Modelling of inter-stop minibus taxi movements: Using machine learning and network theory.

Innocent Ndibatya
School of Computing and Informatics Technology,
Makerere University, Kampala, Uganda
(Exchange student at Stellenbosch University)
Email: indibatya@cis.mak.ac.ug

MJ (Thinus) Booysen
Department of Electronics and Electrical Engineering
Faculty of Engineering
University of Stellenbosch, Stellenbosch, South Africa
Email: mjbooysen@sun.ac.za

***Abstract*—Minibus taxis provide affordable alternative transport for the majority of urban working population in Sub-Saharan Africa. Often, these taxis do not follow predefined routes in their endeavours to look for passengers. Frequently, they stop by roadsides to pick up passengers and sometimes go off the main route in an attempt to fill the taxi with passengers to make the trip profitable. In addition, the destinations are changed from time to time depending on the driver. This uncoordinated movement creates a web of confusion to would-be passengers. The key aspects that are not clear to the passengers include; where to get a taxi, the waiting time and the travel time to the destination. These conditions leave taxi passengers at a very big disadvantage. In this research, we applied the concepts of machine learning and network theory to model the movements of taxis between stops. The model can be used to compute the waiting times at the stops and the travel times to a specified destination. Twelve minibus taxis were tracked for 6 months. Density-based clustering was used to discover the formal and informal taxi stops, which were modelled into a flow network with the significant stops as nodes and the frequency of departures between nodes as edges representing the strength of connectivity. A data driven model was developed. From the model, we can predict the time a passenger will have to wait at a stop in order to get a taxi and the trip duration.**

I. Introduction

The medium-sized minibus taxis which carry 10 to 15 passengers dominate the public transport sector in South Africa. 60 % of South African citizens rely on them and they transport an average of 14 million people every day. However, they are not reliable and not properly regulated by government. If by chance a passenger stands at the right pick up place, they do not know how long to wait before the next taxi will pass by and the travel time to reach their destination. In our earlier study [1], we used machine learning to find the formal and informal stops (Figure 1 (a, b and c)) of taxis operating between towns in the Western Cape. During that study, it was discovered that the rate at which taxis stopped at different stops varied according to the days of the week and the times of the day. We demonstrated that by using historical GPS locations of taxis, we could predict the most probable places where a traveller could get a taxi depending on the time of the day and the day of travel. However, we were not able to tell how long the person would wait at the stop in order to get a taxi and the duration of the journey. We therefore recommended further investigations to find the waiting time at the stops, the routes taken by the taxis and the trip durations. In the current study we considered the waiting time at the stops and the trip durations of the taxis. In this paper we model the movements of the minibus taxis between the stops in order determine (1) the waiting time at the stops, and (2) the trip durations as the taxis move from one stop to the other.

We believe that this information is key to understanding the operations of the informal public transport sector where the minibus taxis are dominant. The rest of this paper is organised as follows. Section II discusses the theoretical basis of the research, section III discusses the methods used, section IV discusses the results and section V concludes the paper. In this paper, the term *asset(s)* is used to refer to the *taxis* that were tracked during the study, and sometimes, the words are used interchangeably. The term "*modelled system*" is also often used to refer the conceptual "experimental component set up" during the process of modelling.

II. Literature review

Researchers have in the past attempted to model transportation systems and many models have been developed. These models were categorised by Zhou and Dai [2] into two, i.e Freight transportation models for goods in transit and passenger transportation models for human passengers. In developed cities, passenger transportation models have traditionally been part and parcel of urban design [3]. The main objectives of the incorporation are (1) to reduce the number of motorized trips, (2) to increase the number of non-motorized trips, and (3) to increase vehicle occupancy. Traditional transportation models have always been theoretical and based on mathematical description of processes. Such models forecast travel demand based on trip generation, trip distribution, transportation mode (public/private) and route choice [4].

The 21st century era of powerful computing introduced a new method of modelling – Data Driven Modelling(DDM). DDM uses vast amounts of data about the process and methods of computational intelligence and machine learning to model processes [5]. Machine learning deals with the construction and study of algorithms that can learn from data. Some of the methods used for machine learning include, Network analysis, Neural networks, Markov models and many more. Network analysis in particular interprets the modelled process into a network graph with discrete objects represented as vertices and the relationships between the objects represented as edges. Combined with machine learning, researchers have applied network analysis to model problems in many fields such as in neural science [6].

 ISBN: 978-0-7972-1533-7

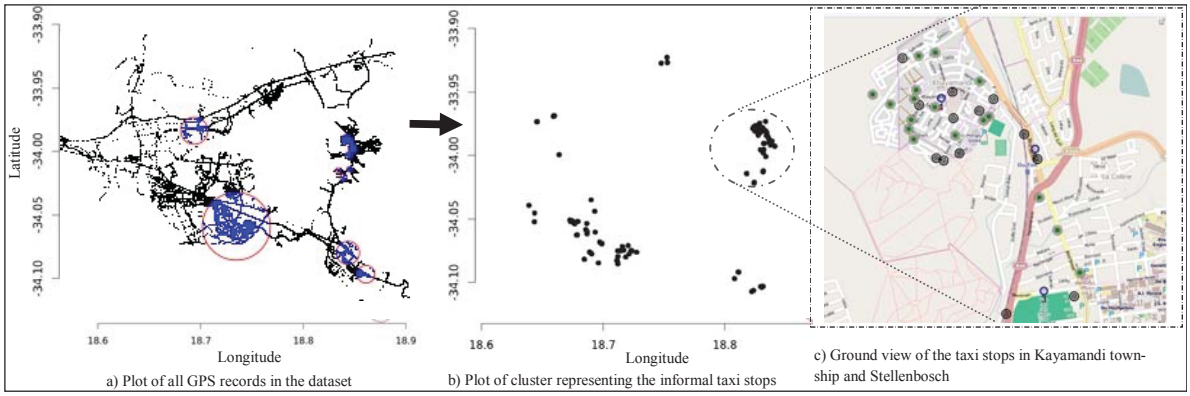


Fig. 1. Figures showing the formal and informal stops discovered during our previous study

The reviewed literature suggests that the network analysis method has successfully been used before to model processes. Furthermore, in the presence of data, DDM can be used to study processes and develop dynamic models that learn from the data. It is on this basis that we sought to apply the methods to model the movements of minibus taxis.

III. Methods used

In this study, we used two methods. (1) Vector overlay of GPS trajectories in a circle to map out the significant nodes/stops, and (2) network theory to model the movements of the taxis between the significant nodes/stops.

A. Data sources

Ten minibus taxis (assets) operating between Stellenbosch, Somerset West, Strand, and Bellville were equipped with GPS (Global Positioning System) tracking devices. Each device had a GSM (Global System for Mobile communications) sim card installed and would log GPS locations of the taxi to a remote server through the GSM network. Data was logged at a nominal frequency of 1Hz. Attributes of the data that was logged included - date and time, GPS location (longitude and latitude), speed and direction among others. This dataset contained a total of 1,842,570 GPS records collected over a period of 6 months (December 2013 - May 2014).

B. Vector overlay of GPS coordinates

Overlay analysis is a technique commonly used in GIS (Geographical Information Systems) studies where layers of features are analysed to find intersects, unions and clips. In this study we performed circular overlays over different regions of dense GPS points – significant clusters (Figure 2b and c). To create the overlays, centres were determined by computing the mean GPS points of eleven significant clusters discovered in the dataset when speeds were less than 2 km/h. The diameters of the overlay circles were determined by measuring the distance of the closest centres. This meant that centres that were very close to each other had smaller radii compared to those that were far apart as shown in Figure 2b. All the Western Cape GPS records (1,842,570) and eleven centres with radii ranging from 0.3 to 5 kilometres were used during the overlay analysis. After the overlay analysis, GPS intersects were obtained (These were records intersecting the overlay circles) and labelled by appending meta-data representing the ID of the intersecting circular layers. A total of 1,573,829 intersects were obtained. These were records in the close proximity of the significant clusters since the mean point was used as the center. From this point on, two assumptions were made. (1) That each of the eleven significant clusters were major sources and destinations of taxi trips. (2) That taxis never stopped in other places during transit from one node to the other. Results from this section opened way for our next section (network flow analysis).

C. Network analysis

During this phase of our experiment, every overlaid region (significant clusters region) was represented as a node on the network (Figure 2c). The objective of this phase was to study the departure times and trip durations between the nodes for all individual dates recorded in our dataset.

An algorithm was developed that processes the movements of the taxis between the nodes. The algorithm takes the GPS intersects, the node IDs, and the periodic interval. The periodic interval is the time (in minutes) that defines the rate at which the algorithm checks on the status of an asset to determine if it has departed from the node or arrived at the node. The interval was estimated such that it was less than the expected duration of travel of an asset between any two nodes. During our experiment, we used the interval of 10 minutes as the expected minimum travel time between any two closest nodes. Figure 3 shows the sample modes of the assets at different periodic intervals. In the table, assets A to L are monitored and their modes recorded at different time intervals. For example, when 3 is recorded in a cell, it means that the asset was registered active at node 3 during that periodic interval and N mean that the asset was recorded inactive a given periodic time.

We then analysed the records obtained in the table (Figure 3) to get a count of departures and the trip durations between nodes. During this analysis, we used a combination of two modes (active/transmitting and inactive/not transmitting) and a transition between them to explain three unique states in the modelled system – "$atNode$","$inTransit$","$inSleep$". An asset in our modelled system can only be in any of the three states at a time. When an asset is $active$ in two

 ISBN: 978-0-7972-1533-7

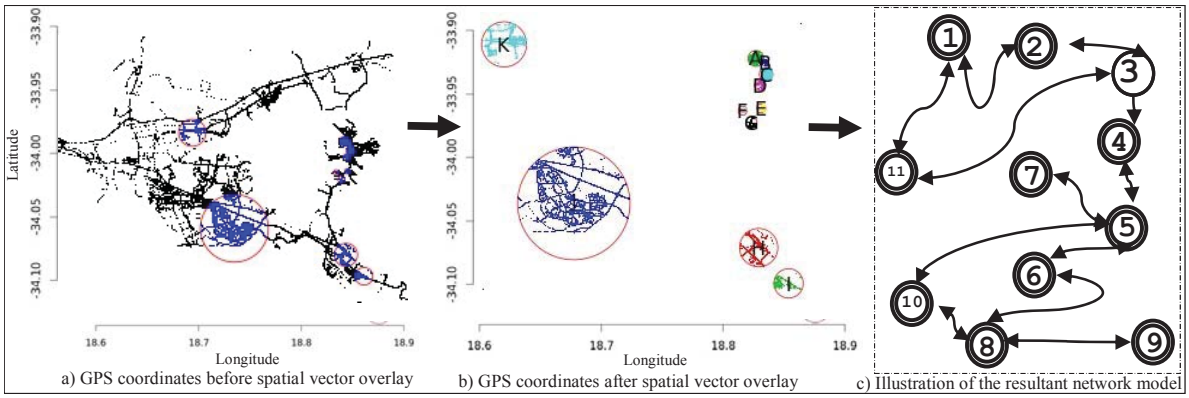


Fig. 2. Figures showing the vector overlay process and the first phase of modelling significant nodes into a graph

Periodic interval (10 minutes)

Assets tracked	15:00	15:10	15:20	15:30	15:40	15:50	16:00	16:10	16:20	16:30	16:40	16:50	17:00	17:10	17:20	17:30	17:40	17:50	18:00	18:10	18:20	18:30	18:40	18:50	19:00	19:10
A	1	3	5	5	N	3	N	4	1	1	1	7	4	3	3	3	1	1	2	N	N	2	2	1	1	N
B	1	1	1	2	N	1	1	3	N	3	1	1	3	3	3	1	1	2	2	1	1	1	3	1	N	N
C	3	3	1	1	3	1	1	5	5	1	1	3	1	1	3	N	N	3	3	3	N	8	8	8	5	1
D	N	3	N	N	N	N	N	N	3	1	1	4	N	N	3	1	1	1	1	1	1	1	N	N	1	N
E	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
F	N	7	N	N	N	N	N	3	1	1	3	3	6	6	3	1	1	1	1	3	2	1	1	1	1	3
G	N	4	3	3	3	N	3	1	1	3	N	N	N	3	1	1	N	3	3	3	1	1	1	1	N	N
H	3	3	3	7	N	8	N	8	N	N	8	N	5	1	1	3	N	3	1	1	1	3	3	4	3	N
I	6	6	3	1	1	3	1	1	1	1	3	2	1	1	3	3	N	3	1	1	3	N	N	N	N	3
J	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
K	2	N	N	N	1	1	N	N	1	1	N	N	N	N	N	N	N	1	1	1	1	N	11	11	N	11
L	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

Fig. 3. Table showing the periodic assessment of the assets (A-L) modes and the nodes(significant stops) where assets are transmitting from.

successive intervals and it is at the same node, then the asset is in "$atNode$" state, e.g asset B at 15:00,15:10 and 15:20. An "$inTransit$" state is recorded if an $inactive$ mode(s) is observed between two $active$ modes at different nodes, e.g asset D during the 17:00 and 17:10 time intervals. An asset is in "$inSleep$" state if an $inactive$ mode is recorded between two $active$ modes at the same node, e.g asset H during the time intervals 16:20 and 16:30.

The results of the table in Figure 3 are analysed by a separate algorithm which deduces two general matrices; (1) the count of departures from the node for every day (identified by the date), (2) the average duration between two nodes (complete trip). Essentially, the algorithm monitors the time and state of all assets for every time interval in the modelled process. It records an asset departure from one node to another if an "$inTransit$" state is realised (e.g in Figure 3, asset D departed from node 4 at 16:50 and arrived at node 3 at 17:20). The algorithm uses the time difference to compute and record the trip duration. For example the trip duration of asset D from node 4 to node 3 recorded between 16:50 and 17:20 was 30 minutes. Figure 4a shows the sample departure count matrix for a selected date and Figure 4b shows the sample average duration matrix for the same day in the modelled system.

IV. Results

Taxis in the Western Cape tend to exhibit a stochastic behaviour. Though the routes taken by the individual taxis tend to change most of the time. It was discovered by cluster analysis that over time these routes go through some specific places (stops) where the taxis pick up and drop passengers. However, some of these stops are not gazetted by the local authorities hence we referred to them as informal stops.

At the stops (significant), taxi departures vary according to the destination. Figure 4a shows a matrix of departure counts from different places (nodes) on a Tuesday. It is clear that there were more departures from node 3 to node 1 (177 departures) while there were no departures from node 3 to node 11. However, it was discovered that while in some cases there are no direct departures between two nodes, these nodes are still reachable through other nodes. For example, if a passenger wanted to use a taxi from node 3 (Stellenbosch) to node 11 (Bellville), he/she can get a taxi from node 3 to node 1 and then from node 1 to node 11.

By spreading the departure times at a single node throughout the day, we can tell what time segment has the most frequency of taxis departing and so we can compute the waiting time depending on the time of arrival of the traveller. Figure 4c shows the graph of varying departure frequencies from node 3 to other nodes during the day. In the plot, every line represents trend of departures to one node in the modelled system. From the graph, the peak departures from node 3 to node 1 happen very early in the morning (05:00) and at around 17:00 hours in the evening. It should be noted that the variations also depend on the destination node. For example, departures to node 6 (Somerset West) on this day have their peak at around midday. Therefore we can compute the waiting time at the node. Similar graphs comparing inter-node durations can be plotted and a similar analysis undertaken.

 ISBN: 978-0-7972-1533-7

Nodes (To)

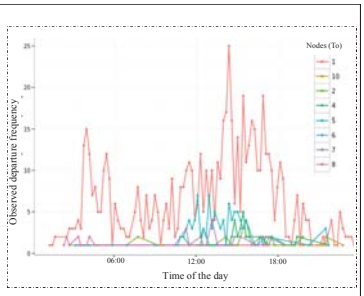
Nodes (From)	1	2	3	4	5	6	7	8	9	10	11
1		218	711	61	61	7	14	15	0	13	68
2	180		33	5	3	0	1	1	0	2	56
3	714	52		61	99	4	48	66	0	12	0
4	62	3	88		17	5	17	32	0	3	0
5	40	2	80	36		8	15	61	0	0	0
6	9	1	5	5	6		0	0	0	0	0
7	11	4	63	19	26	0		44	0	1	0
8	14	2	63	33	32	0	72		5	22	0
9	0	0	1	0	0	0	0	2		0	0
10	12	4	11	7	1	0	0	23	0		0
11	120	2	0	0	0	0	0	0	0	2	

a) Matrix of departure counts on Tuesday

Nodes (To)

Nodes (From)	1	2	3	4	5	6	7	8	9	10	11
1		10.1	12.5	12.6	10.1	10.0	10.0	41.6	0.0	24.1	23.6
2	21.4		10.2	10.0	10.0	0.0	10.0	10.0	0.0	15.0	38.8
3	10.9	19.6		10.0	10.1	10.0	13.8	19.9	0.0	19.5	0.0
4	10.0	15.0	10.8		10.1	10.0	10.0	20.1	0.0	15.0	0.0
5	10.0	10.0	13.6	10.0		10.6	20.0	13.8	0.0	0.0	0.0
6	10.0	10.0	10.0	10.0	15.0		0.0	0.0	0.0	0.0	0.0
7	10.0	10.0	10.4	10.1	13.1	0.0		15.4	0.0	10.0	0.0
8	27.5	20.0	16.4	22.4	11.7	0.0	10.8		39.4	11.1	0.0
9	0.0	0.0	30.0	0.0	0.0	0.0	0.0	15.0		0.0	0.0
10	15.5	85.0	48.6	12.8	10.0	0.0	0.0	16.1	0.0		0.0
11	24.9	35.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	270	

b) Matrix of trip durations on Tuesday



c) Graph showing the variation of departures from node three to other nodes through out the day

Fig. 4. Figures showing the matrices from the network modelling process and a graph of varying departure frequencies with time of the day.

Using the duration matrix, we can compute the duration of the passenger journey. For example a passenger who wishes to travel from node 3 to node 11, would take 10.9 minutes to get to node 1 and 23.6 minutes to get to node 11 hence a total of 35.5 minutes provided there is no delay at the intermediate node.

From our results, time optimisation is possible by choosing the most optimal travel choice. For example, in order to travel to node 11 from node 3, there are two choices, i.e from node 3 to 1 to 11 which takes 35.5 minutes and from node 3 to 2 to 11, which takes 58.4 minutes. Travellers can then choose the most optimal choices that suites their needs.

V. Conclusion

Movement of taxi between stops have been studied and modelled to provide useful information to the users of the informal public transport sector (Minibus taxi users). If utilised, users can reduce the time wasted waiting for taxis and during the journey by planning their trips in advance.

For data driven models to be more accurate and precise, there is need for a lot more data. This would assist researchers in discovering more hidden behaviour regarding the minibus taxis. Particularly there is need to study the variations of the location of stops with particular seasons of the year; the delays at the stops; and a continuous learning mechanism that will always keep the models up-to-date.

Acknowledgement

The authors would like to thank Mix Telematix and Trinity Telecom for providing data for the research, Uganda Christian University for providing computing resources to test and run the experiments.

References

[1] I. Ndibatya, M. J. Booysen, and J. Quinn, "An adaptive transportation prediction model for the informal public transport sector in africa," in *IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)*, Qingdao, China, October 8-11 2014, pp. 2572–2577.

[2] J. Zhou and S. Dai, "Urban and metropolitan freight transportation: A quick review of existing models," *Journal of Transportation Systems Engineering and Information Technology*, vol. 12, no. 4, pp. 106 – 114, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1570667211602146

[3] R. Cervero and K. Kockelman, "Travel demand and the 3ds: Density, diversity, and design," *Transportation Research Part D: Transport and Environment*, vol. 2, no. 3, pp. 199 – 219, 1997. [Online]. Available: http://www.sciencedirect.com/science/article/ pii/S1361920997000096

[4] H. K. transport department, "Third comprehensive transport study (cts-3)," *Journal of Urban Planning and Development*, 2011.

[5] D. P. Solomatine, *Data-Driven Modeling and Computational Intelligence Methods in Hydrology*. John Wiley and Sons, Ltd, 2006. [Online]. Available: http://dx.doi.org/10.1002/0470848944.hsa021

[6] H. B. Jonas Richiardi, Sophie Achard and D. V. D. Ville, "Machine learning with brain graphs," IEEE Signal processing magazine, Tech. Rep., 2011.