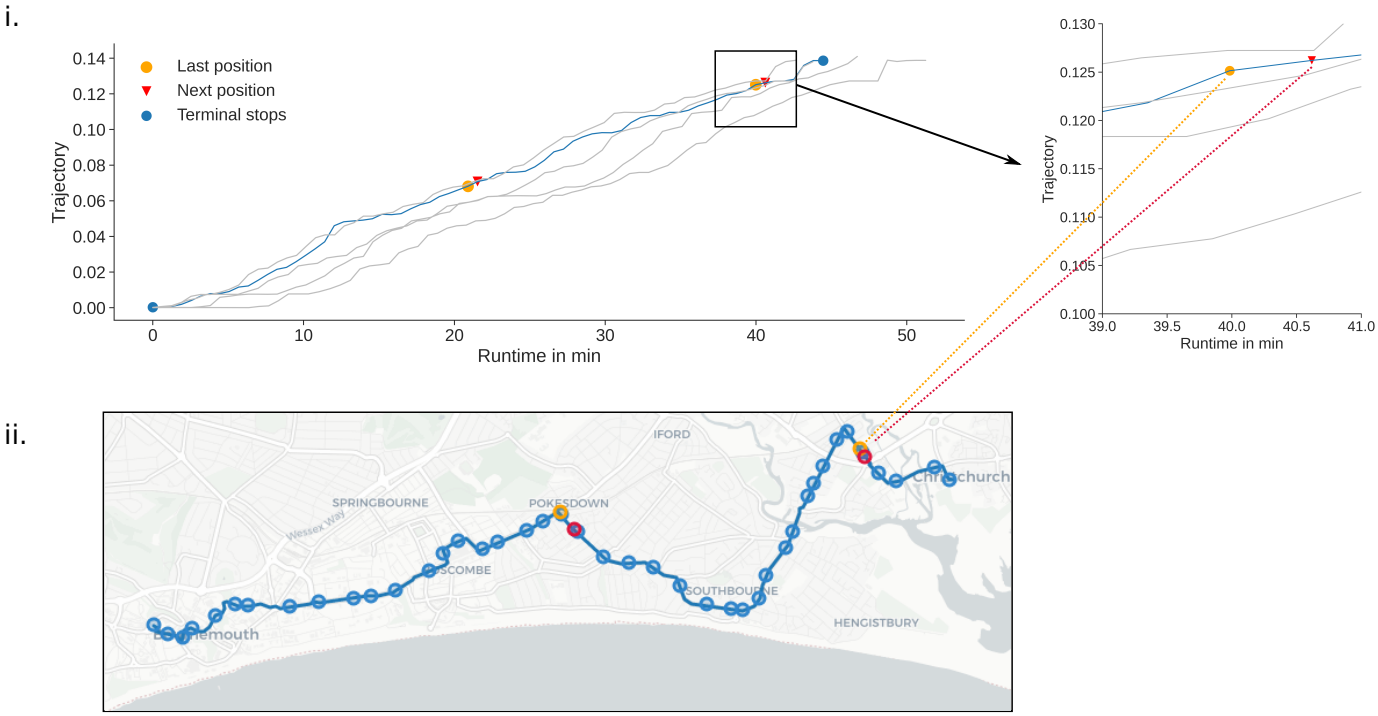


Fig. (2) **i.** The trajectory representation of several journeys, where the progress along the route is represented over time. The difference between several vehicles travelling on the same route is illustrated. As an example, one journey has been highlighted in blue with examples of the input position in yellow and the target position in red. **ii.** The route of the bus line with stops indicated as blue circles. The highlighted trajectory positions are shown as coloured circles on the route

D. Data pre-processing

As described previously the data suffered from quality issues. The problems encountered were misreported identifiers and positions resulting in physically impossible position-changes between recordings such as vehicles travelling over 60 mph. To combat these, several filtering procedures were applied within an ablation study, to clean the data and assess the influence these cleaning steps have on the final results (Table I). These data cleaning steps are referred to as 'sets' and correspond to the experiments described in Section IV:

- *Set 1.0 – minimal processing.* As the line number of a vehicle was unreliable, all vehicles identifying themselves as line 1 were selected. To ensure these are following the correct route, the journeys were filtered by excluding those with reported positions further than 2 x mean-distance from the route. Furthermore, any vehicle which appeared to travel faster than 62 mph (100 km/h) was also removed as this is legally and physically not possible within a city environment. This represents the dataset with minimal pre-possessing and thus has the most data-points. To ensure a fair comparison to the more heavily processed datasets, a randomly selected subset of 1476 journeys of this dataset was used.
- *Set 2.0 – filtering of direction.* As the direction was found to be reported incorrectly, the outbound direction was filtered by ensuring that each vehicle was within 100 m of the first outbound stop at the start of its journey. If this was not the case such journeys were removed.
- *Set 2.1 – removing repetitions from the end.* In practice, a vehicle will stop at the beginning and the end of the journey for operational reasons. Therefore, these positions will be repeated until the vehicle starts the next journey. These repetitions were removed from the end of the journey once the bus has reached the closest point to the final destination.
- *Set 2.2 – removing repetitions from the start.* Stationary repeats were removed from the start of each journey, assuming that a vehicle has started its journey once it has moved more than 10 m between recordings. This removes positions where the bus has arrived at the beginning of the route but is waiting for the timetabled journey start.
- *Set 3.0 – removing all repetitions.* The final set combines all the above-described filters and is, therefore, the most heavily processed dataset.

E. Benchmarks

The literature regarding ETA prediction in public transport often lacks comparative benchmarks, which makes it difficult

Set	Outbound only	End truncation	Start truncation
1.0			
2.0	X		
2.1	X	X	
2.2	X		X
3.0	X	X	X

TABLE (I) Ablation study setup.

to objectively compare different approaches. In other areas of machine learning, it has become the norm to use benchmarks and standard datasets. As there is no appropriate publicly available benchmark dataset available for urban public buses, this study uses benchmarks that can be easily implemented on any dataset. This allows other researchers to compare their solutions to this publication but also gives a threshold to assess any results against. The benchmarks are:

1) *Average speed.* This method uses the average speed of a vehicle since the start of its current journey. Thus, it does not reflect any short-term speed variations. The calculated speed is used to interpolate the position of the vehicle from the trajectory of its journey pattern for the next 40 s.
2) *Current speed.* This method uses the last three transmitted positions of a vehicle to calculate its current speed. The prediction is made by interpolating the position for the next 40 s from the journey trajectory. This method will account for temporary speed variations.
3) *ETA benchmarks.* To calculate the ETA benchmarks both speed-based methods are used to interpolate the arrival time at the next two stops for the ETA based benchmarks. For further details see Section III-F.

F. Target representation

To investigate differences in accuracy three different target representations were used. These all use the same data as input but represent the prediction target differently:

1) *Unconstrained coordinates.* The raw data of bus locations are affected by inaccuracies due to interference of the GPS signal. Therefore, the positions of vehicles are not always directly on the route. This represents the raw target where no pre-processing of the target was applied. The only constraint used was a bounding box framing the city. This approach predicts two normalised values representing coordinates within the bounding box.
2) *Trajectory.* The raw coordinates can be projected onto the route-pattern of a journey by simply using the closest point on the route as position once a journey is successfully matched to a route-pattern. This ensures that inaccuracies locating a vehicle off-route are removed. The route-matched positions can be turned into a trajectory by plotting the distance along the route over time as demonstrated in Fig. 2. In practice, this method predicts a number representing the progress along the trajectory with a max of 1, which is the final destination.
3) *ETA.* This approach predicts the arrival time at the next bus stop instead of the position of a vehicle. As the next stop could be very close to the vehicle, we predict the next two stops instead. The prediction itself is in seconds to the corresponding stop. As we aim to compare different target representations, to make the ETA predictions more comparable to the position based approaches, the error in seconds was translated into an approximate margin of error in meters based on the travel speed, assuming the bus travels at a constant speed from its current position to the two stops. The distance-based errors are approximations for comparison only.

G. Model training and evaluation

All models were trained on an Nvidia GeForce RTX 2060 GPU using the *fastai* library. The experimental setup was:

1) *Input features.* The features included were coordinates normalised to a bounding box, the bearing reported by the AVL system, the time-delta between consecutive recordings, the elapsed time from the start of the journey and time embeddings as described below. The input features were min-max normalised unless stated otherwise.
2) *Time embeddings.* The time information was split into its components to make it possible for the algorithms to learn seasonal patterns. To achieve this the timestamp was translated into minute of the day, hour of the day, day of the week, day of the month and month of the year. These were embedded in a multidimensional space as detailed in architecture description.
3) *Architecture.* Two neural network models were used with identical architecture (Fig. 3) except for the Recurrent Neural Network (RNN) module which was either a Gated Recurrent Unit (GRU) or a Long Short Term Memory (LSTM). The time embeddings were learned by the network in a multidimensional space. The dimensions were chosen as half of the possible number of values for each embedded variable. As an example, the hour of the day was embedded in 12 dimensions as the maximum number of hours is 24. These embeddings with a total of 52 dimensions were fed into a linear layer to reduce their dimensions back to the original number of time based features. The output of the linear layer was concatenated with the remaining input features and fed into either a GRU or LSTM layer followed sequentially by a 1D Batchnorm, a linear layer, a leaky ReLU, a second Batchnorm and a final linear layer. To ensure the outputs were bounded, a sigmoid was also applied.
4) *Hyper-parameters.* To allow for direct comparison between the models all training hyper-parameters were kept constant. It is appreciated that this might not in all cases yield the best performance but will illustrate the influence of the modifications made on the performance. The used variables were chosen through empirical exploration. Each model was trained for 50 epochs using the one-cycle policy [18] with a maximum learning rate of 10^{-3}. Networks with unconstrained coordinate

targets used the haversine distance between target and prediction as loss-function, while all other networks were trained using the Mean Average Error (MAE).

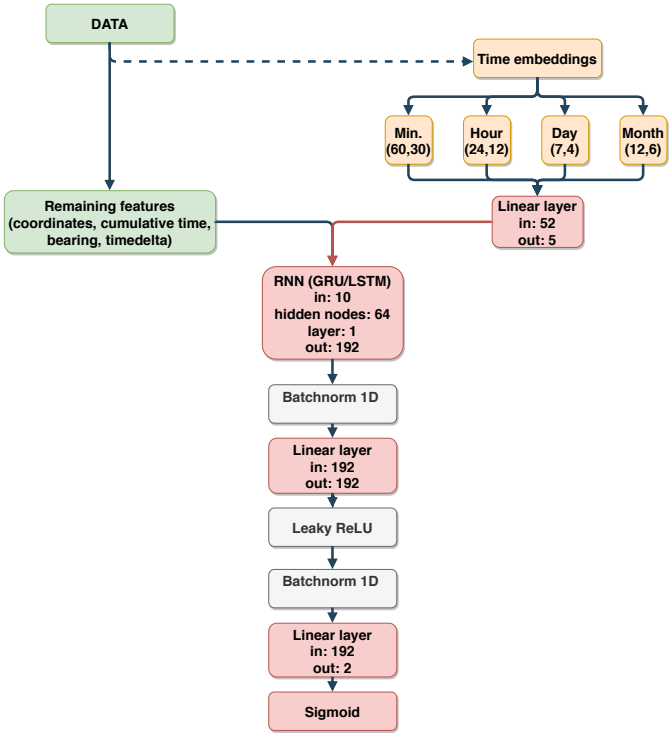


Fig. (3) Network architecture for the two RNN approaches – GRU or LSTM without any other changes to the network.

H. Evaluation

Predictions from approaches producing positional outputs including coordinate and trajectory-based predictions were converted to denormalised coordinates. Errors were assessed via the haversine distance between target and prediction.

As the ETA based approach does not give any location-based prediction, the error in meters was estimated. This was done by using the error in seconds to calculate the number of meters travelled in this time, based on the average speed between the current position and the target stops. This assumes that the vehicle travels at a constant speed and therefore was not used as loss function but rather for comparison.

IV. Results and Discussion

A. Data cleaning

The dataset spans 144 days (12-Oct-2019 to 04-Mar-2020) with an overall number of 1,909,861 instances (bus location records). These correspond to 4080 individual journeys as it can be seen in Fig. 4. This excludes 0.9312% of journeys due to speeds above 62 mph. Filtering by the direction as discussed in Section IV-D1, leaves 1486 (36.42%) of the overall number of journeys.

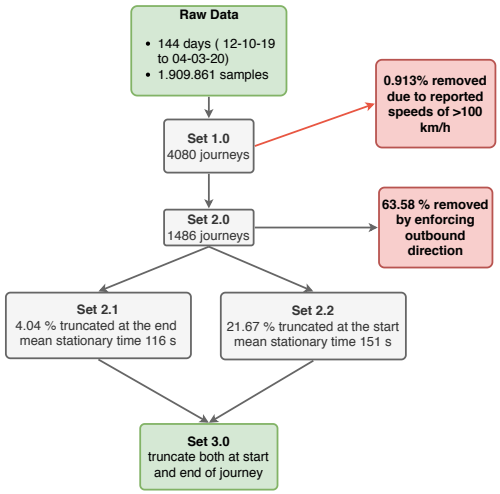


Fig. (4) Step-wise data cleaning sequence.

B. Benchmarking

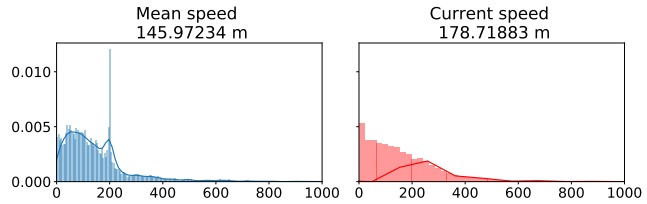
The purpose of the benchmark is to give a baseline to interpret subsequent results. Fig. 5 (a) shows the distributions of errors for each benchmark in meters.

The peak at 200 m stands out in the mean-speed benchmark. This error occurs when a vehicle remains stationary as this method does not allow for a stationary prediction. The average distance travelled along the trajectory corresponds to ~200 m (the error is calculated as straight line distance, therefore corners or loops will cause smaller errors than the same distance along a straight part of the route). Plotting the errors along the route gives a more detailed overview of the performance of the benchmark as shown in Fig. 5(b). This confirms the hypothesis that in general, the benchmark will perform poorly at stationary positions, which is especially evident at the start and the end of the journey when some vehicles remain stationary for extended periods time.

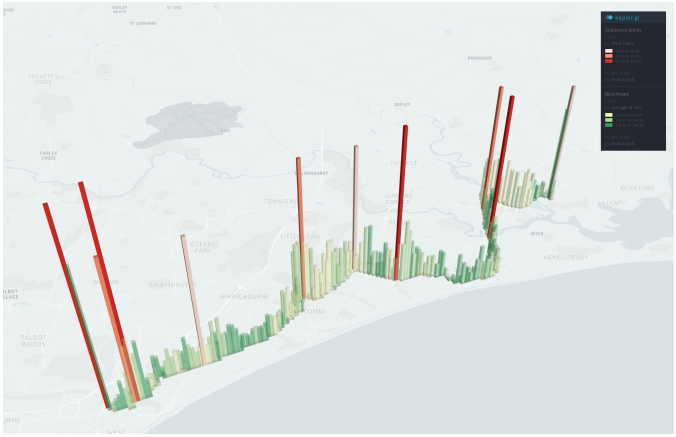
Interestingly set 1.0 which is the least processed and thus affected most by noise, had the best results. This is further discussed in Section IV-D. The mean-speed predictions will be used as a baseline from hereon.

C. Data quality

In addition to the aforementioned problems with the data quality, the collected data contained characteristic circular patterns (Fig. 6i). These occur at bus stops only and are not explainable by artefacts from GPS interference. An empirical investigation showed that the origin of this phenomenon is most likely a side-effect of geofencing that the AVL-system uses to determine if a vehicle has arrived at the stop. Unless the bus has been very close to the stop, the AVL-system 'snaps' the real position of the vehicle to the geofence boundary (Fig. 6 iii). By choosing this exclusion zone to be 10 m in radius, it was possible to simulate data mimicking the artefact seen in the real-life data (Fig. 6 ii). The issue requires further investigation to verify the exact rules this artefact is following.



(a) Benchmarks for set 2.0. The 200 m peak visible in the left plot occurs when a vehicle stops either at traffic lights, pedestrian crossings or a bus stop. This peak is not found in the current speed benchmark as it naturally compensates for variations in speed. Overall the global average does result in lower haversine mean-error.



(b) Performance evaluation of the mean-speed benchmark on set 2.0. The average error is shown in meters as green bars (colour and height indicate the average error along the route). The number of repeated positions are shown in red. The bars show the points at which more than 90% of repeated positions occur, generally at bus stops.

Fig. (5) Assessment of the benchmarks.

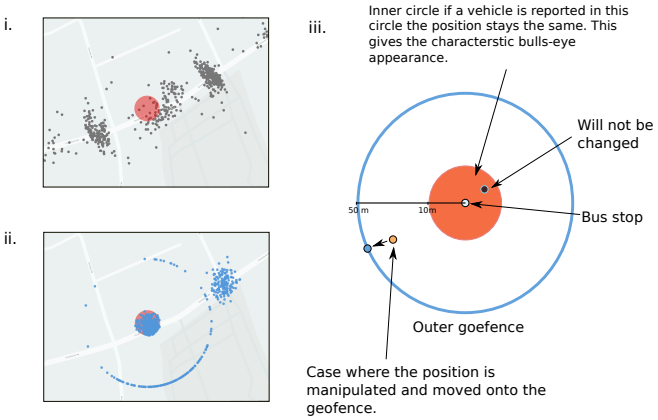


Fig. (6) **i.** The circular artefact recorded from real-life data. The red circle denotes the bus stop. **ii.** The simulated data generated closely resembles the artefact recorded. **iii.** The underlying process used to simulate the data.

D. Effects of data cleaning

To evaluate the effect of the cleaning steps on prediction quality, two target representations were tested: the prediction of unconstrained coordinates and the trajectory-based prediction. As the mean error of the trajectory was ~100 m lower when compared to unconstrained coordinates, only the trajectory-based predictions are shown for clarity.

Errors for both model types are shown in Fig. 7. The performance for all model types is similar in general. Interestingly the benchmark is very robust and only in set 2.2 do both RNNs have a slight advantage over the MAE of the benchmark (GRU: 145.86 m, mean-speed: 146.85 m) but the difference is negligible. However, in practice mean error is not the best metric for model evaluation, as it does not account for the spread of the errors. To better assess the performance, the Sharpe ratio [19] was used (Eq. 1). Widely used in finance, it accounts for the standard deviation of the errors or their volatility. This gives a different picture and the benchmark is outperformed except in the case of set 2.1 (see Fig. 9).

$$S = \frac{MAE - r}{\sigma} \tag{1}$$

(1) Sharpe ratio (S), where r is the risk-free rate which here translates to the best expected error (assumed to be perfect at 0m) and σ is standard deviation of errors.

Comparing the error distributions of the network with the lowest MAE (GRU using set 2.2) to the benchmark (Fig. 8) it becomes apparent that the GRU's distribution is skewed more toward smaller errors and is not bi-modal like the benchmark. This confirms that in practice the GRU will deliver a more reliable prediction even though the mean error is similar.

1) Set 2.0 filter direction: The evaluation of the cleaning steps shows interesting behaviour. The first cleaned dataset 2.0 shows an increase in the mean error of both RNNs and the mean-speed benchmark. It would be expected that limiting the data to a single direction should improve predictability. When assessing the Sharpe ratio the findings are different and the LSTM shows an improvement of 0.189% whereas the other methods decrease in performance with large errors (Fig. 9).

2) Set 2.1 removal of end repetitions: At the end of a journey the vehicle will in some cases repeat transmission of the same position. This occurred in 4.04% of the journeys with an average stationary time of 116 s. As the transmission frequency is 40 s this represents ~3 repeated positions. This only affects a small portion of the journeys but still did worsen the performance compared to the raw dataset (set 1.0) both when assessed using MAE and causing a reduction of -12.755% of the Sharpe ratio. This suggests that both LSTM and GRU had an advantage in those few stationary cases.When assessing the error map of the LSTM for this dataset (Fig. 5) it becomes apparent that the error of the set 2.0 at the final datapoints is ~90 m, whereas upon removing repetitions the error in the same area ranges from 283-818 m. This suggests that the LSTM became exceptionally good at predicting the final

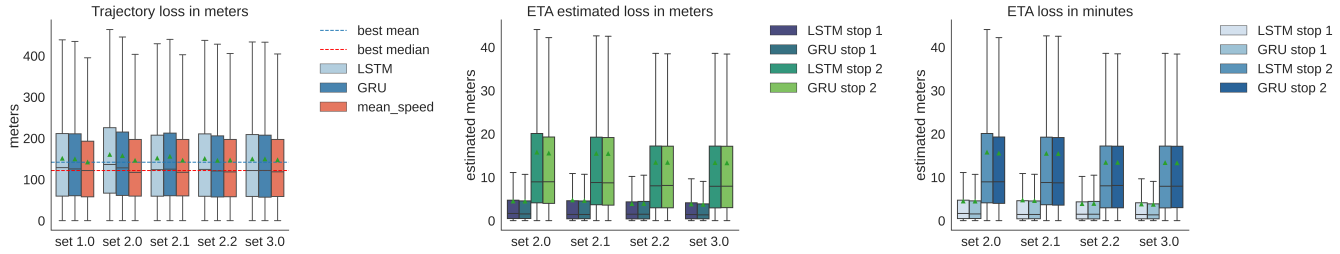


Fig. (7) **Left:** Boxplot showing error in meters for GRU, LSTM and the mean-speed benchmark. Outliers have been removed. Green triangles represent the mean and the median is represented as a horizontal black line. The best benchmark's (based on Sharpe ratio set 1.0) median and mean are shown as red and blue dashed lines respectively. **Middle:** Boxplot showing the estimated error in meters for the ETA prediction. Both networks are shown and errors are given for the first and second stop. Boxplots showing the errors in minutes for the ETA prediction for either network in comparison to the benchmark. The prediction is more accurate for the immediately next stop and the error increases for the second stops. Note the difference in error magnitude. **Right:** Boxplot showing the ETA loss in minutes.

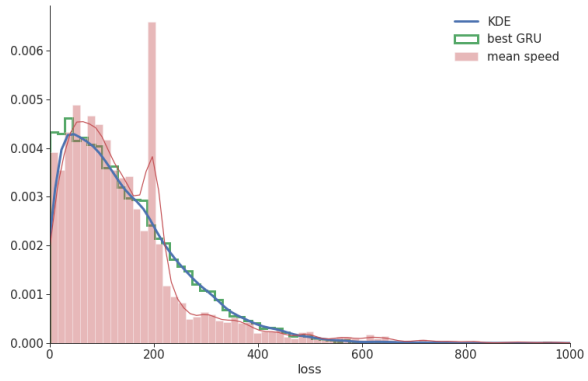


Fig. (8) The error distribution for the best performing GRU and the mean-speed benchmark. The 200 m peak caused by stationary vehicles is apparent. The outline of the GRU errors suggests that the model makes more reliable predictions.

positions if the vehicle remained stationary thus appearing to improve the overall performance.

3) Set 2.2 removal of start repetitions: A large proportion of vehicles idle at the start of a journey, which affects 21.67% of the journeys with an average idle time of 151 s corresponding to ~4 repeated positions. This means a much larger proportion of the vehicles will arrive early at the start of their journey compared to those that remain stationary once the journey is finished. However, also this cleaning step reduced the Sharpe ratio by -4.499% for the LSTM when compared to set 2.0. Interestingly, the GRU performed best in this scenario, whereas in all other cleaning steps the LSTM outperformed the GRU. In addition, under this scenario the best overall MAE was achieved by the GRU of 145.86 m. This suggests that the GRU suffered most from repeated starting points.

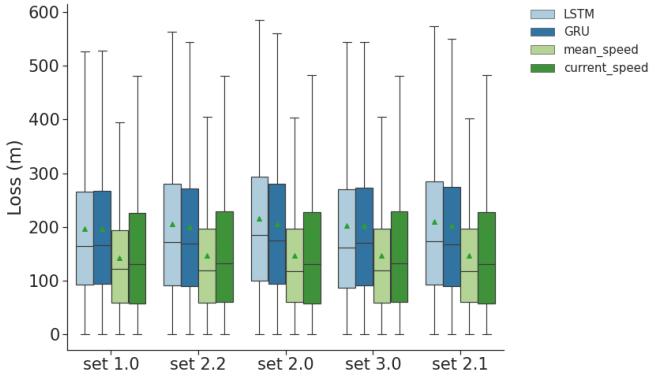
4) Set 3.0 removal of repetitions at the start and end: The final dataset 3.0 is the most processed and reflects a journey most closely with the lowest influence from artefacts. Interestingly, models trained on this dataset did not achieve the overall best performance. The LSTM lies -4.8% below the Sharpe ratio of set 1.0. This might appear counter-intuitive. However, on close inspection, this behaviour can be explained by the fact that the RNNs are most accurate when a vehicle is stationary. This causes the reduction of the Sharpe ratio due to over 20% of the journeys having repeats at the start or the end, making these cases easy to predict. This causes the paradox that the cleanest dataset appears to perform worse yet the reason for this is that the model is no longer able to predict the artefacts of repeated positions at either end of the journey.

5) Interpretation of data cleaning results: It is important to look at error locations and where exceptional performance is achieved. It is obviously easier to predict the position of a bus when it finished the journey because it will not move again, whereas at the beginning the difficulty is to predict when the vehicle will start moving. This could be overcome if the timetable corresponding to a journey was known. It can hence be concluded that although the overall metrics have not improved after data cleaning, it is still beneficial to prevent falsely reduced metrics due to accurate prediction of artefacts.

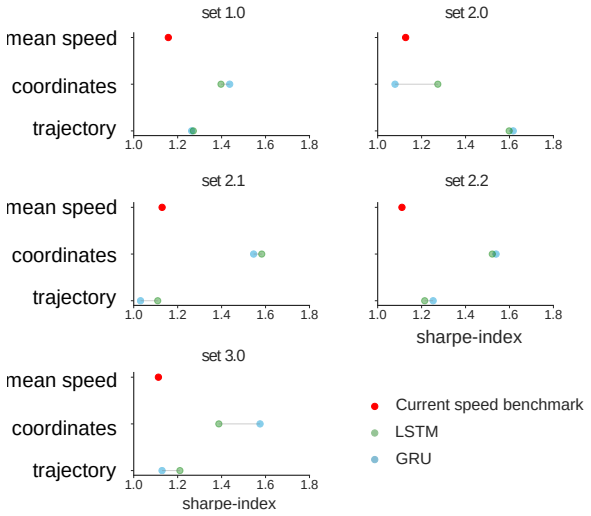
As mentioned before, the mean-speed benchmark fails when a vehicle comes to a stop as it assumes that all vehicles travel at a constant speed. Therefore, the benefit of RNNs is that it will learn locations where this is likely to happen such as bus stops, junctions or pedestrian crossings. This is indeed seen in the data, in areas with the most stationary data points the accuracy of the best RNN is lower than the average accuracy, meaning that the neural network has learned areas where a vehicle is most likely to stop for some time (Fig. 10).

E. The influence of target representation

The prediction of unconstrained coordinates is not an effective approach and results in large errors and therefore is not shown here. This not a surprising finding as this approach has a very large prediction space spanning the entire city. In reality, the possible predictions are along the route thus making this an unnecessarily difficult approach.



(a) The error of the networks and benchmark in meters. Using this evaluation metric the benchmark cannot be outperformed.



(b) Cleveland plot showing the Sharpe ratio (higher number = better performance). Using this more holistic metric compared to MAE the benchmark is generally outperformed.

Fig. (9) Comparison of different ranking approaches.

To combat this complexity issue the trajectory approach was used with much better results. However, this requires linking the data to a specific route pattern, which in practice can be challenging. This again is an expected finding as limiting the prediction space to the route which is the only possible space the vehicle should be travelling on simplifies the problem dramatically, reducing the errors approximately by half from 244.8 m (LSTM) to 141.3 m (LSTM).

The third approach predicting the ETA at the two subsequent bus stops might be the most important method. To make this approach more comparable the error given in seconds was translated into rough estimates of a distance error.

All models trained to predict ETAs used the last 3 positions to interpolate ETA and performed better than the benchmark (Fig. 7). The benchmarking method using the current speed was chosen as it performed substantially better than the overall-speed based method (current speed MAE: 1.2 min, mean speed MAE: 10.7 min). As would be expected the first stop is predicted with higher accuracy compared to the second stop. Both the GRU and LSTM perform better than the benchmark when comparing the MAE. Both the mean and Sharpe ratio are reduced by the data cleaning steps.

Across all sets the performance of the ETA prediction method is ~10 fold more accurate with estimated mean-distance errors of ~4.2 m for the first stop and 14.5 m for the second stop for both RNNs. This is a substantial improvement compared to the position-based method (the best scenario had an error of 145.8 m) and could be further improved by making additional information such as the distance to the next stops available to the neural network. Furthermore, the networks could easily be changed to predict the arrival times for all following stops. The drawback of this method is that in the data used the actual ETA is not known. Therefore, the ETA

Fig. (10) The number of repeated positions generally seen at main bus stops, junctions and crossings shown in red. The error of the best GRU trained on set 2.2 in turquoise. The error is generally low if a vehicle is more likely to stop in an area.

is an approximation and might not fully represent reality. However, to collect accurate ETA information the current technology would need to be upgraded e.g. with proximity sensors at each bus stop as the sample interval is insufficient.

In light of the findings representing short-term prediction-targets as ETA problem rather than a position-based target gives by far the best results with the caveat that the ground truth used to compare the ETA against is an estimation.

V. Conclusions

Bus travel is a well-established mode of public transport and the vehicles are mostly equipped with modern telemetry systems. However, we highlighted data quality issues, which complicate any data-driven solutions. Unreliable or omitted information about the route and timetable a vehicle is following, most likely inhibited the performance of the developed prediction models. Improving the availability and quality of

such data would allow to further advance ETA predictions. Additionally, in this study, the ambiguity of line numbers might have resulted in the loss of some journeys. Circular artefacts were also discovered that can be explained by use of a geofencing method that moves vehicle positions onto geofence boundary unless it has arrived at a stop. Such manipulation of the data stream could hamper prediction efforts although the assessment is difficult without ground truth.

This study used benchmarks to make the findings easily comparable to other studies. We have shown that a simple metric such as mean error cannot be used to objectively compare algorithms. To make an informed decision it is crucial to use several metrics. The Sharpe ratio used to account for the standard deviation in addition to the mean error, proved to be a better measure than a simple MAE. Furthermore, the importance of assessing the error distribution was highlighted where it was possible to see that, for example, the mean speed benchmark performed especially poor if the vehicle was stationary, which is impossible to deduct from simpler metrics.

The extracted journey data was affected by artefacts such as repeated position records at either end of the journey. Therefore, it was necessary to remove such artefacts and assess their impact on the final prediction. Unexpectedly, the step-wise cleaning approach did not improve the overall MAE of the predictions compared to the raw data. This can be explained by the fact that the RNNs perform especially well when predicting stationary positions of buses at stops or idling at either end of the route – over 20% of the journeys have repeated positions at the start. This is a large number of predictions that can be made with exceptional accuracy thus giving the appearance that the predictions are more improved the noisier the data is. In other words, the more stationary points a dataset contains the better will be the overall prediction accuracy. Such a model is naturally not very useful in an operational context where the emphasis is on predicting vehicles in motion. Therefore, it is crucial to assess each developed algorithm in depth by examining errors along the route and focusing on any patterns that might be contained in such data. Even though the overall prediction rate did not improve, the RNNs did perform better than the benchmark at stops along the route and with a general better accuracy along the route while the vehicle is in motion.

Using alternative representations of the same target considerable improvements in accuracy have been made (66 m between the trajectory and unconstrained coordinates). The intuition is that by simplifying the problem and reducing the prediction space the models will achieve better results. In practice, this meant that predicting unconstrained coordinates did not perform well, whereas limiting the prediction-space to the trajectory and subsequently transforming the problem to an ETA prediction improved the results 10 fold. The overall winner was the ETA prediction. Operationally, this could be considered the most important algorithm as ETAs for example displayed at bus stops or in mobile apps could be considered more important than short-horizon predictions at all points along the route. However, a short-horizon prediction compensating for transmission delays in 'live' location representations on the web or mobile apps, will make the user experience better, benefiting those passengers, who rely on such features.

Overall, this study highlighted the urgency to make all available data accessible to develop the best data-driven solutions in public transport. It furthermore illustrated the importance of not only relying on mean-based metrics but using a selection of different metrics in combination with geographical error representation to objectively assess any prediction algorithms. Additionally, even though in theory modern deep learning methods should learn to predict a target in any format, in practice they perform best if faced with the most simple representation of the task. As a conclusion and suggestion for further work, it is necessary to address the highlighted lack of data, as well as the lack of benchmark datasets. This will be addressed in an imminent paper. Furthermore, it is worth to consider the development of an evaluation framework specifically tailored to public transport prediction methods, consisting of a collection of different metrics and a formula to assess the geographic variation of errors.

REFERENCES

[1] M. M. Salvador, M. Budka, and T. Quay, "Automatic Transport Network Matching Using Deep Learning," *Transportation Research Procedia*, vol. 31, no. 2016, pp. 67–73, 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352146518301273

[2] J. W. Grotenhuis, B. W. Wiegmans, and P. Rietveld, "The desired quality of integrated multimodal travel information in public transport: Customer needs for time and effort savings," *Transport Policy*, vol. 14, no. 1, pp. 27–38, 2007.

[3] R. G. Mishalani, M. M. Mccord, and J. Wirtz, "Passenger Wait Time Perceptions at Bus Stops : Empirical Results and Impact on Evaluating Real- Time Bus Arrival Information," *Journal of Public Tr*, vol. 9, no. 2, pp. 89–106, 2006.

[4] Department of Transport and Drivers and Vehicle Licensing Agency, "All vehicles (VEH01)," 2018. [Online]. Available: https://www.gov.uk/government/statistical-data-sets/all-vehicles-veh01table-veh0101

[5] G. Cookson and B. Pishue, "INRIX Global Traffic Scorecard," p. 44, 2017. [Online]. Available: https://media.bizj.us/view/img/10360454/inrix2016trafficscorecarden.pdf

[6] T. Xia, M. Nitschke, Y. Zhang, P. Shah, S. Crabb, and A. Hansen, "Traffic-related air pollution and health co-benefits of alternative transport in Adelaide, South Australia," *Environment International*, vol. 74, pp. 281–290, 2015.

[7] P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, and M. C. González, "Understanding road usage patterns in urban areas," *Scientific Reports*, vol. 2, no. August 2015, 2012.

[8] M. Hickman, "Bus Automatic Vehicle Location (AVL) Systems," in *Assessing the Benefits and Costs of ITS*. Boston: Kluwer Academic Publishers, 2006, pp.

59–88. [Online]. Available: http://link.springer.com/10.1007/1-4020-7874-9_5

[9] N. T. Al Ghifari, A. Setijadi Prihatmanto, R. Wijaya, and R. Yusuf, "Data Quality Measures and Data Cleaning for Pattern Analysis Angkot Transportation in Bandung City," *Proceeding - ICoSTA 2020: 2020 International Conference on Smart Technology and Applications: Empowering Industrial IoT by Implementing Green Technology for Sustainable Development*, 2020.

[10] Y. Li and T. Voege, "Mobility as a Service (MaaS): Challenges of Implementation and Policy Required," *Journal of Transportation Technologies*, vol. 07, no. 02, pp. 95–106, 3 2017. [Online]. Available: http://www.scirp.org/journal/doi.aspx?DOI=10.4236/jtts.2017.72007

[11] H. Yin, S. C. Wong, J. Xu, and C. K. Wong, "Urban traffic flow prediction using a fuzzy-neural approach," *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 2, pp. 85–98, 2002.

[12] B. Agard, C. Morency, and M. Trépanier, "MINING PUBLIC TRANSPORT USER BEHAVIOUR FROM SMART CARD DATA," *IFAC Proceedings Volumes*, vol. 39, no. 3, pp. 399–404, 2006. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1474667015359310

[13] A. Misra, A. Gooze, K. Watkins, M. Asad, and C. Le Dantec, "Crowdsourcing and its application to transportation data collection and management," *Transportation Research Record*, no. 2414, pp. 1–8, 2014.

[14] P. Wepulanon, A. Sumalee, and W. H. K. Lam, "A real-time bus arrival time information system using crowdsourced smartphone data: a novel framework and simulation experiments," *Transportmetrica B*, vol. 6, no. 1, pp. 34–53, 2018. [Online]. Available: https://doi.org/10.1080/21680566.2017.1353449

[15] L. J. P. Van Der Maaten, E. O. Postma, and H. J. Van Den Herik, "Dimensionality Reduction: A Comparative Review," *Journal of Machine Learning Research*, vol. 10, pp. 1–41, 2009.

[16] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[17] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS Medicine*, vol. 15, no. 11, pp. 1–17, 2018.

[18] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay," *arXiv*, pp. 1–21, 2018. [Online]. Available: http://arxiv.org/abs/1803.09820

[19] W. F. Sharpe, "The Sharpe Ratio," *The Journal of Portfolio Management*, vol. 21, no. 1, pp. 49–58, 1994.