For this experiment a set of networks was generated similar to the previous experiment and only their quadtree creation time recorded.

The creation of the quadtree was broken down into the following steps.

1. Cycle through all the nodes in the network and save the minimum and maximum x and y coordinates.
2. Create a quadtree, using standard MATSim functionality, with the x and y extent of the network.
3. Cycle again through each node in the network and add the coordinates of the node into the quadtree.

Since the network was traversed twice to create the quadtree, the effect of an increase in network nodes was expected to have a 2f(n) increase in quadtree-creation duration.

As can be seen in Figure 4.5, the effect of the creation of the quadtree was fairly low compared to the map-matching duration. Based on the linear regression - even on big networks like the Cape Town network - the total creation time of the quadtree was less than a second. Given this and the fact that the quadtree was only ever created once for every network, no further analysis was done on the quadtree creation.

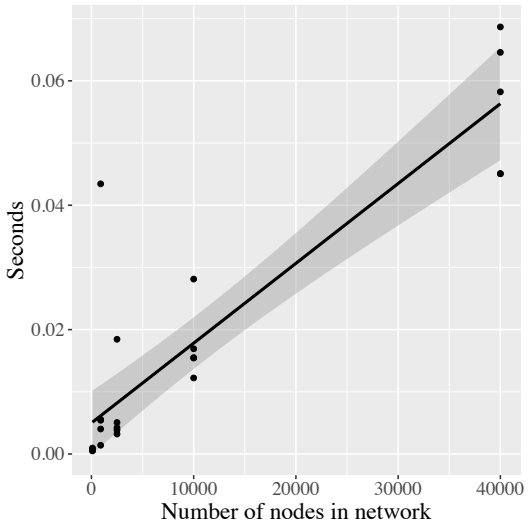


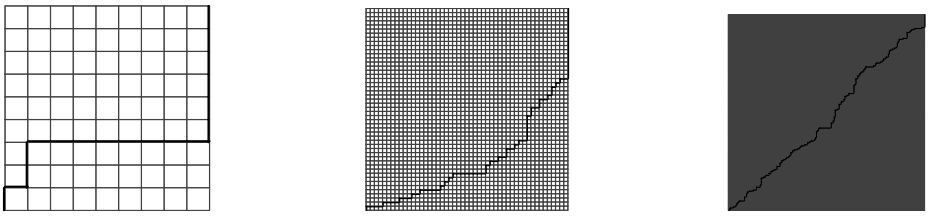
Figure 4.6: Analysis of quadtree creation duration

4.3.3 Network and path increase

This experiment aimed to assess the general outcome of running bigger experiments and to determine what the overall effect on the map-matching algorithm is if the network and path size increased proportionally in size.

For this experiment, the network size was increased by the same increments as the previous experiment, but the true path (TP) was recalculated for every experiment. Figure 4.7 shows three of the networks analysed. Changing the network size and TP lead to a relative increase in the number of GPS points together with segment and node increases in the network. Because the grid network generated for these experiments had segments of equal lengths and the starting and ending node of each true path were the same for each experiment, all the true paths generated for a specific network had the same length

but random routes. The free speed was also constant on all the links and since the GPS trajectory generation used the link free speed to simulate the object traversing the true path, the number of GPS points for each true path on the same network was also constant. The experiment was replicated 10 times for each network size.



(a) 10×10 (b) 50×50 (c) 200×200

Figure 4.7: Network and path increase examples

There seemed to be a non-linear increase in the execution time of the map-matching algorithm as the network and path size increased. This was because the increase in network size not only affected the execution time of the map-matching algorithm in general, as seen in Figure 4.5, but there was also an increase in the number of GPS points to match, because the length of the TP increased. This compounded the effect and led to an exponential increase in execution time.

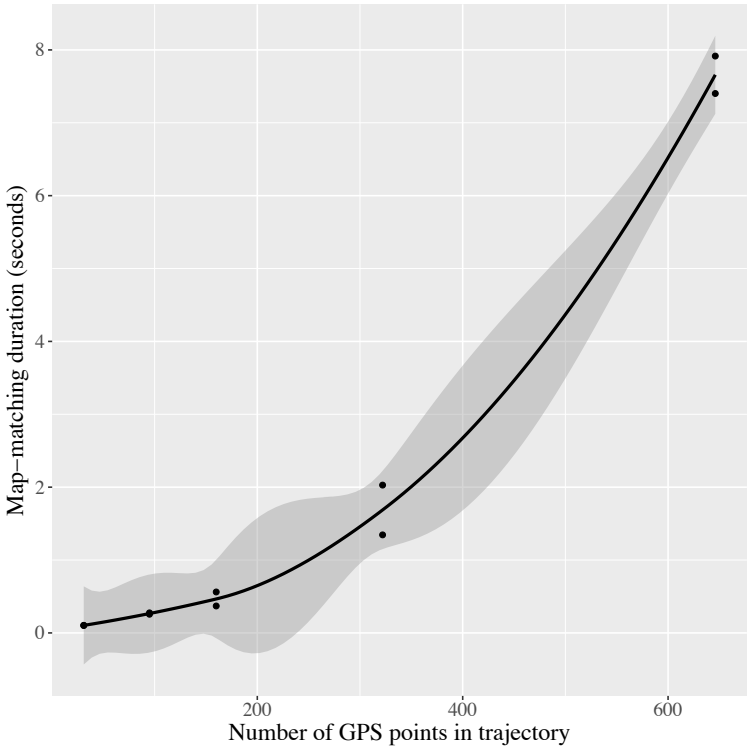


Figure 4.8: Map-matching duration versus network size and true path increase

4.4 Number of GPS points

The number of GPS points not only influenced the efficiency but also the effectiveness of the algorithm, as it had a direct influence on the number of data points available with which to calculate the true path with. Predefining GPS frequency for experimental setups or for picking data sets is key to ensuring that the algorithm will execute in an acceptable time as well as provide an acceptable level of accuracy required for the specific application.

The increase in duration is mostly because every GPS point will have a number of candidate links associated with it. These candidate links make up the nodes in the graph that will determine the most probable path of links that was traversed. As discussed in section 4.3.1, calculating the weight of each link between nodes in the graph is the most expensive part of the algorithm as it requires the calculation of the shortest path between the candidate links in the road network. The shortest path calculation uses the Dijkstra method, which is a resource intensive calculation: the more candidate points in a given calculation, the more permutations the Dijkstra method needs to calculate.

However, the quality of the calculation should increase considerably for an increase in GPS frequency, because the more GPS points being mapped, the higher the probability of not missing the correct links in a path.

But there seems to be a case where, if there are too many GPS points too frequent during the trajectory, then the algorithm erroneously identify unnecessary U-turns and loops. This is expected as the spatial-temporal (ST) algorithm used in this study was specifically designed to work with low-sampling rate GPS trajectories and does not include functionality to identify or prevent instances where too many GPS coordinates provide conflicting possible routes.

The influence of changes in the GPS frequency was assessed by generating different experiments on differently sized networks and varying the GPS points available for the algorithm to use for the same true path. For a specific true path a single GPS trace with a time interval 0.5 s was generated and then from that new traces generated with 1 s, 5 s, 7 s, 10 s intervals. Taking the longer period samples from the short period trace allowed for a direct experiment comparison between GPS periods. The experiment was replicated 10 times across the same five network sizes as used in previous experiments, and the number of closest links was set at 8. For each replication, the five different GPS trajectories were generated based on the GPS periods defined.

Although this algorithm was expected to cater for longer GPS periods, above 30 s, periods above 10 s were not tested because it was found that the accuracy of the algorithm was very low. This was the result of the short segments lengths in the simple grid network and the fact that moving from one corner to the opposite corner in the simple network will always have the same distance; there is no way for the algorithm to choose the correct path based on so few GPS points. This is an important finding for future use of the algorithm on simple grid networks and poses a possibility of future work to develop more complex controlled networks for controlled experiments. However, the longer-period GPS trajectories have been tested on the actual road network, and reported on in section 4.6. Figure 4.9 illustrates the issue experienced with GPS periods above 10 s on a grid network. The figure reflects a portion of the 200×200 grid network where the solid black dots indicate the 50 s period, and the grey crosses indicates the 100 s period trajectory. Since the 50 s and 100 s were sampled from the same original 0.5 s trace that was generated, black dots with grey crosses represent where the 100 s and 50 s intervals overlap and are at the same location. A myriad of possible routes can be drawn to match up on these GPS points and get the same spatial and temporal probabilities as the true path would. Fortunately, most

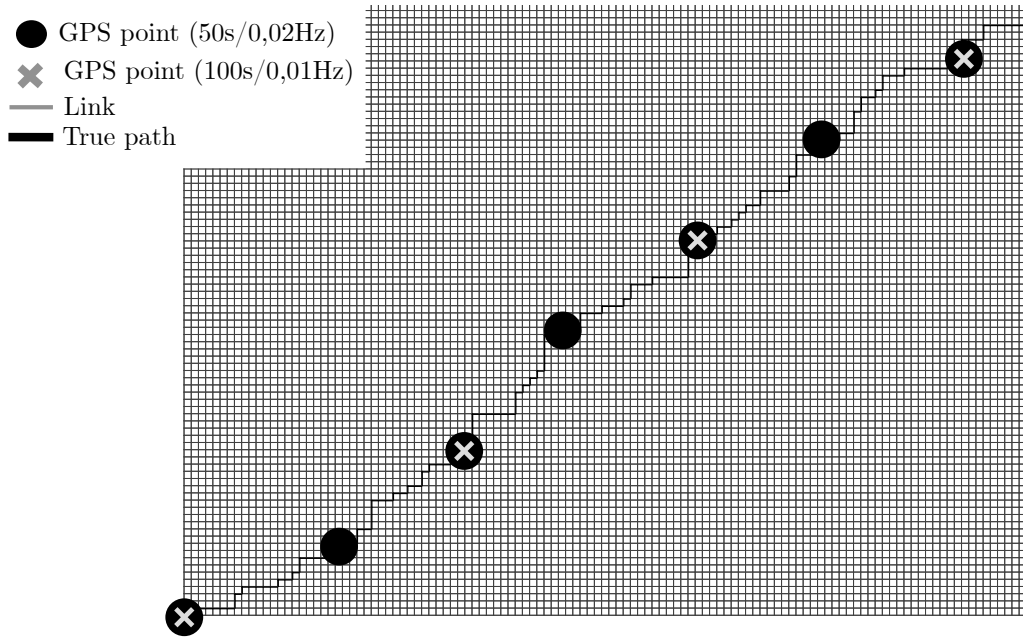


Figure 4.9: 200×200 grid with 50s and 100s GPS trajectory

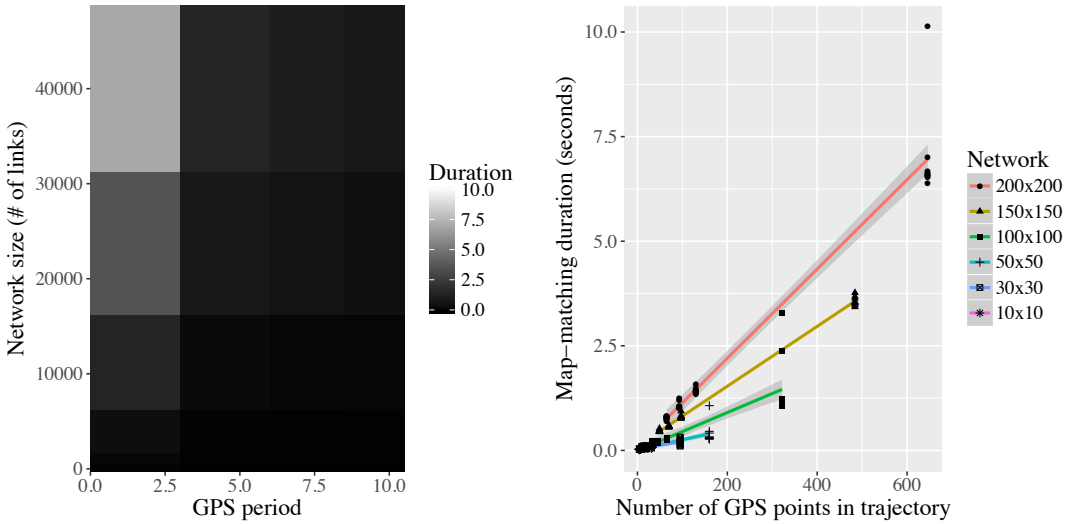
real-world networks have not only different segment lengths to map the contours of roads, but also varying speeds across different sections and it is these abnormalities in the road network that provide higher and lower probability routes for the map-matching algorithm to identify the TP correctly.

4.4.1 Efficiency

The number of GPS points generated is a function of the GPS period and the length of the true path. Since the length of the true path is a function of the network size in these experimental grid networks, Figure 4.10a illustrates the relationship between duration and GPS period split per experimental network size. It is evident that a drastic drop in execution time is experienced when moving from 0.1 s to 0.5 s across all networks and that the decrease in execution time is more profound on the bigger network than on the smaller network.

Figure 4.8 showed an exponential increase in the execution time as the number of GPS points increased; however, Figure 4.10b illustrates that there is a linear relationship between an increase in the number of GPS points and duration of execution but only if the network size is kept constant. This suggests that the rate of increase in execution time, as the number of GPS points increases is also dependent on the size of the network. This then explains why a non-linear increase in execution was reported in section 4.3.3 when increasing both network and true path.

It was concluded that a similar increase in the number of GPS points to use during map-matching will have a higher increase in execution time in bigger networks than in smaller networks. Thus reducing the network size to include only the parts of the network the map-matching is most likely to use, will lead to a decrease in execution time. This is most likely a worthwhile pursuit if a significant number of trajectories are to be mapped



(a) Efficiency; GPS period versus network size

(b) Efficiency versus number of GPS points per network size

Figure 4.10: Efficiency analysis for number of GPS points

on an excessively large network that contains large portions of unused road networks.

4.4.2 Accuracy

Since the lengths of segments in the experimental network are exactly the same, the number of accurately identified segments expressed as a number, ARRn, and in terms of length, ARR, will be the same. In this section only ARR will be presented and analysed.

As expected, Figure 4.11 shows that a decrease in GPS periods leads to a decrease in accuracy, but interestingly enough, not a significant decrease in Al. Evaluating the very low Al score for 0.1 s GPS period revealed that even though the route contains a 100% score for ARR, it is most probable that a significant number of incorrect links have also been identified.

Figure 4.12 illustrates the issue the map-matching algorithm has with high-frequency sampling rates on the specific network. The black lines indicate the matched route, the thick grey line the true path and the black dots the generated GPS points. When the GPS frequency is high, Figure 4.12a, the points are generated in such a fashion that some points might appear to be *behind* the previous point generated due to the GPS standard deviation and error margin. This causes the algorithm to evaluate that the object must have turned around at the next available point and travelled in the opposite direction. Given this relatively small network with numerous nodes this is a feasible option for the algorithm to evaluate given the high number of GPS points and segments available to create a valid and probable route.

If the GPS frequency is lower, for example 1 Hz, as in Figure 4.12b, it becomes less likely for the algorithm to identify spurious segments as there are fewer GPS points to create the possibility for U-turns in the trajectory.

A further problem with the simple grid network is that as soon as the GPS frequency becomes so low that the network allows for numerous paths to provide a suitable route through all the GPS points, the accuracy deteriorates significantly.

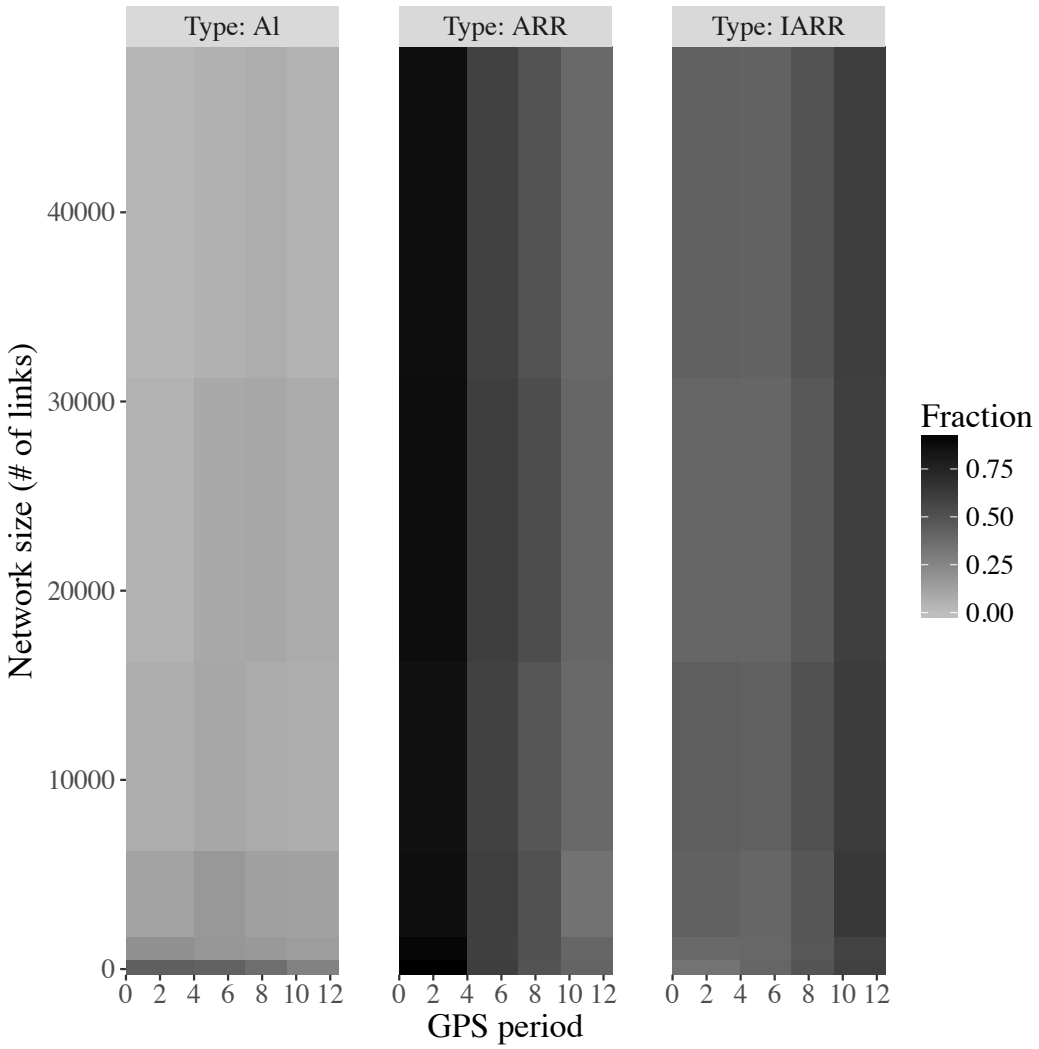
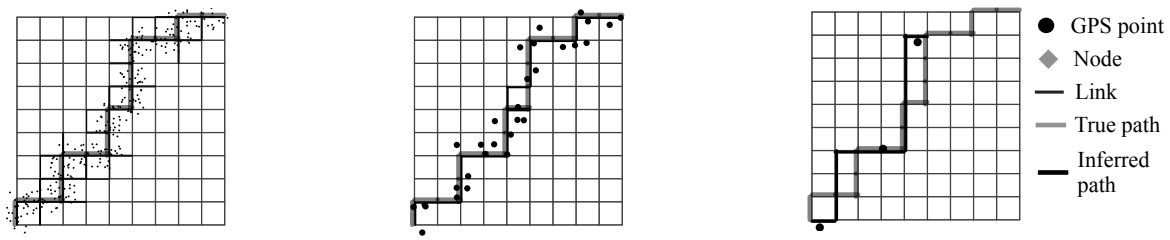


Figure 4.11: Accuracy analysis of GPS period versus network size



(a) 10 Hz map-matching (b) 1 Hz map-matching (c) 0.1 Hz map-matching

Figure 4.12: Examples of true path versus matched routes on simple grid network

4.4.3 Speed analysis

In this section the study took a novel approach to analysing the output of the algorithm by comparing the *inferred speed*, the calculated travel speed of the object across the links of the inferred path (IP) to the free speed of the links. This aimed to be an accuracy measurement or at least a sanity check that can be used in the absence of the existence of a true path. The assumption was that a vehicle will under normal conditions travel at a speed close to the segment free speed.

This measurement can be improved further by making use of historical actual speeds on the links based on time of day and using that as a comparison to the inferred speed. This will provide a more realistic and dynamic baseline to compare the inferred speed to, instead of the fixed free speed which is just the speed restriction on the road.

Figure 4.13 illustrates that with the high-frequency GPS trajectories, the object will regularly travel faster than the free speed of a segment. This is evident by the significant number of U-turns in the matched route. As the GPS frequency decreases, and the restriction to match to specific segments relaxes, the algorithm is able to choose routes that adhere more closely to the speed restrictions.

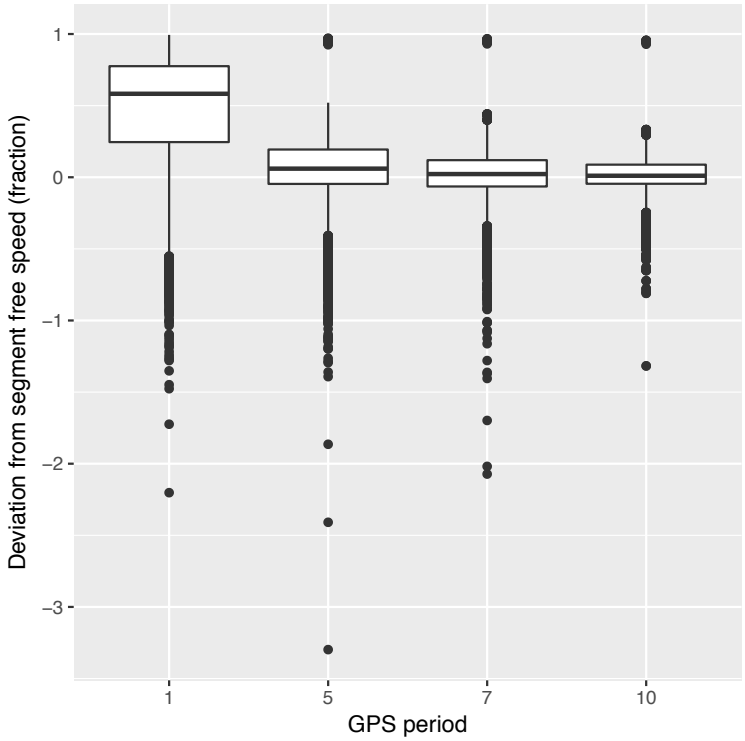


Figure 4.13: Speed analysis versus GPS period

4.4.4 Probability analysis

In this section, this study took another novel approach to analysing the output of the algorithm by investigating whether the ST algorithm's probability value assigned to the graph links can be used to estimate the accuracy of the IP in the absence of a TP. The aim was to give a confidence level to the results when analysing trajectories.

Figure 4.14 illustrates that there does exist some correlation between the overall probability of a match and the accuracy, but that it is not a clear-cut relationship. For example,

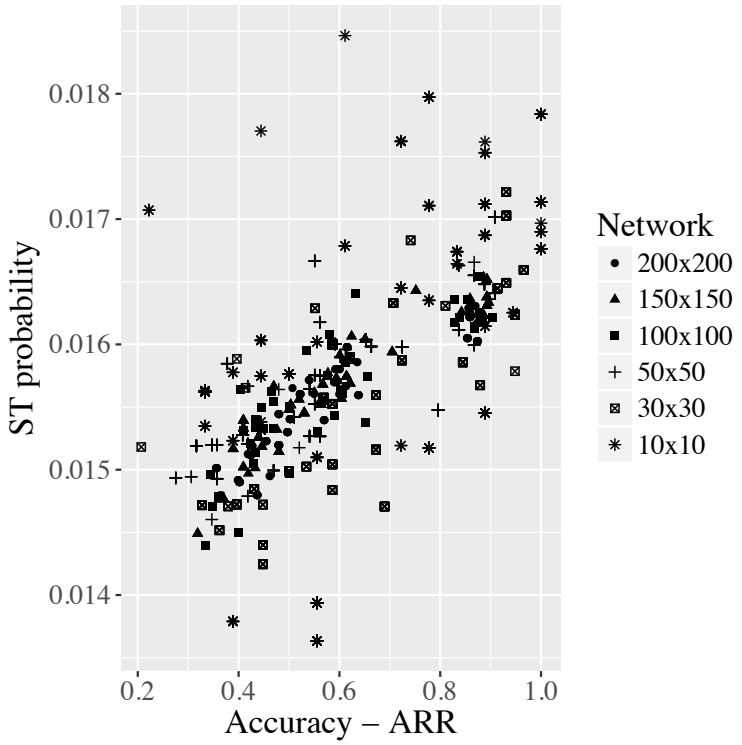


Figure 4.14: Probability versus ARR, grouped by network size

it seems that probability results above 0.017 can be linked to accuracies from 0.40 all the way to 1, for the small network 10×10, indicating no real correlation. However, for the larger networks probabilities above 0.017 can be linked to accuracies from 0.6 to 1, except for a few outliers.

4.5 Number of closest links

Similarly to the number of GPS points, the number of closest links has a significant effect on the algorithm efficiency and accuracy. It determines the number of links that will be assessed as a candidate link for each GPS point in the trajectory. Not only does it determine the ability of the algorithm to ensure it is selecting the right list of possible links to assess, thus affecting accuracy, but it also increases the nodes in the graph that are used to assess the most likely route, thus increasing execution duration.

As discussed in section 3.2, in a complex network that contains long and short links in fairly close proximity, the number of links the algorithm needs to consider increases dramatically. Due to the density of links in urban areas the closest link might not always be the correct link. For these simple grid experiments the conclusions drawn cannot necessarily be extrapolated to other network configurations and one must revert to real-world network tests, similar to those in section 4.6.

For this experiment the number of links in the simple grid network was kept constant at 100×100, with the same free speed and link length as used in the previous experiments. Based on the results from section 4.4, the GPS frequency was set at 0.2 Hz. It was deemed unnecessary to experiment with the number of closest links using two or more factor experiments, e.g. comparing the network size or GPS period effect with number of closest links. The effect of these factors have already been identified in the previous experiments

and there was no reason to believe a compounding effect exists between number of closest links and other parameters.

All experiments had 20 replications for each factor of the number of closest links parameter.

4.5.1 Efficiency

As per Figure 4.15 the number of closest link to assess for each GPS points appears to have a semi-linear impact on the map-matching duration. As the number of closest links to assess increases so does the number of nodes to add to the candidate graph. As found in section 4.3.1, the candidate graph creation was the most time-consuming activity of the algorithm and searching for the shortest path in the candidate graph will increase exponentially with an increase in the number of nodes.

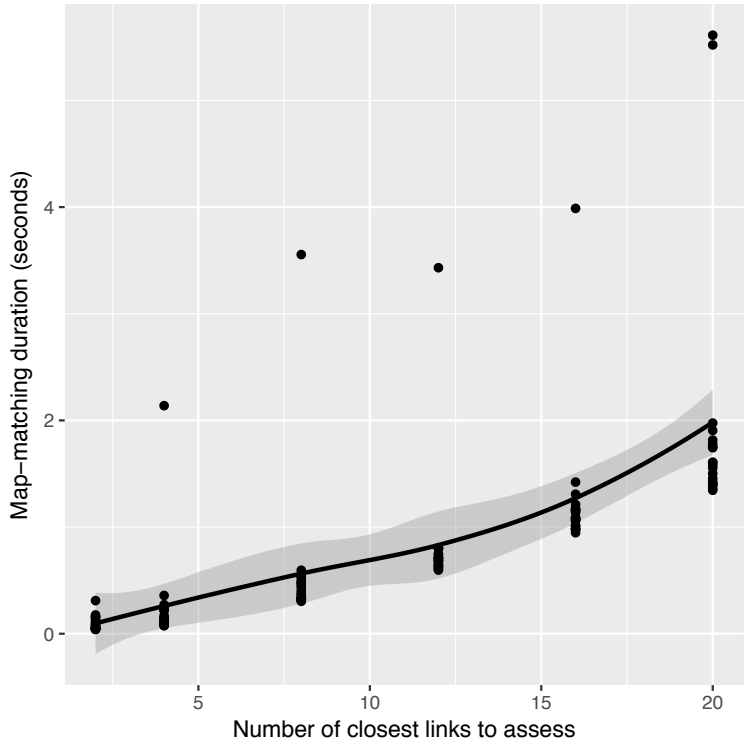


Figure 4.15: Map-matching duration versus number of closest links

4.5.2 Accuracy

Figure 4.16 shows that the accuracy of the algorithm increases and the probability to identify incorrect routes decreases when more links are being considered in the map-matching algorithm. The main purpose of this analysis was to determine at which point does assessing more links stop yielding an improved accuracy and a decrease in inaccuracy, because assessing more links is a costly action in terms of computational power and algorithm efficiency. This appears to occur around eight closest links but this number can be significantly different in real-world networks.

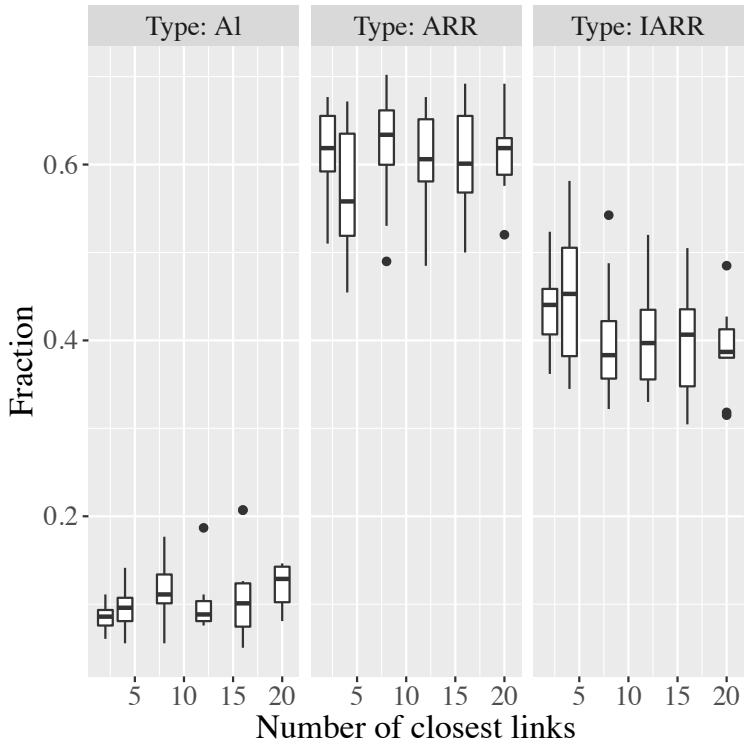


Figure 4.16: Accuracy versus number of closest links

4.5.3 Speed analysis

The speed analysis measures the deviation of inferred speed to the free speed on the segments in the IP. This analysis has no input from the TP but since all of the data was generated based on an object moving at a segments free speed it can be assumed that any outliers indicate a segment that was incorrectly identified. Due to the nature of the simple grid network, all experiment results generated using the same number of links could be combined for statistical calculations since the length of the true paths and number of the GPS points per link was exactly the same. As such all the links identified for all IPs have been grouped based on number of closest links and analysed using a box-plot as per Figure 4.17.

It is interesting to note that given the fixed GPS period and true path for each experiment, changing the number of closest links did not seem to have a significant impact on the adherence to free speed analysis. In other words the algorithm's ability to choose realistic link options for the GPS trajectory was not impacted by the number of closest links, but the accuracy and inaccuracy is affected. This might not hold true at lower or higher GPS rates or in different real-world networks where the links have varying free speeds and lengths.

The simple grid network yielded a very low accuracy level when compared to other references of real-world network map-matching in literature, which sources such as Qudus & Washington (2015), quoting 70% accuracy as very low. However, Lou et al. (2009) quoted 67%, on 10 min sampling intervals, as an improvement compared to previous methods.

In conclusion, using a very simple grid network to assess the algorithm performance seems adequate for efficiency analysis but the universal length and free speed of the links put uncertainty on the results of accuracy and speed analysis. In order to get a more representative analysis of the performance, and to validate the results of the simple grid

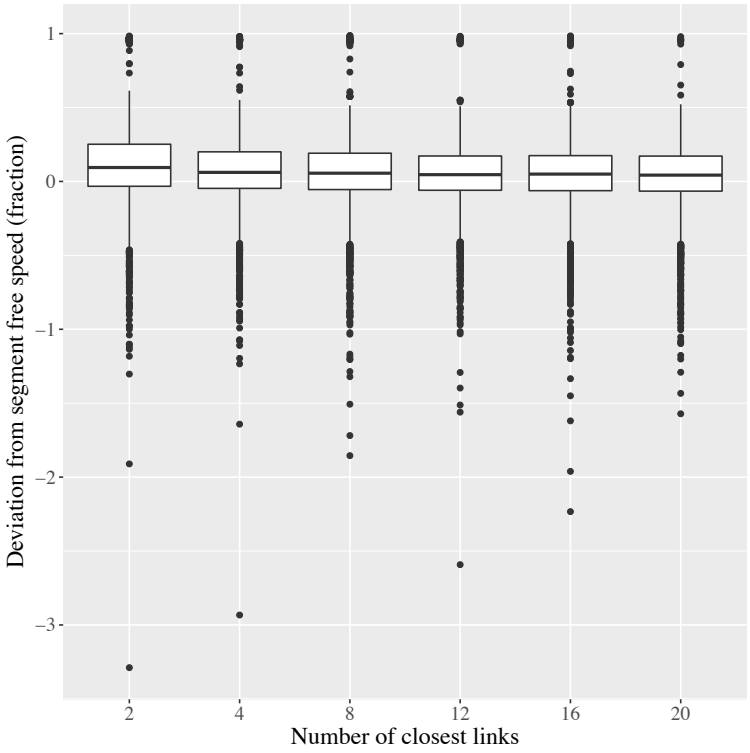


Figure 4.17: Speed analysis versus number of closest links

network some of the experiments from this section where repeated, as reflected in the next section, but instead of a simple grid network, a real-world network was used.

4.5.4 Probability analysis

In this section, the study evaluated whether the same relationship of probability versus accuracy can be observed when experimenting with the number of closest links as was observed in Figure 4.14 when experimenting with network size. Figure 4.18 illustrates that there is less correlation between probability and accuracy when using the two closest links in the map-matching algorithm. The rest of the experiments show, as before, that there exists a minor relationship between probability and accuracy. This does not appear to be meaningful enough to draw a statistically significant conclusion when analysing trajectories without a TP. Further work and investigation was needed to provide a confidence level for such outputs

4.6 Experiments on real-world road network

In this section, the real-world road network of the City of Cape Town is discussed, and the GPS sampling rate and number of closets links experiments repeated. Experiments on yielded satisfactory results but future studies could extend this analysis not only by generating more complex trajectories with varying speeds of movement but also setup a real-world network with different representations of the road topologies. For example using short segment lengths to represent road curves versus long segment lengths approximating the topology.

For the road network, a dataset of true paths was generated together with an accompa-

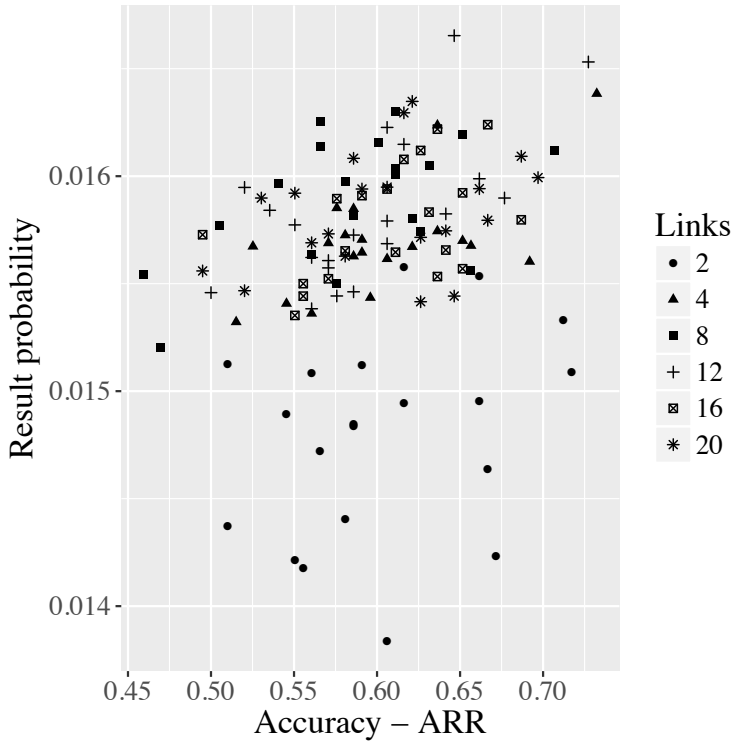


Figure 4.18: Probability versus ARR, grouped by number of closest links

nying set of GPS trajectories for each true path. For each true path, different trajectories where extracted from a single trajectory of 10 Hz so that the results from one trajectory can be compared directly with the other. The GPS trajectories generated where at frequencies of 5 Hz, 1 Hz, 0.2H z, 0.1 Hz, 0.05 Hz, 0.02 Hz, 0.01 Hz.

There was no common characteristic for true paths generated across the network, other than using the same GPS frequency, and thus the results from one true path cannot be compared directly to the results from another true path as per the simple grid network. All true paths were generated based on three random nodes selected within the network and finding the shortest path from node 1 to node 2 and the shortest path from node 2 to node 3. See Figure 4.19 for an example of a trajectory generated, the true path and the inferred path. For this specific example the map-matching algorithm had a ARR 96% and a 4% inaccuracy ratio of route by length (IARR). For this analysis, the standard deviation of GPS error was kept at 20 m and the mean GPS error at 0 m. Each dataset was assessed using the following set of closest links: 8, 12 and 16. Thus, for each true path generated on the network, the eight different GPS trajectories were assessed three times using the different number of closest links. The road network was kept constant during the analysis and was not reduced to include only the links that would most likely be used in the analysis of a specific trajectory as this was outside the scope of the current study.

4.6.1 Efficiency

The efficiency of the algorithm decreased considerably in the real-world network compared to the simple grid network. In Figure 4.20, we see that even experiments with GPS points below 500 points had a duration of over 200 s compared to around 7 s for 600 GPS points in the 200×200 simple grid network experiment. Based on the conclusions reflected in section 4.3.1, this would be mainly due to the significant number of links in the real-world